Machine learning based approaches for identifying sarcopenia-related genomic biomarkers in ageing males and females

Abdouladeem Dreder

Department of Computer and Information Sciences

Northumbria University at Newcastle

This thesis is submitted in partial fulfilment of the requirements of the University of Northumbria at Newcastle for the degree of Doctor of Philosophy

2017

Abstract

Sexual dimorphism of skeletal muscle can occur due to age and many of these age-related changes in skeletal muscle appear to be influenced by gender. In humans, the muscle mass peaks in the second decade while loss of muscle mass (sarcopenia) starts between the third and the fifth decade of life. In system biology, the function of genes still needs to be understood and understanding gene function remains a significant challenge. Several machine learning and computational techniques have been used to understand. However, these previous attempts have not produced enough interpretation of the impact of age on skeletal muscle mass across both gender. Although there are several thousands of genes, very few differentially expressed genes play an active role in understanding the age and gender differences. The core aim of this thesis is to uncover new biomarkers that can contribute towards the prevention of sarcopenia progress in humans according to the gene expression levels of skeletal muscle tissues. The main contributions are the development of machine learning methods based on majority voting of multi-evaluation methods and multi-feature selection methods in order to analyse microarray data and identify subsets of genes related to muscle mass loss in ageing males and females. Previously, statistical methods were used to find important genes related to the impact of age on muscle mass loss. Multi-filter and multi-wrapper based systems are proposed in this thesis to identify different and common sarcopenia-related genes in males and females based on human skeletal muscle. Genes are first sorted using three different evaluation methods (t-test, Entropy and Receiver operating characteristic). Then, important genes are obtained using majority voting based on the principle that combining multiple models can improve the generalization of the system. Experiments were conducted on three different microarray gene expression datasets and results have indicated a significant increase in classification accuracy up to 10% associated with sarcopenia when compared with existing systems.

ACKNOWLEDGEMENTS

I wish to offer my sincere thanks to my supervisors Dr. Ammar Belatreche, Dr Mohammad Tahir and Dr Muhammad Naveed for their support, advice and encouragement throughout the PhD process. Their knowledge and expertise has been invaluable and has made an immense contribution to my research.

I would also like to extend my thanks to the Libyan Ministry of Higher Education and Scientific Research for providing me with full financial support to conduct this research and produce this thesis.

Many thanks to the staff and colleagues at the Department of Computer and Information Sciences, Faculty of Engineering and Environment and the Biotechnology Research Centre (BTRC), who offered valuable advice, support and encouragement.

A special thanks also go to the PGR administration and technical staff for their wide range of guidance and support.

The most important people in my life are my family and they deserve special thanks. Sincere thanks to my mum and family members for their continuous encouragement and help throughout my research. I dedicate this thesis to them.

DECLARATION

I declare that this thesis is the result of my own work and has not been submitted in any form for another degree or diploma at any university or other institution. I also confirm that this work fully acknowledges opinions, ideas and contributions from the work of others. Any ethical clearance for the research presented in this thesis has been approved.

Name: *Abdouladeem Ahmed Sasi Dreder*

Signature: *SASI*

Data: 7 /4/2017

CONTENTS

ABSTRACT……. ……...……..…..I

ACKNOWLEDGEMENTS ……..… ……...……..…II

DECLARATION…… ……..……..…….. III

CONTENTS……..…… ……..……...……... IV

LIST OF FIGURES…… ……..……..…..….. X

LIST OF TABLES ……..……..……XII

LIST OF ABBREVIATIONS ……...…….XV

1. INTRODUCTION……..……...1

1.1 Introduction…….…….…..1

1.2 Motivation……...…….…….…..3

1.3 Aim and objective…….…….5

1.4 Thesis Contributions……...……6

1.5 Thesis scope …….…….……9

1.6 Thesis outline……..…..……10

2. BACKGROUND INFORMATION AND LITERATURE REVIEW….....…...12

2.1 Introduction…….…….…..12

2.2 Microarray technology……13

2.3 Molecular levels…….16

2.4 Skeletal muscle……..…….……...…….19

2.5 Pathways involved in muscle growth…….…….……22

2.6 Muscle atrophy……...……...…….……23

2.7 Age associated loss of skeletal muscle mass…….……..…..…….....…26

2.8 Supervised learning……...……...….……..32

2.9 Cross-validation (CV)……..……..……..…… …...…….35

2.10 Classification…….……..…….…….…….…….……..36

A . Support victor machine (SVM)……..……. …….……36

B . K-nearest neighbor (k-NN)…….…..…… …….…….…..38

2.11 Feature selection…….…….…….…...….39

A. Filter approach…….……..…….…..….40

B . Wrapper feature selection…….…….……...…….44

C . Embedded approach……...……...……...…. 47

2.12 Statistical tests……...…….… 47

A . P-value…….……..……48

B . Fold change family…….…….……..49

C . ANOVA …….…….…..49

2.13 Conclusion …….…….…….…..……..50

3. MAJORITY VOTING APPROACH TO UNDERSTAND GENDER-RELATED SKELETAL MUSCLE ATROPHY…….…….....51

3.1 Introduction…….…….....……51

3.2. Material and proposed method……..……53

3.2.1. Micro array gene expression data set……..…..….53

3.2.2 Gene subset selection using feature ranking techniques ……...……54

3.2.3 Classification ……..……...……55

A . k-NN classifier:…… ……..…….55

B . SVM classifier …….……55

3.2.4. Proposed System…….……..56

3.3 Results and discussion……...……..58

3.3.1 Study A …….…….……59

A . Case study 1: young males versus old males……..……61

B . Case study 2: old males versus old females……..…..63

C . Case study 3: young females versus old females…….…….....……..65

3.3.2 Study B……..…….……68

A . Case study 1: young males versus old males…….……..…68

B . Case study 2: old males versus old females…….……..……..70

C . Case study 3: young females versus old females…….…….....……..73

3.4 Conclusion……..…….…….……76

4. GENE SELECTION FOR MICROARRAY DATA OF SKELETAL MUSCLE MASS USING RANDOM SUBSET FEATURE SELECTION……..….... 78

4.1 Introduction……..……78

4.2 Material and proposed method…….……..…….. 79

4.2.1 Micro array gene expression data set…….……..…….……..79

4.2.2 Gene subset selection using Feature ranking techniques…….……...…….80

4.2.3 Wrapper feature selection…….…….……..81

4.2.4 Classification ……..…….……..82

4.2.5 Proposed System……..……....…..……82

4.3 Results and discussion…….……...……84

4.3.1 Study A……..…….……..84

A . Case study 1: young males versus old males……....…..…84

B . Case study 2: old males versus old females…….…...…..……..88

C . Case study 3: young females versus old females……...……..……...92

4.3.2 Study B…….…….....…...……..…95

A . Case study 1: young males versus old males…….……...……..…95

B . Case study 2: old males versus old females……..……...…….……..98

C . Case study 3: young females versus old females……...…….....101

4.4 Conclusion ……..……..……..……..104

5. MULTI-FEATURE SELECTION BASED APPROACH FOR ANALYSIS OF MICROARRAY DATA FOR SKELETAL MUSCLE MASS LOSS……...…106

5.1 Introduction……...……....106

5.2 Material and proposed method……..…….…….. 107

5.2.2 Gene subset selection using Feature ranking techniques…….……..……..…….107

5.2.3 Classification ……..…….…..……..108

5.2.4 Proposed System……..……....……108

5.2.5 Feature selection……..…….111

A Random subset feature selection (RSFS)…….…… ……..……..111

B support victor machine recursive feature elimination ……111

5.2.6 Functional annotation of top genes……..……..……112

5.3 Results and discussion……...……...……113

5.3.1 Dataset 1…….……...…….…….…..113

A Age-based differential gene expression in males.… …….…….……113

B Age-based differential gene expression in females. ……..…..……116

C Age-based differential gene expression in older adults…….…..…….....…119

5.3.2 Dataset 2……..……..……..…..121

A Age-based differential gene expression in males.… …….…….……121

B Age-based differential gene expression in females. ……..…….…….……124

C Age-based differential gene expression in older adults…….…..……..…126

5.3.3 Dataset 3……...……...…….…..129

A Age-based differential gene expression in males.… ……..…129

B Age-based differential gene expression in females. ……..…..…….……132

C Age-based differential gene expression in older adults…….…..……..…135

5.4 Conclusion …….……...…….…137

6 CONCLUSION AND FUTURE WORK……...……...…140

6.1 Thesis summary ……...…….……140

6.2 Summary of core contributions…….…….141

6.3 Difficulties and solutions……..…..……..143

6.4 Challenges…….……..…144

6.5 feature research directions…….……144

6.6 List of publications……..…….145

References……..……147

List of figures

Figure 2.1: Affymetrix gene chip ……..……27

Figure 2.2 Support victor machine …….…....……33

Figure 2.3: Diagrammatic representation of a cross structure of skeletal muscle…...…….51

Figure 3.1. Proposed Multi-Filter System (MFS) ……..……....70

Figure. 3.2. Performance of the MFS using varying numbers of genes compared with Liu et al, (2013)……...…….75

Figure 3.3. Performance of the MFS using varying number of genes.(old male versus old female)…… .....……..……78

Figure. 3.4. Performance of the MFS using varying number of genes.(young female versus old female)……...…..…….80

Figure. 3.5 Performance of the MFS using varying number of genes (old male versus young male)….. ……...…….83

Figure 3.6. Performance of the MFS using varying number of genes (young female old female). ……..……85

Figure 3.7 Performance of the system using varying number of genes (old male versus old female)……..…….88

Figure 4.1: Proposed Multi-Filter System single wrapper system (MFSWS)…….……..96

Figure. 4.2. Comparison of the performance change against varying number of genes for MFS, Liu et al (2012) and MFSWS (old male versus young male)……99

Figure. 4.3. Performance of the system using varying number of genes (old male versus old female)……...……..…….102

Figure. 4.4. Performance of the system using varying number of genes (old female versus young female)……..……106

Figure. 4.5. Performance of the system using varying number of genes(old male versus young male)……..109
Fig. 4.6. Performance of the system using varying number of genes(old female versus young female)……111
Figure. 4.7. Performance of the system using varying number of genes (old male versus old female)……...….114
Figure5.1 Proposed Multi-Filter System (MFMWS)……..…..…..…121
Figure. 5.2. Performance of the system using varying number of genes (old male versus young male)…….…….....125
Figure. 5.3. Performance of the system using varying number of genes (young female versus old female)……...…….…….…….…….128
Figure. 5.4. Performance of the system using varying number of genes (old male versus old female)……...…….…….132
Figure. 5.5. Performance of the system using varying number of genes (old male versus young male)……..……..135
Figure 5.6: Performance of the system using varying number of genes (old females versus young females)…… ……..……..…….138
Figure 5.7: performance of the system using varying number of genes (old female versus old male)……141
Figure 5.8 Performance of the system using varying number of genes (old female versus old male)……143
Figure. 5.9. Performance of the system using varying number of genes (old female versus young female)……..149
Figure. 5.10. Performance of the system using varying number of genes (old female versus old male)……..……150

List of tables

Table 2.1: Confusion matrix for two-class classification problem…….…..30
Table 2.2: A common types of filters methods……..…….38
Table 2.3: Advantages and advantages of the wrapper me…….……..……42
Table 2.4 Strength of evidence against the null hypothesis according to p-value……..……45
Table 3.1 Majority voting to select important genes. ……..……..71
Table 3.2 Confusion matrix for two-class classification problem……..……72
Table 3.3 microarray dataset……..……73
Table 3.4 Confusion matrix : young men versus old men……..…..…..74
Table 3.5 Classification performance: Young men versus Old men. ……74
Table 3.6 New genes selected by the proposed system. Common genes selected by proposed system and system by Liu et al, (2013)……...…76
Table 3.7 Confusion matrix : old male versus old female…….……77
Table 3.8 Classification performance Old men versus Old women. ……..……77
Table 3.9 show some of the common and different genes that obtained by MFS and Liu et al, (2013) …….…..……78
Table 3.10 Confusion matrix : old male versus old female…….…..……79
Table 3.11 Classification performance: Young female versus old female…...….……..80
Table 3.12 some of the common and different genes that obtained by MFS and Liu et al, (2013) …….……..80
Table 3.13 CLASSIFICATION PERFORMANCE young men versus old men…….……..82
Table 3.14 Confusion matrix : young men versus old men…….……..82
Table 3.15 List of genes identified by the proposed MFS method and those common to both the proposed method and the method by Raue et al, (2012) …….……83
Table 3.16 Classification performance: young female versus old female…….…….84
Table 3.17 Confusion matrix : young female versus old female…….……85

Table 3.18. List of common and different genes between Raue et al, (2012), and the proposed MFS……..……..……86
Table 3.19 Classification performance: old male versus old female……..……87
Table 3.20 Confusion matrix : old male versus old female……..……..……87
Table 3.21 list of common and different genes between Raue et al, (2013) and the proposed MFS ……..……....88
Table 4.1 Majority voting to select important genes. ……97
Table 4.2 Performance comparison between MFS, Liu et al (2013) and MFSWS (young male versus old male)……..……98
Table 4.3 Confusion matrix : young men versus old men……..……..…98
Table 4.4: new genes selected by the proposed system MFSWS (young male versus old male) ……...…..……..……99
Table 4.5: P-values and FC of genes in male with brief descriptions……....100
Table 4.6: Classification performance: old men versus old women. ……..……101
Table 4.7 Confusion matrix : old male versus old female……..……..……102
Table 4.8: P-values and FC of genes in adults with brief descriptions (old male and old female)……...…….103
Table 4.9 Classification performance: young females versus old females. ……..….105
Table 4.10 Confusion matrix : old male versus old female……105
Table 4.11: P-values and FC of genes in female with brief descriptions……107
Table 4.12 Classification performance:Young males versus old males……..…108
Table 4.13 Confusion matrix : young men versus old men……109
Table 4.14: P-values and FC of genes in men adults ……..…110
Table 4.15 Young females versus old females……110
Table 4.16 Confusion matrix : young female versus old female……111
Table 4.17: P-values and FC of genes in adults with brief descriptions. ……112
Table 4.18 Classification performance: old males versus old females……113
Table 4.19 Confusion matrix : old male versus old female……113
Table 4.20: P-values and FC of genes in adults with brief descriptions……..115
Table 5.1 majority voting to select important genes. ……..……122
Table 5.2 CLASSIFICATION PERFORMANCE: old men versus young men…….……125

Table 5.3 Confusion matrix : young men versus old men……..125
Table 5.4 P-values and FC of genes in adults with brief descriptions ……..…..127
Table 5.5 CLASSIFICATION PERFORMANCE: young women versus old women……..……..129.
Table 5.6 Confusion matrix : old male versus old female…….……..129
Table 5.7 P-value and FC for some genes with briefly descriptions…….……..130
Table 5.8 Classification performance: old men versus old women……133
Table 5.9 Confusion matrix : old male versus old female……..……..……133
Table 5.10 P-value and FC for some genes with briefly descriptions……..……133
Table 5.11 Classification performance: old males versus young males. …….……..…..135
Table 5.12 Confusion matrix : young men versus old men…….……136
Table 5.10: P-values and FC of genes in Adults (P-value <0.05) ……..136
Table 5.11: Classification performance of MFMWS, MFSWS and MFS…….……137
Table 5.12 Confusion matrix : young female versus old female…….……..……..137
Table 5.13: P-values and FC of genes in adults……..…….139
Table 5.14: Classification performance: old males versus old females……..…..……140
Table 5.15 Confusion matrix : old males versus old females……..……..……140
Table 5.16: P-values and FC of genes in Adults…..……142
Table 5.16 Classification performance: old males versus young males…….144
Table 5.17 Confusion matrix : young males versus old males……144
Table 5.18 P-values and FC of genes in males ……145
Table 5.19 Classification performance: young females versus old females…….147
Table 5.20 Confusion matrix : young females male versus old females…….……147
Table 5.21 P-values and FC of genes in female with brief descriptions. ……148
Table 5.22 Classification performance: old men versus old women……..150
Table 5.23 Confusion matrix : old male versus old female……150
Table 5.24 P-values and FC of genes in adults human with brief descriptions……151

List of abbreviations

ALL	Acute lymphoblastic leukaemia
AML	Acute myeloblastic leukaemia
A	Adenosine
ANOVA	Analysis of variance
BWL	Body weight loss
COPD	Chronic Obstructive Pulmonary Disease
CLIFF	Clustering via alternative feature filtering
CC	Correlation coefficient
CFS-SF	Correlation-based Feature Subset Selection
CO	Cosine
CV	Cross-validation
C	Cytosine
DM	Data mining
DT	Decision tree
DNA	Deoxyribonucleic acid
DLDA	Diagonal linear discriminant analysis
FN	False negatives S
FP	False positives
FSS	Feature subset selection
FC	Fold change

GEO	Gene Expression Omnibus
GF	Gene function
GA	Genetic Algorithm
G	Guanine
IG	Information gain
IMF	Inter myofibrillar
KNN	K-nearest neighbor
KDD	Knowledge-discovery in Databases
LOOCV	Leave one out cross validation
ML	Machine learning
miRNAs	MicroRNAs
MFMWS	Multi-filter multi wrapper system
MFSWS	Multi-filter single-wrapper system
MFS	Multi-filter system
NB	Naive Bayes
NCBI	National Centre for Biotechnology Information
ANN	Neural network classifier .
ROC-AUC	Receiver operating characteristic-area under curve
RFE	Recursive feature elimination
RBF	Redundancy based filter
RL	Resistance loading
SFBS	Sequential backward floating selection
SBS	Sequential backward selection

SFFS	Sequential forward floating selection
SFS	Sequential forward selection
SAGE	Serial analysis of gene expression
SNR	Signal-to-noise ratio
SNP	Single nucleotide polymorphism.
SM	Skeletal muscle
SMA	Skeletal muscle atrophy
SMT	Skeletal muscle tissue
SS	Sub-sarcolemmal
SVM	Support victor machine
T	Thymidine
TN	True negatives
TP	True positives

Chapter 1

Introduction

1.1 Introduction

The loss of skeletal muscle mass and function due to ageing is known as sarcopenia and leads to functional limitations among the elderly. Muscle atrophy is reported to be a main risk factor of disability and potentially mortality among the elderly. Sarcopenia has many of concepts since it started in 1987 (Rosenberg, 1989). The term 'sarcopenia' has been used to characterize muscle wasting in the elderly. Muscle atrophy is normally starts after fourth decade of life (bell et al, 2016). Sarcopenia is a strong risk factor for decrease mobility, events like falls and fractures in addition it often associated with rates of hospital and long-term care admissions in human (Bauer et al, 2015), it has been observed that the proportion of muscle mass loss is about 6% every ten years (Flynn et al., 1989). It has been proven to be associated with falls among the elderly.

Gene function (GF) is still the main challenge in system biology. Earlier, many machine learning and computational techniques were applied to understand GF (Hiesinger and Hassan, 2005; Pranckeviciene, 2015; Bodine et al, 2001; Sacheck et al, 2007). However, most of previous studies did not comprehensively interpret the impact of genes on skeletal muscle in both males and females because these studies do not use all gender and age classes. For example, Paul et al. (2005), and Rivas et al, (2014) have conducted a comparison between young males and old males only. Raue et al (2007) have investigated muscle atrophy based on only female

while Stephen et al., (2002) investigated only muscle atrophy based on old male and female only.

In recent years, advances in genomics and proteomics have led to the generation of a huge quantity of biological data, such as that related to the gene expression profile, protein and single nucleotide polymorphism (SNP). One of the most important tissues is skeletal muscle, which represents about 50% of the human body mass. As mentioned above, skeletal muscle mass loss and function impairment in the elderly is called sarcopenia and leads to muscle mass decrease due to an impairment in gene function among older individuals.

Microarray experimentation is becoming very important in order to measure gene activity, such as pre-disease screening, and for general health checks. It is used in the bioinformatics field to provide an insight into gene interactions and disease pathways with the potential for disease prognosis, the discovery of new disease groups, molecular marker identification and the prediction of therapeutic responsiveness (Golub et al., 1999; Dupuy and Simon, 2007; Yu et al., 2007; Wang et al., 2008). Microarray consider as basic tools of molecular biology, it allow for simultaneous assessment of thousands of genes expression. But analysis of microarray data is not easy Jaksik et al, (2015). Microarray consists of glass slides of specific oligonucleotide probes Oyelade et al, (2015). It comprises the measurement of thousands of microscopic spots of DNA probes. However, only a small subset of genes is related to the issue of interest. Therefore, techniques for extracting the informative genes, that underlie the pathogenesis of cancer cell spread, from high dimensional microarrays are necessary (Yu et al., 2007; Osareh and Shadgar, 2008; Wang et al., 2008; Zhang et al., 2008). There is also a need to develop algorithms that undertake such a complex task. This puts the computational analysis of microarray studies at the forefront of research.

Microarray gene expression data is characterised by sample scarcity and high gene dimensionality, and the behaviour of genes is complex (i.e. the interaction between genes

within the data). This poses big challenges related to the development of computer algorithms for cluster discovery, prediction of class and biomarker discovery, as well as to derived a biological interpretation of features that can be responsible for the causing disease. This reflects the challenging nature of microarray analysis. In microarray data experiments, using feature ranking is essential in order to reduce the data dimensionality and develop tractable analysis methods. In biological data studies, bringing together a range of machine learning techniques and designing a novel framework lead to better understanding of biological processes and solving certain biological issues. The impact of age on skeletal muscle is still unclear, therefore the discovery of more reliable biomarkers related to skeletal muscle mass loss has now become essential.

1.2 Motivation

In previous years, the primary aim of bioinformatics was to conserve biological information such as gene expression and the sequence of amino acid. However, advances in bioinformatics technology have led to the generation of a huge amount of data with hundreds or even thousands of rows. Thus, the role of bioinformatics has changed; nowadays the aim of bioinformatics is to provide a general overview of different types of data through data analytics in order to produce meaningful information. These requirements have raised awareness among researchers and prompted them to start developing new methods to interpret biological problems.

In the current bioinformatics era, motivated by the rapid development of computers coupled with advances in data mining techniques, this thesis aims to work with biological data in order to contribute towards addressing certain biological issues by identifying a distinct subset of genes related to muscle atrophy. Typically, the purpose of applying data mining techniques to this area is to interpret the relationship between molecular levels (proteins, gene expression

and SNP and biological issues. In the data mining field, there are several methods used to address biological problems, such as modelling techniques, software tools and statistical analysis. Recently, data mining has witnessed remarkable development, not only in terms of techniques but also in terms of applications. In the post-genome era, the accumulation of a huge amount of biological datasets provides great opportunities for bioinformatics and biological researchers, but it is also accompanied with tough challenges.

Much effort has been carried out in this area. However, most work has been concerned with disease cases, while little work done with investigating healthy cases. Although there has been some investigation related to control data, however most previous studies have used basic statistical analysis and have not used complete training data during the analysis (e.g. Liu et al., 2013; Raue et al.,2012), which have affected the results. Moreover, a few studies attempted to determine the impact of age on human skeletal muscle atrophy based on gene expression dataset. For example, Liu et al 2013) is the first study which provided global evidence for presence of extensive gender differences in ageing process of human skeletal muscle.

1.2 Aim and objective

It is believed that there are a subset of genes which have an active role on muscle atrophy across both males and females. This sub set of genes normally spread in older females more than older males, most of previous studies such as (Lui etal, 2013) and (Raue et al, 2012) have used statistical methods such as t-test and ANOVA software in order to identify the significant subset of age-related genes

The aim of this research is to develop a novel framework which able to identify a subset of genes related to sarcopenia across both men and women. Microarray data collection involves thousands of gene expression skeletal muscles. In this research the datasets represent gene expression derived from the skeletal muscle tissue (SMT) of healthy males and females of various ages.

Handling large amount of data and the enormous number of genes in microarray datasets with different variances represent a major challenges for this thesis. In order to overcome these

challenges, various data mining and machine learning techniques have been investigated to identify important subsets of genes which can be used as biomarkers for muscle mass loss in the elderly. They include feature ranking, feature selection and classification. In addition, statistical tests such as P-value and fold change (FC) are considered. For example, fold change is used to determine the level of a gene expression (Huang and Xu, 2007) and the P-value test is used to compare the means of two groups of samples. Each microarray dataset in this thesis has involved the division of the data into three main cases. They include old females versus old males, old males versus young males, and old females versus young females.

The main objectives of this research are as follows:

- To source existing microarray datasets suitable for the proposed research
- To develop feature selection approaches which are able to determine a subset of genes associated with age in both genders with high level of accuracy. When applied on different datasets also these genes can be used to prevent sarcopenia progress in both male and female.
- To identify objective evaluation matrix and use them to evaluate the performance of the developed approaches.
- To compare the performance of the proposed approaches against existing studies and analyse potential biological significance of the subsets of genes obtained through the proposed approaches.

1.4 Thesis Contributions

In order to identify important genes related to specific problem based on microarray data, the decrease of the size of dataset is the first step. Therefore our first attempt was aim to use each evaluation method separately, in order to re-rank the genes. However the results show that each evaluation method has yielded different subset of genes also the classification accuracy was nearly same. Therefore we developed a new framework based on a combination of ML techniques such as feature selection and

classification methods with evaluation methods such as t-test, entropy and ROC-AUC in order to improve the classification accuracy and to identify more reliable subset of genes related to sarcopenia

The thesis contributions are as follows:

1) A Multi-filter System (MFS) for ranking and selecting subsets of sarcopenia related genes in males and females is proposed. The system is based on majority voting of three evaluation methods including t-test-entropy and Receiver operating characteristic-area under curve ROC-AUC. A single evaluation method has its own selection criterion and consequently it yield different ranked genes. As a result, some important genes were excluded in each method. Using majority voting based on a combination of different evaluation methods (namely T-test, entropy and ROC-AUC) has provided more reliable sarcopenia-related genes with high performance of accuracy in each case study related to age and gender differences (i.e young men male versus old male; young female versus old female; and old male versus old female). Findings from this research work have been published in the following conference and journal papers.

2) A Multi-Filter Single Wrapper based system (MFSWS) is proposed to deal with the limitations found in the previous MFS approach. Although the MFS was able to achieve competitive results compared to the previous studies by Liu et al (2013) and Raue et al (2012), some important genes were excluded which affects the classification accuracy. Therefore, in order to improve the quality of the selected subset of genes and subsequently improve the system classification accuracy, we propose to use wrapper feature selection due to its ability to deal with such microarray datasets which involve several thousands of features. The main advantages of the wrapper feature selection are the prevention of over-fitting and the consideration of the interaction between genes.

In this contribution we adopted random subset feature selection (RSFS) where, unlike sequential floating forward selection (SFFS), all features are evaluated in terms of their

average efficacy in the context of many other feature combinations. The RSFS approach was proposed to filter-out the list of genes that are produced by a majority voting based on the combination of three evaluation methods and then the final list of genes produced by RSFS are fed to the SVM and KNN classifiers. Results show that the proposed MFSWS system based on RSFS is able to achieve higher classification performance for each case of study in comparison with the MFS.

3) A third system based on the combination of multi-evaluation and selection methods instead of a single feature selection method is proposed. Although the MFSWS is able to achieve high accuracy using the new list of genes which is obtained based on RSFS, using different single feature selection methods yielded different subsets of genes. Therefore, that there are still important genes that were excluded using this system. Therefore the proposed multi-filter multi-wrapper system (called MFMWS) considers the majority voting based on a combination of different feature selection methods to overcome the MFSWS shortcomings. In this approach, the final list of genes was obtained based on two majority voting systems. In the first one, the top ranked genes based on multi evaluation methods are given to different feature selection methods RSFS (Räsänen, and Pohjalainen 2013) and. The SVM-RFE is an advanced version of support vector machine recursive feature elimination SVM-RFE presented by Guyon et al (2002). SVM-RFE is an approach for gene selection. Then the majority voting is applied to the two lists that are obtained by RSFS and SVM-REF. Later, the final list of genes are classified using KNN and SVM and the obtained results revealed that the proposed MFMWS achieved excellent classification performance for each case of study compared with the MFSWS. The genes selected by this approach are considered sarcopenia-related genes as shown in the biological interpretation results. The obtained genes are determined using statistical tests such as (P-value ($P<0.05$) and fold change (FC) and the pathway of these genes was examined by the DAVID software which “provides a

comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes" (https://david.ncifcrf.gov/) (Huang et al., 2007). Results show that the proposed MFMWS system is able to identify distinct age-related genes with a higher level of accuracy for each case of study compared with existing systems. Moreover, P-values of these genes were less than 0.05 which that there are statistically significant differences between the means of two groups

The systems proposed in the above-mentioned contributions were evaluated using three different microarray datasets, all of them are publicly available on the Expression Omnibus (GEO), a public functional genomics data repository (http://www.ncbi.nlm.nih.gov/). Each dataset contains 54,623 genes, as outlined below:

- First dataset (GSE38718): this dataset includes a total of 22 healthy males and females of various ages distributed as follows: 11 males (7 young and 4 old) and 11females (7 young and 4 old).NCBI (2013). were young (20-29 years old), and old (61-81 years old)
- Second dataset (GDS5216): this dataset includes a total of 36 healthy subjects distributed as follows: 19 females aged 8 to 11 and 17 males aged 7 to 10. NCBI (2012). Were young (24 ±1 years), and old (84 ±1 years).
- Third dataset (GDS5218): This dataset includes a total of 55 subjects distributed as follows: 16 younger males, 15 younger females, 12 older males and 12 older females. NCBI (2012). Were young (24 ±1 years), and old (84 ±1 years).

1.6 Thesis outline

This thesis comprises six chapters.

Chapter 2 provides a general overview of data mining and how it can be used in an effective way to address such problems. The advancement of machine learning techniques is also presented. Bioinformatics relevant to such issues are discussed along with the importance of gene expression skeletal muscle to address such issues. Age-associated gene selection in previous studies is also described.

Chapter 3 describes the proposed multi-filter system (MFS) system which is designed to reduce the number of genes from the original data based on majority voting of three evaluation methods and to increase gene selection system efficiency. The idea is based on a combination of three different types of evaluation methods including T-test, ROC, and Entropy based on majority voting of three evaluation methods.

Chapter 4 discusses the multi-filter single wrapper system (MFSWS) framework. In addition, it provides a brief description of the wrapper based feature selection which is used in this proposed system and is called random subset feature selection method RSFS.

Chapter 5 introduces the proposed system multi-filter multi-wrapper (MFMWS) which aims to discover a more reliable subset of genes related to a particular case study. Moreover, this chapter gives details about two different feature selection methods. Also it describes the three different datasets that are chosen to evaluate the proposed system.

Finally, the sixth chapter draws conclusions and discusses future research that could potentially extend the current work Next chapter will review the background of datamining, machine learning and statistical tests also it gives information about microarray technology and it discuss skeletal muscle atrophy.

Chapter 2

Background and literature review

2.1. Introduction

Data mining emerged in the research community in 1980 (Shafique and Qaiser,2014).. By the early 1990s, data mining had become known as Knowledge-discovery in Databases (KDD) or the science of analyzing datasets or databases to gain useful information; for example, to discover patterns in data that can lead to obtaining precise predictions. The huge amount of biological data available provides both opportunities and challenges in relation to development of new KDD methods. Mining of biological data uses a combination of computer science and statistical tools to gain useful knowledge from gathered datasets (Raza, 2012). The aim of data mining applications is to find meaningful patterns in data and accurately interpret them. Data mining consists of supervised and unsupervised techniques. Classification, estimation and prediction are examples of supervised techniques. In the case of an unsupervised techniques, such as clustering and description & visualization no variable is singled out as the goal; the target is to establish the relationship between all

the variables. Unsupervised learning used to identify patterns without the use of a given target field (Pranckeviciene, 2015). Artificial intelligence algorithms, such as genetic algorithms and neural networks, have witnessed great developments in terms of obtaining non-linear relations. In the field of biology, data mining can be exploited to predict gene relations in genomes (Farooqi and Raza, 2012), hence it provides an opportunity to interpret or discover

new and useful information related to biological data and contributes towards solving certain biological problems.

Bioinformatics was established by Paulien Hogeweg in 1979. The genomics and genetics fields were the first fields to use bioinformatics techniques in the 1980s, in particular large-scale DNA sequencing (Raza et al., 2012). Bioinformatics can be defined as a way to analyse and manage genomic data which have been produced in previous decades. It has allowed the integration of research fields such as computer science, biology and information technology, and has facilitated the process of analyzing, extracting and interpreting biological data. The increased power of computers and advances in genomics have spurred bioinformatics researchers. Nowadays, it is possible to get genomic structure information and to apply an analysis process in order to understand gene expression patterns. last decade have witnessed a noticeable increase in the analysis of biological data because this is freely available through websites (Bresell et al., 2008).

2.2 Microarray technology

The advancement of technology has motivated researchers around the world to determine the interactions between multiple components during a signalling pathway. In muscle atrophy studies, the most common method is serial analysis of gene expression (SAGE), which permits analysis of a huge number of transcripts at the same time. There is another approach which uses Affymetrix GeneChip microarray analysis (Sasik et al., 2004). Microarray technologies allow the monitoring of expression levels for thousands of genes. This novel technology has introduced a new form of biological classification on a genome-wide scale. Nowadays, biologists and other specialists have changed the strategy of their experiments due to new technology which has allowed them to investigate all organism genes in one experiment. In the genomic area, microarray technology is prevalent and it has various

applications that are used in medicine and biology. Some of these studies focus on cancer classification (van't Veer et al., 2002).

The microarray technique was created in the 1990s as a result of many efforts to reduce the time taken for the drug discovery process (Lenoir and Giannella, 2006). In earlier drug discovery, the candidate medications were used only one-by-one against diseases (Babu, 2004).

The first type of microarray was presented by the sister company Affymetrix (as shown in Figure 2.2) and was called GeneChip. Nowadays, the GeneChip microarray is one of the most important tools in biology. Indeed, it has become an indispensable technique for many biologists and is used to monitor gene expression levels in specific organisms. A microarray consists of a slide of glass called spots and the DNA molecules are sorted in these spots. It comprise several thousands of genes, each one of them comprising a million or more copies of corresponding DNA molecules that are uniquely identical to the gene. In order to make a comprehensive view of the cellular function, the traditional method gene-by-gene is not enough, while the microarray can measure the gene expression activity from a whole genome in only one experiment for example it can simultaneously measure the expression level of thousands of genes within a particular mRNA sample(Tarca et al,2006).

Microarray technology was developed to cover many thousands of genes. Typically, the gene expression profiles are stored in microarrays. This helps researchers to examine gene patterns at any time to address a particular gene expression issue. Microarray analysis contributes towards distinguishing between a normal cell and an infected one. The results of analysis can be exploited to recommend biomarkers to diagnose a disease. There are two main types of microarrays, the Stanford Microarrays and Array Express (NCBI), both of which are publicly available online (Kashyap et al., 2015).

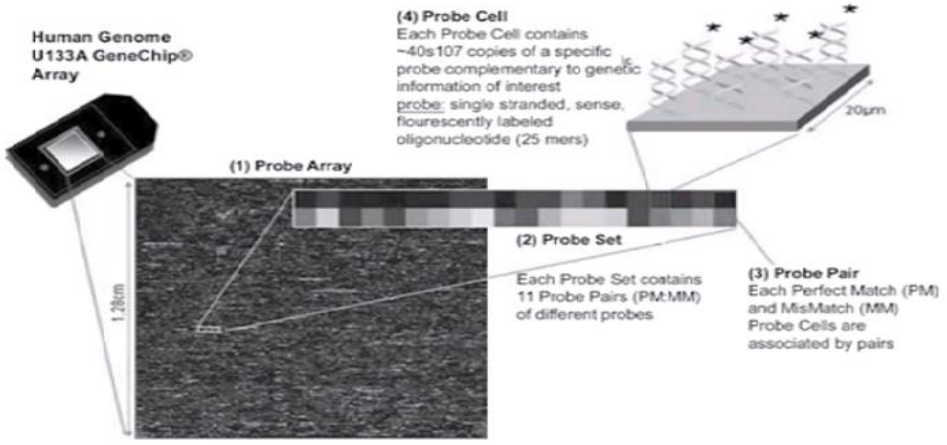

Figure 2.1: Affymetrix gene chip copyright the Oceon Ridge biosciences

To summarise, microarray is the most important source of biological information. It can be exploited to provide answers about a wide range of biological questions, including:

- Identifying genes that are highly expressed (Enrightal, 2000; Feil et al., 2003).
- Providing information about the progress of a disease at different times (Vos et al., 1995; Van et al., 2004).
- Comparing genes to distinguish between normal and diseased cells (Koning et al., 2002).

The main function of class prediction is to develop a multivariate function for accurately predicting class membership of new patients (Sánchez et al, 2008). It is used to determine whether two genes have the same expression patterns or not.

In this thesis, we will place emphasis on classification and prediction. Firstly, we will use supervised learning methods (Miller et al., 2002). Secondly, different feature selection techniques will be proposed to identify discriminatory features, concurrently. Classification

algorithms will be proposed to create models on training samples and prediction. Information sources can be divided into sequences of protein, DNA and genome. Thus, the researchers can pass through these resources to the several products of a gene. Bioinformatics techniques provide a good opportunity to understand or solve several molecular biology problems.

The aims of using bioinformatics can be summarized as follows:

Firstly, it allows researchers to access biological data such as the Protein Data Bank and to analyze them. Secondly, it can improve algorithms that are used to analyze datasets. Finally, such algorithms may help researchers to discover novel and meaningful information and to interpret biological processes (Luscombe et al., 2001). In terms of genomes, few genes are completely understood; however, biological processes and gene functions are still unclear

2.3 Molecular levels

MicroRNAs (miRNAs) are small fragments of RNA. They can interact with messenger RNAs (mRNA). As a result of this interaction, mRNA transforms into protein. During this transformation certain proteins contribute greatly towards regulating gene expression. The molecular levels include gene expression, protein and single nucleotide polymorphisms (SNP) (Tessler et al, 2011). Deoxyribonucleic acid (DNA) consists of a chain of nucleotides that has four types: Adenosine (A), Guanine (G), Cytosine (C) and Thymidine (T). The DNA molecule stores cells genetic information and includes several genes. Cell function uses diverse genes to create proteins by copying the code of the gene into messenger ribonucleic acid RNA (mRNA) in a procedure called transcription. The copying process of a DNA sequence of genes into mRNA sequences is called Gene expression. The mRNA sequences are then converted into proteins amino acid sequences while the number of created RNA

copies is called gene expression level. Human body cells have the same genes, some of them are more expressed than others.

Gene expression is the process of transcribing a gene's DNA sequence into mRNA sequences, which are later converted into amino acid sequences of proteins. The number of copies of produced RNA is called the expression level of the gene. Gene expression level regulation is considered significant for proper cell function. Microarray technologies are considered an effective way to investigate the regulation of gene expression, providing the opportunity to concurrently measure the levels of expression in several thousands of genes in cells. Gene expression data from DNA microarrays includes a huge number of genes in a few samples. The expression profile of samples is the expression levels for all genes under one empirical condition for all genes. Gene expression profiles produced by microarrays could help researchers understand the cellular mechanism of biological processes. To understand the cellular mechanism of biological process and which genes are more active compared to other genes in specific issues. Gene expression can be located by measuring mRNA levels using the microarray technique that contains thousands of genes which can be exploited to identify the differentially expressed genes. It also helps researchers to interpret the biological process of cellular mechanisms, for example, according to the information provided by a gene expression profile in the cell's tumorous mutation, researchers can identify the significant genes that could have led to this mutation (Edgar et al, 2002). In recent years, the throughput of genome technology has become very popular because many experiments of molecular biology use this technique in order to measure the molecular abundance of mRNA and DNA.

Gene expression profiling is considered as a powerful and reliable approach. It empowers genetic researchers to add new prospects to their capability of transforming genomic

information into useful information. for example there were more than 30 million independent measurements of gene expression between 1996 and 1998. This indicates that the tools available to deal with data processing are insufficient (Antonov et al, 2014).

Identifying the small subset of genes that can attain good accuracy represents one of the main issues of microarray data classification. Several studies have considered clinical researches that have used gene expression profiling as a key during their investigations.

In bioinformatics, the size of data is dramatically increased due to the huge number of studies. For example the European Bioinformatics Institute (EBI) is one of the biggest biology-data stor. The availability of high throughput of numeric data contributed to the growing volume of data. The advanced bioinformatics technologies motivate the researchers in this field to extend their studies in order to solve complicated biological issues. For example, many attempts were carried out to identify important biomarkers related to given biological. The results of experiments that use bioinformatics techniques to analyse biological data are considered more accurate than just used statistical analyse (Kashyap et al, 2015).

Recently, bioinformatics data volume has risen dramatically, for example one of the big sources of biological data is The European Bioinformatics Institute (EBI). In 2013 it had 8 petabytes of molecular level data such as protein and genes. This data volume increased to approximately 40 petabytes in 2014 (Kashyap et al, 2015). Each year the size of data increases to double that of the previous year. There are other sources of big biological data such as the National Institute of Genetics, the National Centre for Biotechnology Information (NCBI),in Japan and USA . These big datasets are publicly available through the centres website. In bioinformatics investigation, there is a huge volume of several types of data that are widely used such as DNA, RN and gene expression data.

2.4 Skeletal muscle

Skeletal muscles consist of several thousand muscle fibres (muscle cells), all of these fibres are multinuclear. Skeletal muscle mass represents more than 40% of the human body. Basically, the skeletal muscle function generates force to provide movement. The regulation of this function depends on particular proteins and gene expression.

Skeletal muscle is characterised with a high degree of flexibility. One main feature of muscle is its plasticity. Notwithstanding any decrease or increase in muscle mass, skeletal muscle can adapt in response to stimuli with different situations, such as with a decrease or increase in speed of movement, or a metabolic disorder which usually happens because of disease or ageing. Testing skeletal muscle in human beings is challenging due to experimental restrictions. The number of studies is also limited due to ethical issues that surround these experiments.

In 1953, Hanson and Huxley presented the structure of skeletal muscles, and reported that each repeating sarcomere involves 2 types of protein, myosin and actin. In recent decades, the skeletal muscle type has been classified based on biochemical properties, each muscle involves high levels of myoglobin; mitochondria is classified as red fibres, while those muscles containing low myoglobin and little mitochondria are classified as white fibres I (Dubowitz and Pearse, 1960) and (Gauthier and Padykula, 1966). In mammals, skeletal muscle is composed of several myosin heavy chain MHC proteins, and their expression is different. In 2011, Schiaffino and Reggiani reported that MHC types such as IIb, I and IIa, have the main expression in the trunk and limbs. There are several ways to regulate the MHC gene and protein expression; for example, in the elderly or during mechanical loading. Usually, the elderly are associated with a loss of skeletal muscle mass which leads to a decrease in the ability to force production (Deschenes and Rooyackers, 2004). The main

function of a particular skeletal muscle in mammals is identified by the ratio of the types of fibre in the muscle. There are four types of fibre: type I, type IIA, type IIX and type IIB. Each type differs in contractility speed, mitochondrial content and energy metabolism. Moreover, each type expresses a distinct myosin heavy chain isoform.

Slow twitch type I fibres have a greater oxidative/aerobic capacity, increased mitochondrial content and are defined by the expression of the slow myosin heavy chain isoform MYH7. Type II fibres, in general, are fast twitch fibres which have a more glycolytic/anaerobic energy metabolism and decreased mitochondrial content. Type II fibres have higher ATPase activity levels compared to type I and produce lactate during energy production, thus are more easily fatigued (Scott et al., 2001).

In general, type I fibres have a larger aerobic capacity and increased mitochondrial content. While type II fibres are faster twitch fibres compared to type I fibres because they have more anaerobic energy metabolism and decreased mitochondrial content. Scott et al. (2001) reported that type I fibres have lower ATPase activity levels compared to type II, therefore type II fibres are more easily fatigued. Type II fibres also have an upper maximal velocity of shorting (Vmax), unlike type I fibres.

In humans, changes in muscle mass can be the result of many physiological effects on muscle degeneration and synthesis. Normally these influences may appear over decades of life, such as through progress to adulthood or in the elderly. The period of muscle mass changes continues and is considered a part of normal growth in adults until the third decade of life. Janssen (2011) declared that the loss of muscle mass and strength starts after the fourth decade, with about six percent loss per decade in humans. There have been several studies conducted to address the loss of muscle size.

D’Antona et al. (2003) documented that Vmax in ageing skeletal muscle is slower in type IIa. There is also evidence to suggest that the size of type II fibres decreases with advanced age (Larsson et al., 1978). Recent findings of D’Antona et al. (2003) established that young and ageing fibres differ and these can be distinguished. Moreover, it has been recently discovered that Vmax might be a disorder in old people due to inherent alterations.

2.5 Pathways involved in muscle growth

Skeletal muscle is characterized by adaptation and changing muscle mass and type of fibre according to functional needs. Thus, muscle mass regulation is not simple and includes pathways. The result is atrophy (decrease of muscle mass) or hypertrophy (increase of muscle mass). Mitch and Goldberg (1996) documented that synthesised protein balances with protein degeneration in healthy adults, therefore the muscle mass in these people is comparatively steady. The change rates in muscle mass loss or muscle mass gain due to an increase in protein degradation or protein synthesis.

In hypertrophy, the main processes, protein synthesis and satellite cell recruitment have an active role in the increase of muscle mass size. Disorder of these processes is observed in Chronic Obstructive Pulmonary Disease (COPD) muscle atrophy.

Hawke and Garry (2001) reported that satellite cells are located in the space between the basal lamina and the sarcolemma; often these cells are quiescent. Satellite cells bail up to the damaged fibre in order to repair the muscle.

Despite these cells having an active role in fibre growth, their contribution towards hypertrophy in muscles is still unclear (O'Connor and Pavlath, 2007; McCarthy and Esser, 2007). However, in 2012, Kudryashova et al. reported that the deterioration of satellite cell function may cause muscle atrophy. Several studies, such as Goldspink et al. (1983) and

Bolster et al. (2004), have suggested that protein synthesis has an active role in muscle growth and hypertrophy. In addition, it may lead to an increase in protein degradation.

2.6 Muscle atrophy

The loss of muscle size and weight is called skeletal muscle atrophy; it is often accompanied with a decrease in muscle fibre. In humans there are several factors that can lead to skeletal muscle atrophy, including tumours, laziness and extreme training.

The skeletal muscle represents more than 45% of the human body, it helps during movement, and the cell organization of this organism is highly structured. Mitochondria are cellular organelles and the main function is to convert metabolic fuels, such as glucose and fatty acids, into energy (Sandri, 2010). The latest data reported that autophagy could lead to sarcopenia and the excessive damage of skeletal muscle mass which occurs in ageing (Wohlgemuth et al, 2010). With advancing age there is a progressive disorder of mitochondrial function with activation of autophagy.

2.6.1 Body weight loss

Body weight loss (BWL) or skeletal muscle atrophy (SMA) are normal and common in elderly individuals. A measure of weight in kilograms divided by height in metres squared is called the body mass index (BMI). If the BMI is less than 22 this associated with a higher mortality risk in both males and females older than 65 years of age. Skeletal muscle mass is harbinger of poor outcome (Thomas et al, 2007). There is a relationship between mortality and BMI (Calle et al, 1990), where for example losing more than ten percent after the fifth decade of age compared to people with stable weight is associated with sixty percent increase in mortality (Thomas, 2005). The Loss of about one or two percent per year in people after the fifth decade is normal because this decline muscle weight happens in both sedentary and

active ageng adults (Tzankoff and Norris, 1978). The side effects of muscle loss are not limited to increase a mortality risk. Also it contributes towards a decline in functional status. For example for women 60–74 years old who lose more than 5% of body mass this leads a doubled risk of disability (Thomas, 2003). Skeletal muscle mass loss is accompanied by muscle fibre loss, and this reduction appears in the type IIa fibre. (Doherty et al, 1993; Brown and Hasser,1996). DNA methylation has also been observed in a few human tissues and other species have been documented and decreases in methylation in numerous groups of aged mouse tissues (Calvanese et al., 2009; Wilson et al, 1987; Zykovich et al, 2014). The three main categories of skeletal muscle loss involve sarcopenia, cachexia and starvation. Sarcopenia was first presented at 1988 by Irwin Rosenberg31as an age-associated with decrease in muscle mass. The acute loss of both fat-free and fat mass as is called cachexia which usually it is accompanies disease, for example immunodeficiency or cancer. Sarcopenia prevalence rates range from 6% to 24% depending on the definition and measure of muscle mass used (Newman et al, 2006; Park et al, 2006).

2.6.2 Mitochondrial

Usually muscle mass loss is coincident with functional impairment. The disuse muscle leads to atrophy. Muscle atrophy is associated with a loss of mitochondria and any changes in mitochondrial morphology lead to mitochondrial dysfunction. There are 2 types of mitochondria in skeletal muscle fibres, inter myofibrillar (IMF) mitochondria and subsarcolemmal (SS) mitochondria. The function of the IMF is to provide energy in order to support of muscle force production, while the function of SS mitochondria is to produce energy for membrane-related events. Physical indolence leads to several changes in mitochondria function (Powers et al, 2012). In skeletal muscles, there is a wide range of mitochondrial-related diseases because of the mutations in nuclear DNA, named

mitochondrial myopathies. Recently, there has been awareness about the autophagy role towards muscle mass control. There are several circumstances which can lead to loss of muscle mass such as indolence, microgravity and several diseases such as diabetes, cardiac problems and cancer. The former circumstances make proteins breakdown and exceeds protein synthesis resulting in muscle atrophy. However, the link between skeletal muscle mass and inflammatory markers is still unclear (Cai et al, 2004; 2005).

The increase in mutation of mitochondrial DNA (mtDNA) and decrease in activity of mitochondrial occurs in elderly tissue. This means that the dysfunction of mitochondrial increases with advancing age, with the quickening of damage of oxidative to macromolecules. Despite the fact that the mutant mtDNA quantity is not high in normal elderly tissue compared to young patients that have mitochondrial diseases, however, it plays an active role in the decline in capacity of oxidative of particular tissues. Evidence from many studies reported that dysfunction of mitochondrial is linked with the elderly. Moreover, in muscles, the increase of mitochondrial dysfunction is possibly significant in the decline of lipid oxidation and related to a rise in the level of insulin resistance with advanced age.

The key features associated with mitochondrial diseases, whether stemming from mutations of nuclear DNA or mtDNA, tissues like the skeletal muscle, heart and brain, are the most affected by defects of mitochondrial. There is evidence that there is a relationship between changes in muscle insulin resistance and a decline in mitochondrial genes expression (Mootha et al,2003; Patti et al, 2003).

2.6.3 Gene expression

Utilizing gene expression in various models of muscle atrophy could lead to obtaining a subset of significant genes that are usually down- or up regulated in atrophying muscle. This subset of genes could regulate the components of muscle. Through the atrophy process there is a need for transcriptional regulation in order to rebuild the damaged muscle components (Bodine et al, 2001) and (Sacheck et al, 2007).

2.7 Age associated loss of skeletal muscle mass

The loss of muscle mass due to ageing is called sarcopenia, an architectural and molecular change that alters muscle quality and manifest functional limitations in skeletal muscle mass.

The loss of muscle mass in healthy humans normally appears after the forth decades of life, with a proportion of approximately less than 2% (Hughes et al., 2002). Walrand et al. (2011) suggested many contributing factors, however the main reasons for sarcopenia remain unclear. Age-related hormonal alterations may lead to sarcopenia, as a decline in growth hormone has been observed in the elderly as well as a decrease in testosterone (Vermeulen et al., 1996). In addition, the response to pro-inflammatory cytokines leads to muscle protein disorder. Both old and young humans seem to have the same resting proportions of muscle protein disorder and synthesis (Volpi et al., 2001) and both old and young muscle responds to exercise. Even if there is an impairment in type II fibres, exercise will consistently help to improve muscle function (Klitgaard et al., 1990).

Karsten and Martin (2014) proposed an approach to determine the effect of age on muscle mass and muscle strength for both males and females. In their investigation, they divided microarrays data into two groups. The Wilcoxon-Mann-Whitney-test was used to identify differences between the groups and the statistically significance threshold was fixed to a p-

value <0.05. Their result showed that there is loss of muscle mass associated with age. However, their approach was based on only one case study (old versus young) and used basic statistical analyses (Russo et al, 2003).

Stephen et al, (2002) analysed microarray data and the aim was to identify the impact of age on gene expression patterns. 40 samples from healthy humans consisting of young and old men and women where each group involves 5 samples, different statistical tests were used such as t-test. Results indicated that fifty genes were identified as differentially expressed (>1.7-fold) in relation to age.

Welle et al. (2003) proposed an approach to obtain significant genes that have an important role in skeletal muscle changes in men and to produce more comprehensive gene expression profiles. Dataset involves healthy young and old men, it divided into two groups, namely 8 young and 8 older men. Statistical analysis include t-test, results show that there are 700 probe sets for which t-tests indicated a difference ($P < 0.01$) in mean expression between two groups of age. In this study.

Paul et al. (2005) presented a method to investigate male samples in various age groups. The main goal was to identify subset genes associated with age for young and old males used to discriminate the two groups (older and young). 14 young (19–25 yr. old) and 14 older (70–80 yr. old) healthy men were participants in this study. All samples were in good health as proved by normal tests at a clinical laboratory. The microarray data used was from the National Centre for Biotechnology Information (NCBI)'s Gene Expression Omnibus; the data can be accessed at www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc GSE1428. The supervised classification method (k-nearest neighbour (KNN)) (Theilhaber et al, 2002) with the feature selection method of (Golub et al, 1999) and a holdout cross-validation (HOCV)

have been used. This study is the first attempt to identify a molecular signature of sarcopenia. There are 45 marker genes identified in this study.

Simon et al. (2007), proposed a system to evaluate whether healthy elderly men were associated with transcriptional profile reflecting mitochondrial. In this study, skeletal muscle from the control of 51 young and old men were distributed into two groups: 25 old and 26 young males and females. Several bioinformatics analyses were applied and results indicated that transcriptome profiles revealed a dramatic enrichment of genes associated with mitochondrial function with age. Melov et al,(2007) analysed microarray which is Skeletal muscle gene expression from healthy human. Data divided into two groups: younger (N = 26) and older (N = 25) adult males and females. Two groups ware compared using gene expression profiling. The aim of this study is assess if transcriptional profile in elderly shows mitochondrial disorder or not. Followed by bioinformatics tests results showed that there are 596 genes are differentially expressed based on a false discovery (FDR) rate cut-off of 0.05 in healthy elderly.

Micah et al. (2011) presented a method to profile miRNA expression patterns in healthy elderly people skeletal muscle with a miRNA array followed by bioinformatics analysis. In this investigation, t-test is used and p-value<0.05 was accepted to monitor changes in miRNA expression between groups. As a result, 75 differential expressions have been detected between the old and young group. The study found that the gene expression is higher in older human skeletal muscle.

In 2013, Sifakis et al. presented an approach to analyse microarray data, the proposed approach is based on statistical methods, functional analyses, and machine learning techniques. The proposed system has been applied on publicly available microarray datasets obtained from healthy men and women. According to the presented framework, a subset of

46 genes that involve a candidate gender-independent aging signature was identified. K-NN classifier (k=1, 6, 10) (Ianakiev and Govindaraju, 2002) was used and it outperformed the Decision Tree (DT) (Breiman et al, 1984) and the Random Forest (RF) algorithms (Breiman, 200). Leave-one-out resampling technique was applied to measure the performance.

The aim of an investigation by Dennis et al. (2007) was to identify differential expression of genes due to aging in human skeletal muscle. In order to achieve this goal, several statistical tests including the Mann-Whitney test, the t-test and the Wilcoxon signed-ranks test have been used. The dataset included 20 healthy subjects (old, n = 10) and (young, n = 10). The results indicated that elderly muscle did not show any important changes in gene expression.

Drummond et al (2011) presented a new method to profile miRNA expression patterns in elderly skeletal muscle using bioinformatics and statistical test such as the t-Test. A microarray data set from 19 young and 17 older humans was tested and according to experimental results the authors recommended that a higher expression of genes is an indicator of disorders in cell functioning and may contribute to reduced muscle mass in elderly individuals.

Melov et al (2007) analysed microarrays for skeletal muscle gene expression from healthy humans. Data was divided into two groups: younger (N = 26) and older (N = 25) adult males and females. The two groups were compared using gene expression profiling. The aim of this study was to assess if transcriptional profiles in the elderly show mitochondrial disorders or not. Bioinformatics test results showed that 596 genes are differentially expressed based on a false discovery rate (FDR) cut-off of 0.05 in the healthy elderly.

In 2014 Gheorghe et al, proposed new method to describe how aging progresses in parallel with muscle loss and function disorder, followed by novel statistical methodology to identify

age-regulated expression. To achieve this goal, six expression profiles data sets from diverse human tissues are used include muscle, brain and kidney, data suggested that there are changes in muscle and kidney tissues at two age-positions, first change was noticed during in fifth decade second change was found in eighth decade. To understand the effect of aging in humans,

Thalacker-Mercer et al. (2009) compared changes in skeletal muscle transcriptome that were caused by unusually high intensity resistance loading (RL) inducing muscle damage in older and young adults. There were 46 older and young males and females studied in the investigation. Results revealed that there were 318 genes expressed at a higher level in older individuals compared with only 87 in younger individuals. It was reported that the number of samples/subjects analysed via microarray did not allow them to statistically test interactions between age and RL. Therefore, they suggested that the results provided a basis for future studies targeting age differences in specific cellular processes that could help to better understand the effect of aging on muscle regenerative function.

Della Gatta et al. (2015) proposed a new method to determine if mitochondrial impairment contributed towards age-related muscle loss. Subjects were aged 18 years and over and were healthy men. A t-test was used to check for differences between age groups. Statistical significance was accepted at $P < 0.05$. Results showed that age did not influence the response of specific mitochondrial transcripts. In 2013, Day et al. provided more evidence that skeletal muscle tissue is more reliable than other tissues for identifying age related genes based on differently expressed genes. They compared between different types of tissues including human blood, brain, kidneys, and skeletal muscle. The main aim was to identify tissue-specific age effects, followed by statistical analyses. Results indicated that skeletal muscle only showed age-dependent methylation associated with tissue-specific expressed genes.

Raue et al. (2007) analysed microarray data in order to identify the impact of age on muscle. Different tests were carried out covering 14 females (8 young and 6 old). In conclusion, the results show that the older females' genes were expressed at a higher level compared to the young adults, which means that they are more susceptible to muscle atrophy than young adults.

Resistance exercise plays an active role in raising the proportion of muscle protein synthesis, according to Chesley et al. (1992). Increased muscle protein synthesis does not occur during resistance exercise but increases after 2-3 hours and may remain elevated for 48 hours after exercise.

In humans, an increase in muscle protein synthesis is normally. this increases between two or three hours and it may stay elevated for two days after exercise (MacDougall et al., 1995; Dreyer et al., 2006), but it does not occur through sporting activities (Durham et al., 2004). After aerobic exercise, a change in gene expression, this temporally change is accompanied with a decline in myostatin and a rise in protolithic (Harber et al., 2010). The exercise-induced muscle breakdown may represent a significant part of muscle regeneration. The muscle impairment may lead to assemblage of abnormal mitochondria followed by a rise in and induction of catabolic pathways (Masiero et al., 2009).

Rivas et al. (2014) hypothesised that microRNA (miRNAs) impairment could lead to reduced muscle plasticity among the elderly. This hypothesis was tested among old and young men after an acute bout of resistance exercise (RE) t-test and one-way ANOVA followed by the Student-Newman-Keuls method were carried out on microarray the Significance was set at ($P < 0.05$).. In conclusion, the study identified a miRNA role in the adaptation of muscle to anabolic stimulation and showed an important disorder in exercise-induced miRNA regulation among the elderly.

Ding et al. (2005) presented a new feature selection system called minimum redundancy — maximum relevance (MRMR), this system idea was based on Naıve Bayes (NB), linear discriminant analysis (LDA), logistic regression (LR) and SVM class prediction methods, leave-one-out cross validation (LOOCV) applied and its accuracy computed. Two expression datasets including leukemia data of (Golub et al, 1999). and the Colon cancer data of (Alon et al, 1999). The main aim of the study was to obtain age-related genes with high accuracy. Results showed that the system was able to produce high accuracy compared to previous studies using the same dataset.

Ruiz et al. (2006) presented a novel gene selection system, the main target of which was to identify a subset of features with high predictive power. The algorithm was called BIRS (best incremental ranked subset). The idea of the system was to select features based on a filter approach and then compare the top list of candidate features using a t-test with coefficient (0.1). The analysis was carried out on 4 microarray gene expression datasets, covering a global cancer map, leukaemia, colon cancer, and lymphoma.

All experiments were conducted using a 10-fold cross-validation technique. Results indicated that the BIRS method was able to obtain relevant subset of features from the original set (0.0018% on average) and compared to others it obtained the same predictive performance.

Table 2.1: comparison between different studies related to impact of age on skeletal muscle mass.

Year	Authors	Case Study	Data type	Test type
2002	Stephen et al	young men vs old men	Healthy	Statistical : t-test
2003	Welle et al.	young men vs old men	Healthy	Statistical : t-test
2005	Paul et al.	young men vs old men	Healthy	Machine learning:SFS
2005	Ding et al	----------	leukaemia data	Naıve Bayes (NB), linear discriminant analysis (LDA)
2006	Ruiz et al	----------	leukaemia, colon cancer, and lymphoma	Statistical : t-test
2007	Melov et al.	young men vs old men	Healthy	Statistical : t-test false discovery rate
2007	Dennis et al	Young vs Old	Healthy	statistical tests including the Mann-Whitney test and t-test
2007	Raue et al	Young vs Old	Healthy	Statistical : t-test
2009	Thalacker-Mercer et al.	Young vs Old	Healthy	Anova and Tukey's honestly significant difference (HSD)
2011	Micah et al.	Young vs Old	Healthy	Statistical : t-test
2013	Sifakis et al	Young vs Old	Healthy	Statistical : t-test and clustering method
2014	Rivas	Young vs Old men	Healthy	t-test and one-way ANOVA followed by the Student -Newman-Keuls method
2014	Gheorghe et al	Young vs Old	Healthy	statistical methodology
2014	Karsten and Martin	old versus young	Healthy	Statistical : Wilcoxon-Mann-Whitney-test p-
2015	Della Gatta et al	Young vs Old	Healthy	Statistical : t-test

2.8 Machine learning

Over recent years, machine learning (ML) has played a significant role in bioinformatics research. The goal of ML is to find a global solution for hidden data and to obtain new knowledge and new ability (Kodratoff et al., 2014; Nyberg et al., 2011; Wang et al., 2009). The rapid advances in data mining (DM) technology can be exploited in different areas, such as in medical diagnosis, information management and other applications. The classification approach is used in various fields including statistics, machine learning, and pattern recognition. Classification methods are typically supervised learning methods.

Classifiers typically use separate instances in two or more classes according to the information that is represented by the training instances (Mitchell et al., 1997). ML training and the job of the classification is to separate the different variables and join each of them to the right class. For example, ML might be used to learn whether the patient belongs to the normal class or abnormal class. The classification rules are generated by the training samples themselves without any additional data it can be used to predict unclassified samples (Yan et al., 2013). In unsupervised learning, where the class for the training instance is not pre-defined, the learning is used to group each instance in the most suitable class; in other words, it is used to create the clusters. The classifier simply process the data presented/fed to it, the into two groups, the evaluation stage is very important for making actual progress in data mining., in order to evaluate the classification performance, the common way is divide data into two sets training instances and testing instances.

Training instances are used to train the learning model, whereas the model accuracy is evaluated using both training and test sets. The training accuracy is evaluated using the training set, and the testing set is used to evaluate the testing accuracy and generalisation ability of the classifier. Also, cross-validation is used to evaluate the accuracy and

generalisation of the classifier. In order to calculate the accuracy of prediction, the original labels of the testing samples are compared against those predicted the trained model. If the predicted label of a sample and its original label are the same, this means that the prediction for this sample is correct, otherwise it will be considered incorrect (Xing and Karp, 2001). Confusion matrix can be used to display the total number of correctly classified samples, i.e. true positives (TP) and true negatives (TN) , as well as the incorrectly classified samples, i.e. false positives (FP) and false negatives (FN), as shown in Table 2.1.

Table 2.2: Confusion matrix for two-class classification problem.

Predict class	**A**	**B**
Real class	True Positive **TP**	False Negative **FN**
	False Positive **FP**	True Negative **TN**

A false negative (FN) means that an instance from class A was incorrectly predicted as a class B instance, while a true positive (TP) means that a class B instance was correctly considered as an instance belonging to class A, The confusion matrix rows represent the real class, while the columns represent the predicted class. Therefore, the error rate is the result of the sum of the number of false positives and false negatives divided by the total number of samples:

Error rate = TP+FN/ (total number of sample) (2.1)

This means that the error rate simply represents the overall classification performance. Therefore a low value error does not mean high classification performance. For example,

the total number of samples is 80, where 10 samples belong to class A and 70 samples belong to class B. Let us assume that the TN = 65 while the TP = 5. This means that the error rate is 10% and 50% of samples are incorrectly classified. To be more accurate in evaluation of the results of classification, evaluation matrix are constructed, as shown below.

True positive rate (TP) = TP/(TP+FN): It measures the proportion of misclassified samples in class A and is called 'sensitivity' or 'recall'.

True negative rate (TN) = TN/(TP+FN): It measures the proportion in class and is also called 'specificity'.

False negative rate (FN) = FN/(TP+FN). It is the proportion of positives which yield negative test outcomes with the test

False positive rate (FP) = FP/(FP+TN). It is the proportion of all negatives that still yield positive test outcomes

2.9 Cross-validation (CV)

This approach is often used to define multiple training and test sets, and calculate the accuracy of a model. It is essential to measure the model predictive performance. In cross-validation, the dataset is divided into two subsets named the 'training set' and the 'testing set'. The aim of the training set is to train the learning model, while the trained model is validated by the test set (Kohavi et al., 1995).

There are three types of cross-validation (CV) methods:

- K-fold cross-validation: this means that data will be partitioned into k disjoint subsets. This process will be repeated k times. Each time, one of the data subsets is used as a testing set,

while the remaining subsets will be used to train the classification model. The error of the classification is calculated in each iteration by feeding the testing subset on the classification model. Many studies have confirmed that 10-fold cross-validation is more suitable than others (Olman et al., 2002).

- Holdout cross-validation: it often designates 75% of data as training and 25% as testing, it is also known as 2-fold cross-validation or simply holdout method
- Leave one out cross validation (LOOCV): this strategy of cross validation is able to provides an almost unbiased estimate of generalization ability of classifier

  It is n fold CV, where n equals the number of samples. It produces a balanced estimate of the classifier's ability. (Cawley et al., 2003)

2.10 Classification

Classification is one major type of learning, the learning process of classification is to find models that can separate instances in the several classes using the information delivered by training instances, by other meaning the classification goal is to obtain the definition of a general category given a set of the negative class and other instances of the positive class. The classification learning process is to identify model which able to split instances into two classes or more. there are many classifiers such as the neural network classifier (ANN). Also, the naive Bayes (NB) classifier is not effective when the number of features is greater than the number of samples (Asyali et al., 2006).

A. Support victor machine

Support victor machine was created for the binary classification problem (Figure 2.1). It is widely used in domain of cancer studies, protein identification and especially in Microarray data (Gunavathi et al, 2014), including in genomics, text categorisation, and speech recognition. The strategy of this model of classification is to find a hyper-plane which maximizes the margin between the two classes. As a result, the generalisation capability of the SVM classifier is improved and over-fitting is prevented. The goal of building the discriminant function is to choose the most suitable training data instances which are referred to as support vectors (Hsu et al., 2002).

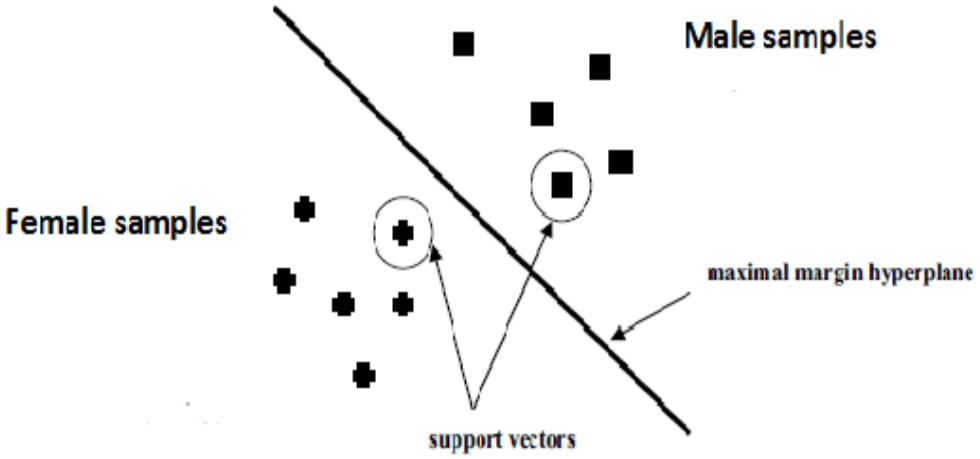

Figure 2.2: Support victor machine, Hsu and Lin,(2002)

Over recent years SVM widely used in bioinformatics and natural language processing (Tessler et al., 2011; Joachims et al., 2002; Scholope et al., 2004). Also, SVM is now widely used in microarray experiments (Yeoh et al., 2002) because it is capable of dealing with large high-dimensional datasets. Brown et al. (2000) determined the performance of three classifiers: support victor machine, Fisher's linear discriminant, and decision tree (DT). Results showed that the SVM achieved the highest accuracy compared with Fisher's linear

discriminate and decision tree. Furey et al. (2000) used 97,802 genes in their experiments and these genes represented cancer datasets. The results showed that SVM achieved the highest performance compared to previous studies that have been conducted on the same datasets. Peng et al. (2003) analyzed four different cancer datasets, including the NCI60 dataset, the GCM dataset, the leukaemia dataset, and the colon dataset. In their study, genetic algorithms (GA) and all paired support vector machines (SVMs) were combined for multiclass cancer identification. The datasets were investigated using 8 and 14 classes. GA (Li et al., 2001; Ooi et al., 2003), recursive feature elimination (RFE) (Ramaswamy et al., 2001), and ranking (Su et al., 2001) approaches were employed in order to identify gene subsets whose expression typifies each cancer class,. Using leave-one-out cross-validations the classification performance was 87.93% and 85.19%, respectively for the NCI60 data set, and 85.19% for the GCM data set. With some problems the SVM classifier has been successfully applied (e.g., handwritten digit recognition). Other studies have applied SVM and another classifier, such as NN or diagonal linear discriminant analysis DLDA, to gene expression in leukemia (ALL/AML) dataset. However, the SVM performance was less than that of both NN and DLDA classifiers (Golub et al., 1999; Chow et al., 2001).

B. K-nearest neighbor (k-NN)

k-nearest neighbor (k-NN) is one of the most common methods of pattern classification (Cover and Hart, 1967). It provides competitive results for example Gunavathi, Premalatha (2014) used GA for effective feature selection. Informative genes are identified based on the T-Statistics, Signal-to-Noise Ratio (SNR) and F-Test values. In this study two classifiers are used including, SVM and K-NN, results revealed that GA-k-NN method achieves 100% accuracy in 4 datasets (Gunavathi and Premalatha, 2014).

The new instance will be classified by a set of similar related instances which are recovered from memory. The majority classes of the training sample will decide the class of the new testing sample. There are many matrices used to measure the distance between two samples and the Euclidean distance is the most common. Therefore the Euclidean distance $d(x_1, x_2)$ between two samples x_1 and x_2 as below:

$$d(x_1, x_2) = \sqrt{\sum_{n=1}^{n}(f_i(x_1) - f_i(x_2)^2)} \qquad (2.1)$$

Where n represents the features, x represents sample and f_i is he value of the i feature k-NN has been commonly used to classify biomedical data such as gene expression data (Xing et al., 2001; Yeoh et al., 2002; Darden et al., 2001; Dudoit et al., 2002). The common advantages of k-NN are that it is computationally simple and less prone to noise. However, it is not scalable and also its performance is reliant on the number of k points used in the classification process (Lu and Han, 2003). There are several other advantages of the k-NN classifier such as:

- Its ability to learn nonlinear relationships among features;
- Its simplicity;
- It considers all features when finding similar training samples from memory. This makes them very sensitive to feature selection.
- Compared to nonparametric techniques such as kernel methods when the data are multidimensional, KNN has been provided better perform (Terrell and Scot, 1992)

2.11 Feature selection

The main aim of using feature subset selection is to remove features that are redundant or not relevant. For example, in cancer studies (cancer-related gene), meaning feature selection technique could lead to identify new knowledge (Guyon et al., 2002). Feature selection is often used to find a distinct subset of features from original data. It is used to improve the

performance of classification (Krishnapuram et al., 2004), to reduce computational complexity and to avoid the over-fitting risk (Abeel et al., 2010). As microarray gene expression datasets consist of tens of thousands of genes, studying these genes or investigating them all at once is considered a cumbersome task. As there is a large number of genes and a small sample size in this type of dataset, using the feature selection method is essential in order to overcome the risk of the over-fitting (Guyon et al., 2002). Since 1970, feature selection has become an important technique and it has proven its ability to exclude redundant and irrelevant features. Recently in several fields there have been many experiments carried out on high dimensional data where the dataset contains hundreds or thousands of features, such as in text categorization (Yang and Pederson, 1997) and genome projects (Xing et al., 2001). Applying feature selection is necessary in order to exclude irrelevant and redundant features. Using feature selection (FS) techniques in bioinformatics has becoming essential where the data includes a large number of genes. Feature selection techniques can be characterized into three categories (Saeys et al., 2008): filter methods, wrapper methods, and embedded methods

A. Filter approach

In this thesis three evaluation methods are used: t-test, entropy, and receiver operating characteristic-area under curve (ROC-AUC).

- **t-test**: a statistical test where the statistic follows a student's t distribution (Thouleimat et al, 2010).It is usually used to evaluate if the averages of two classes are not statistically similar by computing the variability and difference between two classes. It used to detecting differential expression by comparing the mean between samples.
- **Entropy**: it is commonly used in the information theory measure. It considered a measure of the system's unpredictability (Novaković et al, 2016). Entropy represents a measure of

uncertainty of the probability distribution of a random variable x by a variational relationship dI=dx−dx (Wang, 2008) (aharma et al, 2015). Entropy measure is effective for identifying discriminating features (Liu et al, 2004). In this method, the distance between the probability density functions is measured by divergence, which means that the features with higher divergence are considered more suitable for discriminating classes (Sharma et al, 2015)

- **Receiver operating characteristic-area under curve (ROC-AUC):** It is similar to balanced error rate (BER) in that it weights errors differently on the two classes (Chen et al, 2008). This statistical used for determining the efficacy of clinical diagnostic and prediction tests in correctly classifying healthy and diseased individuals (Wray et al, 2010) also it offers an active method to characterize the classifier sensitivity versus specificity. It is drawn between sensitivity and 1-specificity, in other words, the curve is created by plotting the TP rate versus the FP rate at various threshold settings. TP rate is also called sensitivity or recall. The FP rate is also called the fall-out (Fawcett, 2006). In genomic it used to define the efficacy of clinical diagnostic and prognostic tests in correctly classifying control and disease persons (Jakobsdottir et al, 2009; Luet al, 2008; van et al, 2009)

A filter approach, as an alternate type of feature selection technique, has both advantages and disadvantages, as shown below:

Table 2.2: Common types of filter methods used as feature ranking

Fisher Score	supervised, filter, univariate, feature weighting
t-score, F-score	supervised, filter, univariate, feature weighting
Chi-square Score	supervised, filter, univariate, feature weighting
Kruskal Wallis	Kruskal Wallis supervised, filter, univariate, feature weighting
Gini Index	supervised, filter, univariate, feature weighting
Information Gain	supervised, filter, univariate, feature weighting
Fast correlation based filter (FCBF)	supervised, filter, multivariate, feature set
mRmR	supervised, filter, multivariate, feature set
Chi-Square	It is one of the common methods of feature selection based on statistics and filter and it is simple
Relief f	Attribute evaluation, it use Manhattan distance for finding the nearest miss and nearest hit rather than Euclidean distance. And it is easy to use and fast
T-test	is an analysis of two populations means through the use of statistical examination, it is simple and fast

Bioinformatics datasets are highly-dimensional and usually involve thousands of features. In some cases all features are important, but in specific problems only a few subsets of features are important. In such a situation, ranking features is the first step to analysis microarray. In previous studies there have been several feature selection methods used (Chi, 1993; Mántaras, 1991; Setiono and Liu, 1996; Shridhar et al., 1998; Guyon and Elisseff, 2003).

Ming-Chi et al. (2007) suggested a new method to select subsets of genes that had more productive power to check which persons were malignant. To achieve this, breast cancer microarray datasets were used. The strategy of their approach was, based on Markov blanket filtering, a t-test and information gain used to define an information-based measure of correlation

(Zheng et al, 2007). Proposed feature selection approach based on a feature selection repository, it is designed to join the most common algorithms that have been used in the feature selection study to serve as a platform for comparison and joint study, to determine the performance of this platform, 10 benchmark data sets are used such as microarray data, image data and text data. Similarly, Mishra, and. Sahu (2011) presented an approach to identify more reliable genes. Six datasets were used: LEU data (Golub et al., 1999), COL62 data (Alon et al., 1999), BRER49 data (West et al., 2001), LYM77 data (Shipp et al., 2002), PROS102 data (Singh et al., 2002), and LUNG182 data (Mishra et al., 2011). Feature ranking based on the signal-to-noise ratio (SNR) score was used to select the top rank features. The genes were clustered using k-means clustering while SNR is implemented to rank the get top ranked. Next the top scored feature are given to the to KNN and SVM classifiers and validated. Results indicated that the system achieved 99.3% accuracy for both k-NN and SVM classifiers compared to previous works. Xing, and Karp (2001). Proposed an algorithm,

clustering via alternative feature filtering (CLIFF) to select significant subsets of genes, data was a collection of 72 Leukemia patient samples reported in (Golub et al., 1999) the idea of this approach was based on a combination of feature evaluation experts based on independent feature modelling, information gain ranking, then Markov blanket filtering (Koller and Sahami, 1996) to exclude the irrelevant and redundant genes from dataset. Results show that CLIFF is outperforms standard clustering approaches that do not consider the feature selection issue,

Hu et al. (2006) suggested model to select genes with high accuracy. The proposed model was based on different types of feature ranking methods, namely signal-to-noise ratio (SN), cosine (CO) and correlation coefficient (CC) were used The idea was to use a filter approach firstly and then this subset is to filtered out using wrapper feature selection. To select important genes, the misclassification error is estimated using ten-fold cross-validation. Experiments were conducted on cancer datasets covering breast cancer, lymphoma and lung cancer (Van't et al, 2002). Results showed that the model was able to achieve 99.3% classification accuracy based on the k-NN and the SVM classifiers. This result was compared with other approaches that were conducted on a leukemia dataset with LOOCV.

In 2005, Wang et al. presented gene selection method based on filter and cluster. Acute lymphoblastic leukemia (ALL) and acute myeloblastic leukemia (AML) datasets were used in their study. This type of dataset was first introduced by Golub et al. (1999) and has become a benchmark for classification methods which are conducted in relation to leukaemia. Three different, relief-f (Kononenko, 1994), information gain (IG), and $\chi 2$-statistic were used to rank genes. Results indicated that a small dataset usually did not produce high prediction accuracy. Results also showed that there was very good accuracy using relatively small subsets of genes; 100% on an ALL/AML leukemia dataset using 5 genes and an MLL

leukemia dataset using 26 genes, while the system was only able to achieve 91.9% on a colon tumour dataset Golubet al, 1999). I using 3 genes.

In 2004, Yu et al. presented method that can effectively remove redundant genes and select relevant biomarkers. Four microarray datasets were used in their investigation, covering colon, breast cancer, lung cancer and leukemia. This system algorithm representing feature ranking methods based on relief-f, a CFS algorithm denoted by CFS-SF (sequential forward), and a redundancy based filter (RBF). Results indicated that the method was more effective than others found in literature that used the same datasets.

B .Wrapper feature selection

A wrapper feature selection method is an induction algorithm. It selects the best subset of features according to its predictive power using a supervised classifier. Unlike a filter method, it conducts a search for features using the training algorithm itself as part of the evaluation function., wrapper approach The algorithm applied on the dataset which often separated into different sets of features, the feature subset which achieves highest evaluation will be chosen as the final set, then The resulting classifier is then evaluated on an independent test set that was not used during the search (Kohavi and John, 1997; Cadenas and Martínez, 2013). However, a main drawback of the wrapper feature selection methods, such as a genetic algorithm (GA) they have a higher risk of over-fitting than with filter techniques and they are very computationally intensive (George and Raj, 2011).

Wrapper methods have been applied in biomedical data (Blanco et al., 2004; Inza et al., 2004; Jirapech-Umpai and Aitken, 2005; Ruiz et al., 2006). The most common wrapper feature selection methods are sequential forward selection (SFS) (Colak and Isik, 2003) and sequential backward selection (SBS) (Cotter et al., 2001). Table 2.3 shows some properties

of these feature selection methods. The main drawback of both methods is the nesting effect. That is, in SFS when the top-down search is applied the excluded features cannot be re-selected. Likewise, in SBS when the bottom-up search is applied, any features that are selected cannot be excluded. Both of them are suffer from nesting effect. In 1994, Pudil et al. proposed a new way to improve the SFS and SBS methods. The improved SBS method is called sequential backward floating selection (SFBS). while the improved SFS method is called sequential forward floating selection (SFFS). SFFS applying SFS procedure followed by a set of successive conditional to exclude worst feature in the newly updated set provided a further improvement can be made to the previous sets.

. The results show that the SFFS and SBFS are achieved similar results to the SFS and SBS but they are computationally much faster. than both methods. Table 2.3 presents the advantages and disadvantages of the wrapper methods.

Table 2.3: Advantages and advantages of the wrapper methods.

Benefits	**Drawbacks**	**Examples**
Both SBS and SFS Easy Considers the interaction among feature dependencies	Both SBS and SFS Over-fitting is expected nesting effect	Backward elimination (SBS) Sequential forward selection (SFS)

In 2012, Bermejo et al, introduced new approach to reduce the number of wrapper evaluations without degrading the performance of accuracy, filter ranking and wrapper feature subset selection (FSS) are used. To achieve this goal Wilcoxon Matched-Pairs Signed-Ranks Test applied to check if there statistical difference or not and the strategy was

to re-rank the candidate features that already provided by filter ranking then this subset of features will be investigated by wrapper feature subset selection (FSS). This proposed method tested on several dataset. The results show high reduction in the number data dimensionality with achieving high performance of classification for those genes that are the obtained (Rosenberg, I. H. ,1989).

In 2012, Bermejo et al, introduced a new approach to reduce the number of wrapper evaluations without degrading classification accuracy, the proposed approach was based on filter ranking and wrapper feature subset selection (FSS). To achieve this goal, the Wilcoxon matched-pairs signed-ranks test is applied to check for statistical differences and the strategy used was to re-rank the candidate features already provided by filter ranking and then this subset of features would be investigated by wrapper feature subset selection (FSS). The proposed method was tested on several datasets, and the results showed significant reductions in data dimensionality with high classification performance achieved compared to those genes obtained in Rosenberg, (,1989) study.

C . Embedded approach

In this third category of feature selection technique, which perform feature selection in the process of training and are usually specific to given learning machines, the strategy to obtain important subsets of features is similar to the wrapper feature selection method, in that both of them are specific to a given learning algorithm. Embedded feature selection is considered more effective than the wrapper feature selection technique because embedded methods are not divided training data into a training and validation set also they avoided retraining a predictor from scratch for each variable subset investigated which allowed them to reach the solution faster (Guyon and Elisseeff, 2003). (Huang et al., 2007).

Some examples of embedded methods are decision tree learners, such as induction of decision trees ID3 (Quinlan, 1986, 1993, 1996), and the recursive feature elimination (RFE)

approach, which is proposed feature selection algorithm derived from the support vector machine (SVM) theory that has shown high performance in relation to selecting genes using microarray data (Guyonet et al., 2002; Rakotomamonjy, 2003). It provides high performance when it is used to solve certain gene selection problems with cancer microarray data.

2.12 Statistical tests

Statistical tests are usually helps to making decision and to define if there is enough evidence to accept or reject the conjecture about the process. There are several statistical tests that can be used in order to calculate the difference between two groups based on means, including the t-test and z-score, P-value, fold change and ANOVA (Box, 1987). Normally, the data format which is extracted from the microarray experiment is as shown below:

$$\mathbf{I}_n = \underbrace{\begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & \cdots & x_{2,n} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ x_{y,1} & x_{y,2} & \cdots & \cdots & x_{y,n} \end{pmatrix}}_{\mathrm{X}} \Big\} \mathrm{Y} \qquad (2.1)$$

Where X represents samples and Y represents genes. In this research the datasets are characterized such that the number of genes is greater than the number of samples.

A. P-value

It is the probability of finding the observed results when the null hypothesis IS true, it uses to compare two means of groups samples, and for example male versus female, p-value can

be found. To accept or reject a null hypothesis the p-value can be used. The p-value is obtained via a t-test and the purpose is to identify statistically significant differences between the means of two groups. The null hypothesis (no important difference between specified populations) states that the means are equal. Generally, an important P-value level is denoted as α or alpha as show in able 2.4

Table 2.4: Strength of evidence to reject or accept hypothesis according to p-value.

P -value< 0.001	Very strong evidence
P -value < 0.01	Strong evidence
P -value < 0.05	Moderate evidence
P -value < 0.10	Very weak evidence
P -value > 0.10	No evidence

B. Fold change family

Relative indices are assigned to genes according mean rating of the expression levels through various samples of groups (Lazar, et al 2012). Most current biology studies use fold change ratios when they mention fold change.. There are two main types of fold change, namely the fold change ratio and the fold change difference. Both of them are calculated according to the gene expression levels mean through various groups of patients per gene. typically provide the log2 of the ratio between the means of group A and group B, the result of applying fold change should be either positive or negative. The purpose is to show the change of expression, both up or down, based on the mean, when positive means up and negative means down (Huang and Xu, 2007).

C. ANOVA

Analysis of variance (ANOVA), is a set of statistical tools for hypothesis testing or statistical importance testing of the difference in means and divergence of two groups. There are two types of ANOVA, as mentioned below:

- **One-way ANOVA**

One-way ANOVA is normally used to compare the means of 3 or more patient groups. It is used to determine the divergence between sample groups and within samples.

- **Two-way ANOVA**

Two-way ANOVA it used to compares the mean differences between groups which are split on two variables. It is an extension of one-way ANOVA (Doncaster and Davey, 2007).

2.12 Conclusion

The above section provides background about data mining and bioinformatics techniques, and also it reviewed some techniques such as feature selection and classification. In addition it reviewed several example about how the previous studies that were used these techniques. Moreover this section describes microarray dataset. Skeletal muscle atrophy is described and reviewed several studies that were concerned in sarcopenia filed. Most of previous systems were based on the statistical methods such as t-test p-value and fold change however they are not made a combination between statistical methods and sophisticated machine learning techniques such as feature selection methods, moreover most of them are used whole dataset to identify distinct subset of genes which leads to poor generalization, finally most of previous studies are related to cancer data. Next chapter will describe the proposed system multi filter system MFS which chosen to identify atrophy-related genes which have important role toward muscle decline in both males and females.

Chapter 3

Majority voting approach to understand gender-related skeletal muscle atrophy

3.1 Introduction

Sexual dimorphism in skeletal muscle can occur due to age (Welle et al, 2008). Many of these age-related changes in skeletal muscle appear to be influenced by gender (Liu et al, 2013), (Yunet al, 2015; Kalyani et al, 2014) For example, the muscle mass of men is larger than that of women, especially for type II fibers, while the proportion of type I muscle fibers is higher in women (Liu et al, 2010). Welle et al. reported that the muscle mass of men is larger than that of women (Welle et al, 2008; Visser et al, 2003; Hughes et al, 2002), due to higher levels of testosterone and the anabolic effect of testosterone is well known. However, previous studies have failed to identify which genes are responsible for anabolic effects. The molecular biases related to gender difference are still unclear (Welle et al, 2008), but 50% of the cell mass of the human body is muscle and so skeletal muscle is considered an important tissue. There are several changes in skeletal muscle related to age that seem to be influenced by gender (Sifakis et al, 2013). Changes in gene expression could be responsible for the decline in muscle function (Roth et al, 2002). In fact, microarray data involves a large number of genes but only few of them may help to interpret the impact of age in skeletal muscle mass (Liu et al 2010) For example there are few studies that investigated datasets which included

men and women gene expression in various ages (Roth et al, 2002). Janssen et al (2000) reported that the reduction in skeletal muscle (SM) mass related to age starts in the third decade of life. This decrease starts to appear in the lower body SM. To find differences between men and women, they used t-test, Pearson correlation and multiple regression based analysis to determine the relationship between age and skeletal muscle. Liu et al (2013) used basic statistical analysis to conduct a comparison between males and females in each age group using gene expression profiles from skeletal muscle tissue. They identified important gender and age related gene functional groups using intensity-based Bayesian moderated t-test and logistic regression. This was the first study that offered proof of the occurrence of extensive gender differences in the aging process of human skeletal muscle. Although most of previous studies showed interesting results based on gene expression dataset, but they had used genes belonging to X and Y chromosomes, which can easily discriminate genders. Experiments were conducted using three groups, namely old women versus old men, young women versus old women, and young men versus old men. However the main problem in these studies is that important genes are identified using the whole training data. This can lead to poor generalization because one of the fundamental goals of machine learning is to generalize beyond samples in the training data. The main goal of the proposed study is to extend the work reported by Liu et al, (2013) and to identify important genes with good generalization ability. The goal of proposed multi-filter system MFS is to find age-related genes based on three evaluation methods that can used as sarcopenia biomarkers in human.

The proposed approach is inspired by the ensemble of feature ranking methods for data intensive applications (Haury, et al, 2011), where genes are first sorted using three different evaluation methods using t-test, Wilcoxon and the ROC-AUC. Later, important genes are determined using majority voting based on the principle that combining multiple models can

improve the generalization of the system. The scope of this study is the selection of the most reliable genes and the evaluation of the classification power of selected genes. Experiments were conducted on microarray gene expression dataset and the results have indicated a significant increase up to 10% in classification accuracy when compared with the genes obtained by Liu et al (2013) system. In this thesis the proposed technique applied on two datasets and our system is able to identify differentially-expressed genes for the following three case studies in relation to age and gender differences

- Young Women versus Old Women
- Young Men versus Old Men
- Old Men versus Old Women

Next section will describe the microarray datasets also will explain how the proposed system is working and what are the machine learning technique that are used in.

3.2. Material and proposed method

3.2.1. Micro array gene expression data set

In this study, two microarray datasets of gene expression of skeletal muscle are used. Datasets are publicly available in the Gene Expression Omnibus (GEO) database. A total of 58 individuals were involved in this investigation, and 22 healthy males and females of various ages, seven of the males and seven of the females were young (20-29 years old), and 4 males and 4 females were old (61-81 years old), were included in the first study A, in study A we used microarray dataset (Liu et al, 2013) and this dataset is divided into three case of study as shown in the above section. Whole Ribonucleic Acid (RNA) was extracted and gene expression profiling was implemented utilizing the Affymetrix human genome U133 Plus

2chip. As in (Liu et al, 2012), this data set is divided into three cases. The first case involved 11 females (7 young and 4 old), the second case consisted of 11 males (7 young and 4 old) and the last case contained 8 samples (4 old men and 4 old women). In study B, gene subset selection was conducted using feature ranking techniques. Bioinformatics data have extremely high dimensionality, and around 55,000 genes from only 36 samples from 15 young people (7 men, 8 women) and 21 older people(10 men and 11 women) were included in study B.

3.2.2 Gene subset selection using feature ranking techniques

The datasets used in this study have extremely high dimensionality. The first dataset consists of around 55,000 genes from only 22 samples, while the second dataset involves 55,000 genes from 36 samples. This is considered a significant challenge to machine learning methods, as there are a large number of features than samples. To address this problem, it is important to select a small subset of relevant features to reduce processing time and avoid the over-fitting problem (Saeys, et al, 2007). One possible solution is feature selection using feature ranking methods. In this study, three different evaluation methods are used for feature ranking.

3.2.3 **Classification**

The selected subset of genes is used for testing the generalization ability of supervised classifiers. The k-nearest neighbor (K-NN) classifier (k=1,3) and SVM are used to evaluate the systems performance. The leave-one-out cross validation (LOOCV) technique is used for evaluation.

A. k-NN classifier: The main objective of the K-NN classifier is to discover a set of k objects in the training set that are similar to the objects in the test group.

B. SVM classifier: The support vector machine was introduced by Vapnik in 1998 to address several data mining problems. It uses a suitable hyper-plane to identify classes

3.2.4. Proposed System

In order to decrease the size of dataset and TO re-rank the genes we used different evaluation methods such as t-test, entropy, ROC, relief and f-test, the results shows that each on of them has yield a different new list of genes due to each one of them has its own criterion therefore we tried so many combination between them different times for example we chose three evaluation methods, four evaluation methods and all of them together in order to adopt the best combination. The results indicated that using combination between t-test, entropy and ROC achieves best performance of accuracy. Therefore the framework of the proposed system MFS was designed based on the majority voting of the combination of three different evaluation methods including t-test, entropy and ROC. The aim was to identify subset of age-related genes that interpret the impact of age on skeletal muscle mass. The data set is first divided using leave-one-out-cross validation into T folds. In other words, there are 20 folds for 20 samples where each fold consists of 19 samples for training and one sample for testing. For each fold a multi-filter system (MFS) is applied based on three different evaluation methods: the t-test, the ROC-AUC and the entropy as shown in figure 3.1. Each of them is responsible for sorting genes according to specific criterion. From these sorted genes, a unique subset of genes are obtained based on majority voting, as shown in Table 3.1. Assume that there is a total of 10 genes and the objective is to select the first top 5 genes. Genes 9 and 10 are selected by different evaluation methods and so these are considered the most important genes. Genes 1, 4, 5 are selected twice and thus are also considered by the system to be important genes. It should be noted that, due to the majority voting genes 2, 3 and 6 are not selected by the system. Later, k-NN is applied to the new subset of genes in order to check the predictive performance. Leave one out cross validation LOOCV, is applied on two different

microarray datasets and the classification repeated 100 times, firs time based on the first top rank gene that appear in the majority voting list, in the second time LOOCV applied again and the classification is applied based on the 2 genes the first and second top rank genes and so on until last time when the classification is applied based on 100 gene. For example if we have 10 samples this means that each sample will be classified 100 times because in the first time of classification just first top rank gene will be used and the number of genes incrementally increase one by one until the gene number 100 in the top rank gene list and in each time of classification the selected genes were counted to know which genes have high score of repetition then the gene s sorted based on their repetition score.

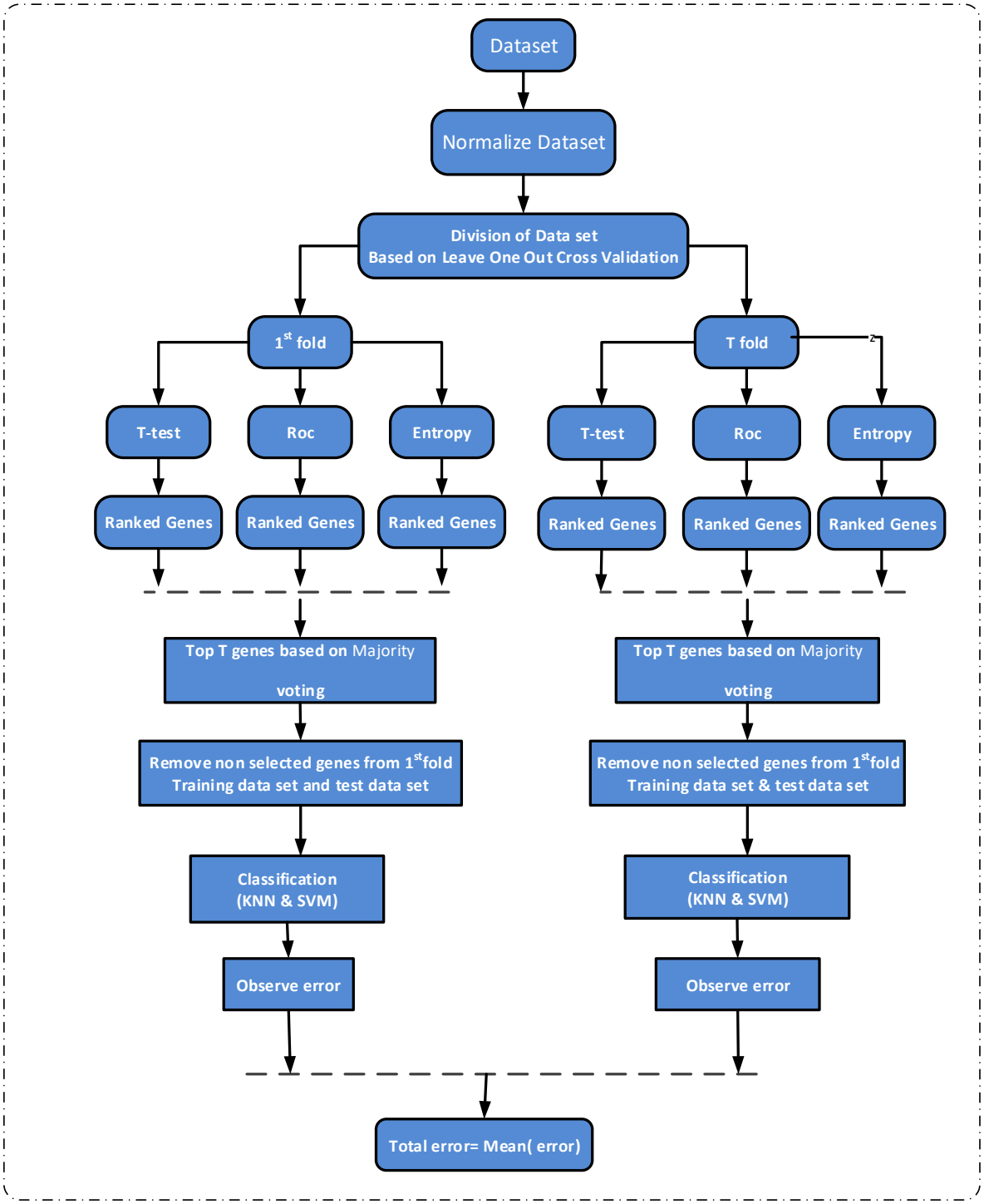

Figure 3.1. Proposed Multi-Filter System (MFS).

Table 3.1 Selected genes using majority voting.

Evaluation method	**Top Ranked Genes**
t-test	1, 4, 6, 9, 10
ROC-AUC	1, 2, 5, 9, 10
Entropy	3, 4, 5, 9, 10
Majority Voting	1, 4, 5, 9, 10

The limitation of using majority voting is if there is no common genes between all evaluation method as shown in table in this case we adopt t-test rank because it was yield the best performance of classification.

Table 3.2 Selected genes using majority voting.

Evaluation method	**Top Ranked Genes**
t-test	1, 4, 6, 17, 100
ROC-AUC	8, 2, 50, 91, 90
Entropy	3, 6, 44, 66, 140
Majority Voting	0, 0 0 0 0

Next section will show the results of multi-filter system experiments.

3.3 Results and discussion

This section, evaluates the performance of the multi-filter system (MFS). The proposed system is also compared with the system presented in previous studies (Liu et al, 2013; Raue et al, 2012) in which genes were identified for three categories (male young versus male old, female young versus female old and male old versus female old) from a total of 54623 genes. In order to make a fair comparison, the same number of genes are selected from the proposed MFS and compared with the genes identified in the earlier researches (Liu et al, 2013; Raue

et al, 2012). The evaluation matrix used in this study are: classification accuracy, sensitivity and specificity.

3.3.1 Study A

In the present study the dataset is a microarray dataset which is publicly available at the Gene Expression Omnibus (GEO) dataset NCBI (2013).

The subjects consist of 22 healthy males and females of various ages. There are 11 males and 11 females which are distributed as follows:

- Males: 7 young (between 20 and 29 years old) and 4 old (between 61 and 81years old)
- Females: 7 young (between 20 and 29 years old) and 4 old (between 61 and 81years old)

The first columns represents the genes names, first row represents the samples while the gene expression values are represented by numbers in dataset as shown in table 3.3. The big variation between genes shows that there is a significant down regulated > 1.5 fold change which lead to mitochondrial dysfunction. In addition applying p-value between old /young shows significant difference $p < 0.05$ between these two groups of samples based on the mean of genes.

Table 3.3 dataset

Gene. symbol	20 – 29 yrs old males	20 – 29 yrs old males	20 – 29 yrs old males	20 – 29 yrs old males	61 - 81yrs old males	61 - 81yrs old males	61 - 81yrs old males	61 - 81yrs old males
RFC2	110	57.518	86.684	86.121	323.658	706.342	650.904	715.802
HSPA6	91.422	72.174	75.344	87.011	109.304	511.561	632.571	809.395
PAX8	90.21	27.9153	43.1126	40.8638	52.6655	627.0215	412.4168	628.0461
GUCA1	93.848	85.141	53.845	79.454	380.377	569.465	283.207	849.21
UBA7	77.0766	55.746	76.8394	59.7029	498.745	400.7278	901.4931	746.6841
CYP2A	72.65	79.886	66.829	78.346	548.846	577.059	525.034	662.238
SCARB	49.6456	55.3403	64.4303	68.6059	527.034	732.8631	620.2522	746.3319
TTLL12	60.715	87.967	96.695	92.238	331.495	506.743	901.2891	616.323

A. Case study 1: young males versus old males

This case study consists of 11 male samples (7 young and 4 old). In this case the true positives represent the correctly classified young males, true negatives represent correctly classified old males , false positive represent the incorrectly classified young males and false negatives represent incorrectly classified old males. Table 3.5 shows the performance of the MFS compared with the genes identified by Liu et al, (2013). It is shown that the best performance is obtained using the 3-NN classifier which achieves 90.9% while the approach by Liu et al (2013) was only able to achieve 81.8%. This improvement is mainly due to higher specificity. Further analysis has revealed that, out of 75 genes, only 9 genes are common in both systems.

Table 3.4 Confusion matrix: young men versus old men

6	0
1	4

Table 3.5 Classification performance: Young males versus old males.

classifier	**Accuracy**		**Sensitivity**		**Specificity**	
	Proposed MFS	**Liu et al (2013)**	**Proposed MFS**	**Liu et al (2013)**	**Proposed MFS**	**Liu et al (2013)**
KNN(K=1)	**81%**	**63%**	**0.714**	**57%**	**100%**	**75%**
KNN(K=3)	**90%**	**81%**	**85%**	**85%**	**100%**	**75%**
SVM	**72%**	**63%**	**75%**	**57%**	**75%**	**50%**

New genes selected by the proposed system. Common genes selected by proposed system. Based on figure 3.2, it is observed that the best classification accuracy was obtained by first 5 genes approximately 100% due to the smaller number genes in this comparison and afterwards there is a 10% drop in performance. This increase of accuracy due to the selection of some new genes that can degrade the performance of the system.

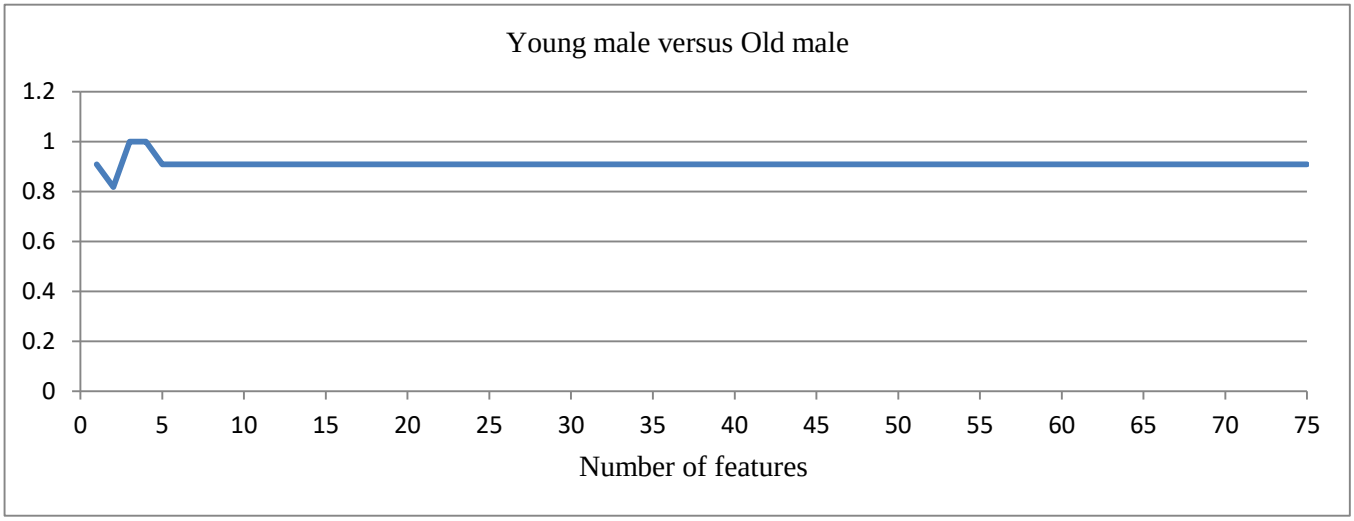

Figure. 3.2. Performance of the MFS using varying numbers of genes compared with Liu et al, (2013)

Some new genes are identified, that have an important role in interpreting age differences of males and females. Some of the new genes are shown in Table 3.6 along with the 9 genes selected by both systems.

Table 3.6: New genes selected by the proposed system. Common genes selected by Proposed system and system by Liu et al, (2013)

New genes selected by MFS Common Genes	New genes selected by MFS different Genes
• Caveolin 3 (CAV3) • Eukaryotic translation elongation (EEF1B2) • FBR-MuSV ubiquitously expressed (FAU) • RNA binding motif protein 15 (RBM15) • Ribosomal protein L4 (RPL4) • Cytochrome c-1 (CYC1) • Mitochondrial ribosomal protein S30 (MRPS30) • Pyruvate dehydrogenase kinase, isozyme 2 (PDK2) • Phosphoglycerate mutase 2 muscle (PGAM2)	• Toll-like receptor 4 (TLR4) • UDP-GlcNAc:betaGal (B3GNT6) • TGF-beta activated kinase 1 (TAB3) • Myozenin 3 (MYOZ3) • Olfactory Receptor (OR5P3) • Thioesterase superfamily member 4 (THEM4) • RAN binding protein 3-like (RANBP3L) • Fc receptor-like 3 (FCRL3) • Rhomboid, veinlet-like 3 Drosophila (RHBDL3)

These genes functional groups reflected three overriding biological themes including, lipid synthesis such as Caveolin 3 (CAV3), Gene transcription and translation such as FBR-MuSV ubiquitously expressed (FAU), RNA binding motif protein 15 (RBM15) and storage also showed a female-specific transcriptional up-regulation with aging while mitochondrial function such as CYC3, Phosphoglycerate mutase 2 muscle (PGAM2) and Mitochondrial ribosomal protein S30 (MRPS30) are down-regulated. This indicate to presence of sarcopenia in old males.

B. Case study 2: old males versus old females

Another objective of this investigation was to examine basal level gene expression among old men and old women. In this case the true positives represent the correctly classified old males, true negative represent correctly classified old females, false positives represent the incorrectly classified old males and false negatives represents incorrectly classified old females. This case study consists of 8 adults (4 old men versus 4 old women). Positive class represented by old men and negative class represented by old women. Table 3.8 shows the performance of the MFS when compared with the genes identified by Liu et al, 2013. It is found that genes selected using MFS have a classification accuracy of 100% using both 1NN and 3NN with high sensitivity and specificity.

Table 3.7 Confusion matrix : old male versus old female

4	0
0	4

Table 3.8 Classification performance: Old men versus Old women.

classifier	**Accuracy**		**Sensitivity**		**Specificity**	
	Proposed MFS	Liu et al (2013)	Proposed MFS	Liu et al (2013)	Proposed MFS	Liu et al (2013)
KNN(K=1)	100%	75%	100%	50%	100%	100%
KNN(K=3)	100%	75%	100%	50%	100%	100%
SVM	75%	62%	75%	50%	75%	75%

According to the figure 3.4 the first gene achieve 40% of classification and then there is a dramatic increase until gene number 5 where the best performance was achieved stating which approximately 100% and the performance accuracy is stable on same level for all remain genes this due to smaller number of genes.

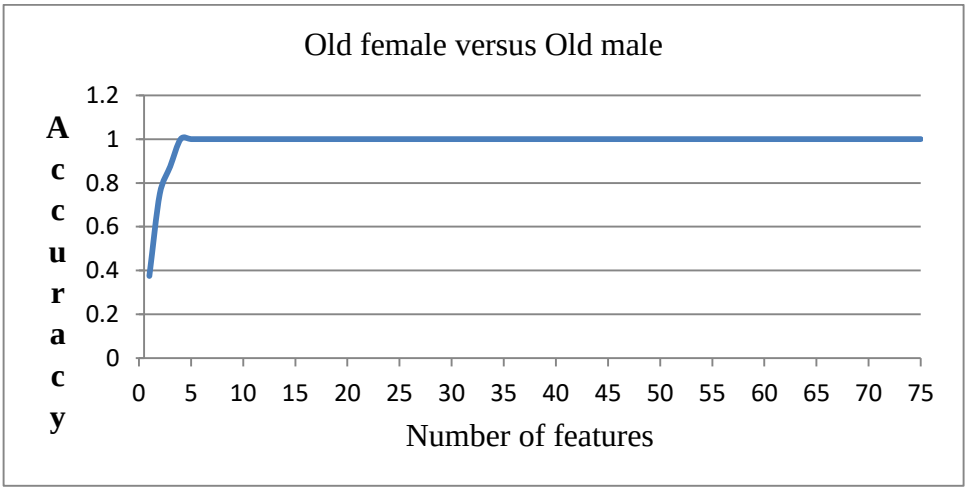

Figure 3.3. Performance of the MFS system using varying number of genes.(old males versus old females)

Table 3.9 show some of the common and different genes that obtained by MFS and Liu et al, (2013)

Different genes	Common genes
activating transcription factor 3 (ATF3)	Peroxisome proliferator-activated receptor gamma (PPARG)
solute carrier family 46 (folate transporter), member 1 (SLC46A1)	Glycerol-3-phosphate acyltransferase, mitochondrial (GPAM)
Poly (ADP-ribose) polymerase family, member 15 (PARP15)	Stearoyl-CoA desaturase (delta-9-desaturase) (SCD)

The above table 3.9 show some of the common and different genes that obtained by MFS. All of genes were in low P-values $P < 0.05$ and fold change was more than 2 fold in addition the pathway of some of these genes reflected mitochondrial function, in old female genes were down-regulated with about 4 fold change compared to young. This refer to muscle atrophy in women

C. Case study 3: young females versus old females

This case study consists of 11 female samples (7 young and 4 old). In this case the true positives represent the correctly classified young females, true negatives represent correctly classified old females , false positives represents the incorrectly classified young females and false negatives represent incorrectly classified old females. Table 3.11 shows the performance of the MFS when compared with the genes identified by Liu et al (20012). Again, the best performance is obtained using the 1NN classifier at 91%. Meanwhile, the genes identified by Liu et al. (2012) are only able to achieve 72.2% accuracy which indicates

the important improved generalization ability of the proposed system. We argue that the improvement in performance is mainly due to high specificity as sensitivity which is same in the both systems.

Table 3.10 Confusion matrix : old male versus old female

6	0
1	4

Table 3.11 Classification performance: Young female versus old female

classifier	**Accuracy**		**Sensitivity**		**Specificity**	
	Proposed MFS	Liu et al (2012)	Proposed MFS	Liu et al (2012)	Proposed MFS	Liu et al (2012)
KNN(K=1)	91%	45%	85%	57%	100%	25%
KNN(K=3)	72%	72%	85%	85%	50%	50%
SVM	70%	63%	71%	75%	75%	50%

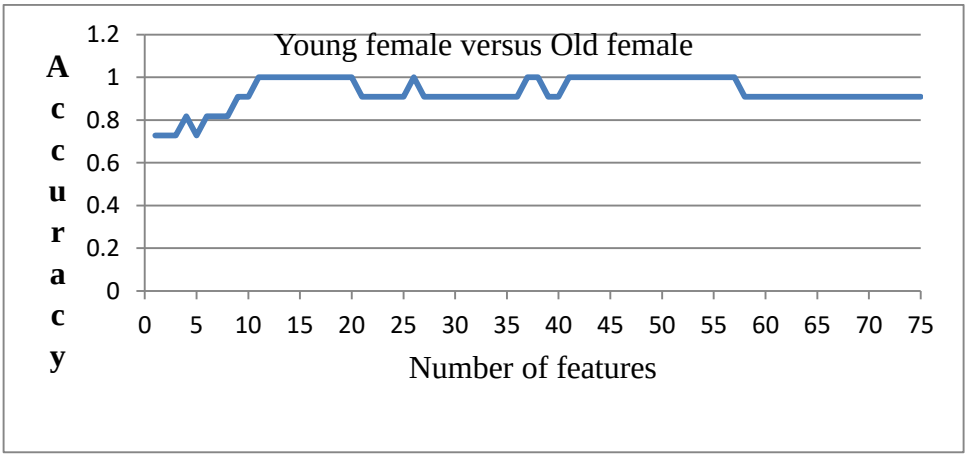

Figure. 3.4. Performance of the MFS using varying number of genes.(young females versus old females)

Figure 3.4 indecates that the performance of classification accuracy was between 75% and 100% for the first 10 genes after that the best performance has achieved which approximately 100% between gene number 10 and 20, then the performance fluctuated between gene number 20 and 40 due to there are some new genes are selected which affected on the classification accuracy, while after the gene number 43 the performance is increased and fixed for 100% between gene number 40 and 60due to some genes are selected finally the slop drop by 10%

Table 3.12 lists the names of some of the new genes identified by the proposed system along with some common genes selected by both systems MFS and Liu et al, 2013 these genes are sarcopenia-related genes due p-values were less than 0.05 and pathway reflected mitochondrial function such as UCP3 it was down-regulated in older women.

Table 3.12 Some of the common and different genes that obtained by MFS and Liu et al, (2013)

Different genes	**Common genes**
PTEN induced putative kinase 1 (PINK1)	Stearoyl-CoA desaturase (delta-9-desaturase) (SCD)
Mitogen-activated protein kinase 1 (MAPK1)	Phosphorylase kinase, gamma 1 (muscle) (PHKG1)
microfibrillar-associated protein 3 (MFAP3)	Phosphoglycerate mutase 2 (muscle) (PGAM2)
TGF-beta activated kinase 1/MAP3K7 binding protein 3 (TAB3)	Uncoupling protein 3 (mitochondrial, proton carrier) (UCP3)

3.3.2 Study B

This second dataset involves 54,623 genes, which is publicly available at the Gene Expression Omnibus (GEO) dataset NCBI (2012). The subjects consist of 36 healthy males and females in various ages, 19 females (8 young, 11 old) and 17 males (7 young, 10 old), the young (24±1y) The old (84±1y). Total RNA was extracted and gene expression profiling was performed using the Affymetrix Human Genome U133 plus 2 chip.

A. Case study 1: young male versus old male

This case study involves 17 male samples (7 young and 10 old). In this case the TP represent the correctly classified old males, TN represent correctly classified old males, FP represent the incorrectly classified old males and FN represents incorrectly classified old males. As shown Table 3.13 the best performance is obtained using 1-NN and 3-NN classifiers which achieved 88% whereas the genes obtained by Raue et al, (2012) are only able to achieve 82%. There are 39 genes which are commonly identified by both systems. Some new genes, identified by the proposed method.

Table 3.13 Classification performance young male versus old male

classifier	Accuracy		Sensitivity		Specificity	
	Proposed MFS	**Raue et al (2012)**	**Proposed MFS**	**Raue et al, (2012)**	**Proposed MFS**	**Raue et al (2012)**
KNN(K=1)	**88%**	82%	85%	71%	90%	90%
KNN(K=3)	**88%**	82%	85%	71%	90%	90%
SVM	82%	80%	80%	70%	80%	80%

Table 3.14 Confusion matrix : young men versus old men

6	1
1	9

Figure 3.5 shows that performance of classification accuracy was fluctuated after between 90% and 88% for all genes this due to include/exclude some genes to the subset of genes

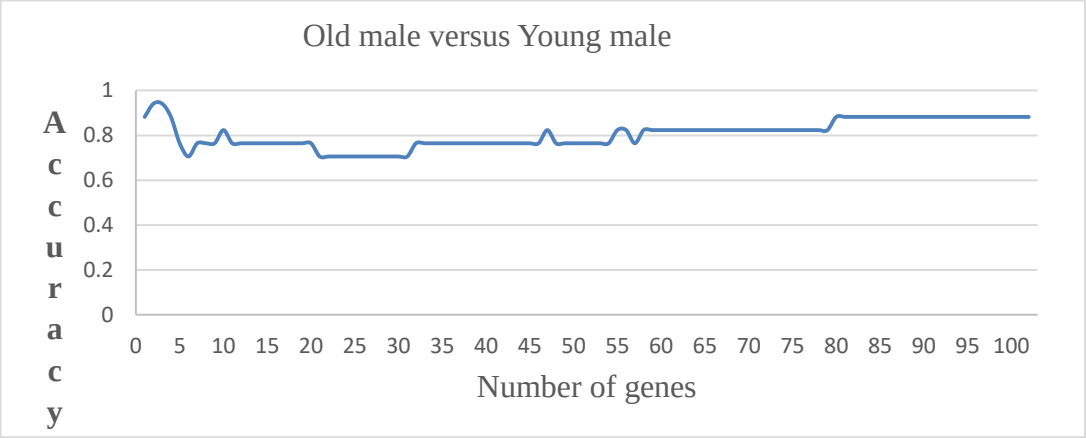

Figure 3.5. Performance of the system using varying number of genes. (old males versus young males)

Table 3.15 List of genes identified by the proposed MFS method and those common to both the proposed method and the method by Raue et al, (2012)

Different genes	**Common genes**
Mannosyl (alpha-1,3-)-glycoprotein	ATPase, Na+/K+ transporting, beta 4
DnaJ (Hsp40) homolog, subfamily C,	Cyclin-dependent kinase inhibitor 1A
	Chloride intracellular channel 5 (CLIC5)
Attractin-like 1 (ATRNL1)	ATPase, H+ transporting, lysosomal
Tripartite motif containing 40 (TRIM40)	Doublesex and mab-3 related
Zinc finger protein 791 (ZNF791)	Rythrocyte membrane protein band 4.1-
	Family with sequence similarity 171,

Table 3.15 show common and different genes between proposed system MFS and Raue et, (2012)All of these genes are consider significant due to the p-value < 0.05 and fold change greater than 2 fold. In addition to provide insight about genes, David software which is normally serve as functional annotation tool is used for both common and different genes. Results revealed that the pathway of some genes were differentially expressed down-regulated such as DMRT2 and others were up-regulated, for example cyclin-dependent kinase inhibitor 1A (p21, Cip1)-CDKN1A, these genes were reflected three biological themes including mitochondrial function, immune function and transcription, it refer to the mitochondrial dysfunction which leads to muscle atrophy in older men.

B. Case study 2: young female versus old female

This case study involves 19 male samples (8 young and 11 old). In this case the true positives represent the correctly classified young females, true negatives correctly classified old females , false positives represent the incorrectly classified young females and false negatives represents incorrectly classified in old females. Table 3.16 shows a comparison of the performance of the proposed MFS (using the second dataset) with the genes identified by Raue et al, (2012) The best performance is obtained using the 1-NN and 3-NN classifier at 100%, while the genes obtained by Raue et al, (2012)were only able to achieve 81.8%. Further analysis has shown that, out of 102 identified genes, only 8 genes are common in both systems. Some new genes are identified, that might play an important role in age differences between young and old men. Table 3.18 lists the names of some new genes identified by the proposed system along with some common genes selected by both systems

Table 3.16 Classification performance: young female versus old female

Classifier	Accuracy		Sensitivity		Specificity	
	Proposed MFS	**Raue et al (2012)**	**Proposed MFS**	**Raue et al (2012)**	**Proposed MFS**	**Raue et al, (2012)**
KNN(K=1)	**100%**	89%	100%	87%	100%	90%
KNN(K=3)	**100%**	84%	100%	62%	100%	100%
SVM	89%	89%	90%	90%	90%	90%

Table 3.17 Confusion matrix: young female versus old female

8	0
0	11

According to the figure 3.6 the performance of classification accuracy fluctuated for the first 56 genes between 50% and 95%, while the best performance of classification accuracy was achieved after the gene number 57 with approximately 100% due to there are some new genes.

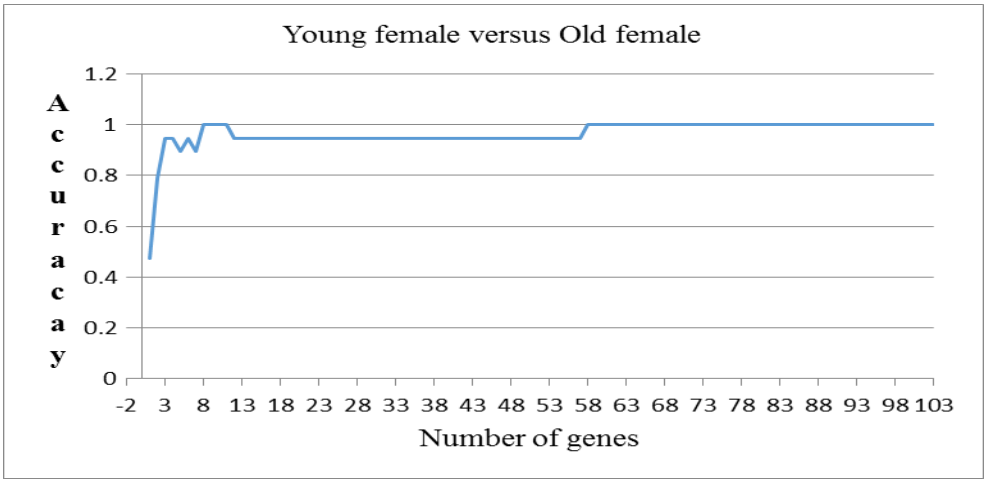

Figure. 3.6. Performance of the system using varying number of genes (young female old female).

Table 3.18. List of common and different genes between Raue et al, (2012), and the proposed MFS

Common genes	**Different genes**
complement component 1, q subcomponent, B chain (C1QB	mitogen-activated protein kinase 1(MAPK1)
damage-specific DNA binding protein 2, 48 kDa (DDB2)	lumican (LUM)
fasciculation and elongation protein zeta 2 (zygin II) (FEZ2)	myosin light chain kinase family, member 4 (MYLK4)
Kruppel-like factor 5 (intestinal) (KLF5)	mitogen-activated protein kinase kinase 1 (MAP2K1)
pleiomorphic adenoma gene 1(PLAG1)	BC016339
src kinase associated phosphoprotein 2 (SKAP2)	thioredoxin reductase 2 (TXNRD2)
Common genes	**Different genes**
zinc finger protein 385B(ZNF385B)	hydroxyprostaglandin dehydrogenase 15-(NAD) (HPGD)
pyruvate dehydrogenase kinase, isozyme 4(PDK4)	RNA binding motif protein 25 (RBM25)

The above genes were observed that have low p-value and high fold change more than 1.5 and some of them are down-regulated such as DDB2 and PDK4 in old women moreover in

term of biological interpretation based on DAVID software the pathway of these tow genes reflected mitochondrial function which refer to the sarcopenia in both male and female.

C. Case study 3: old males versus old females

This case study involves 21 male samples from 10 men and 11 women). In this case the true positive represented by the correctly classified in old males, true negative represented by correctly classified in old females, false positive represents the incorrectly classified in old males and false negative represents incorrectly classified in old females Table 3.19 shows a comparison of the performance of the MFS (using the second dataset) with the genes identified by Raue et al, (2012). Based on Table 3.11, the best performance is achieved using the 1-NN classifier at approximately 95% while the genes obtained by Raue et al were only able to achieve 76%. Only 9 genes are common in both systems. Some new genes are identified, these contribute to interpret the age differences between old men and old women. Table 3.21 lists the names of some new genes identified by the proposed system along with some common genes selected by both systems.

Table 3.19 Classification performance: old male versus old female

classifier	Accuracy		Sensitivity		Specificity	
	Proposed MFS	Raue et al (2012)	Proposed MFS	Raue et al (2012)	Proposed MFS	Raue et al (2012)
KNN(K=1)	95%	61%	100%	63%	90%	60%
KNN(K=3)	80%	76%	71%	81%	90%	70%
SVM	75%	71%	75%	70%	80%	75%

Table 3.20 Confusion matrix : old male versus old female

10	1
0	10

Based on the figure 3.7 it is clearly that after gene number 5 the performance of classification accuracy was stable for 95%. While the lowest performance was between gene number 1 and 4which was less than 95 %. Which due to some of them are effect on the performance

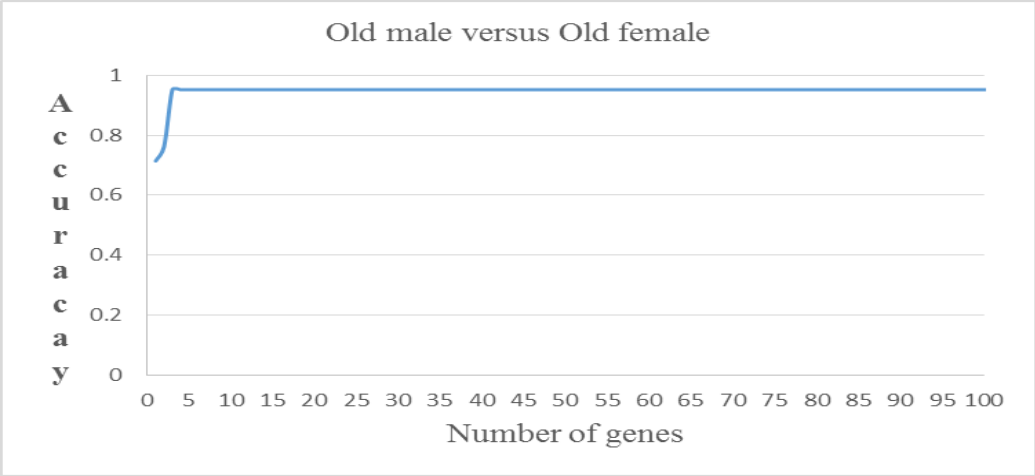

Figure 3.7 performance of the system using varying number of genes (old male versus old female)

Table 3.21 list of common and different genes between Raue et al and the proposed MFS

Different genes	**Common genes**
Protocadherin-related 15 (PCDH15)	Cyclin G2 (CCNG2)
Cytochrome c oxidase subunit IV isoform 1 (COX4I1)	Collagen, type IV, alpha 6 (COL4A6)
DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, Y-linked (DDX3Y)	Ffibronectin leucine rich transmembrane protein 2 (FLRT2)

Ribosomal protein S4, Y-linked 1 (RPS4Y1)	Folate receptor 2 (fetal) (FOLR2)
Eukaryotic translation initiation factor 1A, Y-linked (EIF1AY)	Membrane-spanning 4-domains, subfamily A, member 4 (MS4A4A)
DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, Y-linked (DDX3Y)	Micotinamide nucleotide adenylyltransferase 1 (NMNAT1)
Ubiquitin specific peptidase 9, Y-linked (USP9Y)	Secretory leukocyte peptidase inhibitor (SLPI)
lysine (K)-specific demethylase 5D (KDM5D)	Tubulin polymerization-promoting protein family member 3 (TPPP3)
ATPase, H+ transporting, lysosomal 16kDa, V0 subunit c(ATP6V0C)	Pyruvate dehydrogenase kinase, isozyme 4 (PDK4)

3.4 Conclusion

In this chapter, a multi-filter system (MFS) is proposed to identify important age-related genes which have important role toward muscle atrophy in males and females using skeletal muscle microarray data. Genes are first sorted using three different evaluation methods including t-test, entropy and ROC. The system is then evaluated using publicly available microarray datasets of the gene expression of skeletal muscle tissue.

Subsequently, the most important genes are determined using majority voting based on the principle that combining multiple models can improve the generalization ability of the system. Each microarray dataset divided into three case of study including (young male versus old male; young women versus old women; old men versus old women) and the classes were based on case of study for example old men represent positive class and old women represents negative class. The results have indicated that the classification performance achieved by the proposed system yields the best classification performance when compared with similar numbers of genes identified in previous studies Liu et al, (2013) and Raue et al, (2012).

The aim of next chapter is to further improve the performance by identifying sarcopenia-related genes through wrapper feature selection method rather than only depend on the evaluation methods. In order to increase classification performance, therefore, future work is based on wrapper feature selection which helps identify more reliable sarcopenia-related genes because one of the feature selection method is that it considers the interaction between genes. This is unlike the filter approaches which do not consider such interactions. The limitation of this method was the proposed approach was based on evaluation methods which are not consider the interaction between genes therefore utilizing wrapper feature selection contribute towards the discovery of new significant genes. The next chapter will present wrapper feature selection based system and discusses its viability in identifying more reliable sarcopenia-related genes.

Chapter 4

Gene selection using random subset feature selection

4.1 Introduction

This chapter describes an extension of MFS system, presented in the previous chapter, where a subset of atrophy-related genes was identified using multiple evaluation methods including an ensemble of the t-test, entropy and ROC. Although, the MFS is able to significantly reduce the number of genes and achieve high levels of accuracy compared to previous studies (Liu et al,2013; Raue et al, 2012) , there are many insignificant genes that need to be identified, and many redundant genes should be removed. The main novelty of the study proposed in this chapter is to further improve the MFS in order to discover more reliable genes with higher levels of accuracy. To achieve this, a wrapper feature selection is proposed to be added to the previous system. In the proposed system, wrapper-based random subset feature selection (RSFS) has been chosen because, unlike the evaluation methods, the wrapper method considers the interactions between features. Therefore the proposed multi-filter system with single wrapper improve the accuracy and more reliable subset of genes obtained.

The proposed approach has been applied to the same dataset used in the previous chapter and the MFS results have indicated a significant increase in classification accuracy of 100% when compared with existing work Liu et al, (2013). In this study the proposed technique is used on two microarray datasets and the proposed system MFS is able to identify atrophy-related

genes in older male and female. All datasets are divided into following three case studies in relation to age and gender differences.

- young men versus old men
- old men versus old women
- young women versus old women

4.2 Material and the proposed method

4.2.1 Micro array gene expression data set

The proposed system was evaluated using two microarray datasets of the gene expression of skeletal muscle which are publicly available in the Gene Expression Omnibus (GEO) database. The first dataset includes 22 healthy subjects of various ages distributed as follows:

- 7 young males and 7 young females (20-29 years old),
- 4 old males and 4 old females (61-81 years old).

The second dataset consists of around 55,000 genes from only 36 samples distributed as follows:

- 7 young men and 8 young women
- 10 old men and 11 old women

As per Lui et al. 2013, In the study A dataset is divided into three cases. The first case involves 11 females (7 young and 4 old), the second case consists of 11 males (7 young and 4 old) and the last case contains 8 samples (4 old men and 4 old women).

gene subset selection was conducted using feature ranking techniques where biomedical data have extremely high dimensionality.

4.2.2 Gene subset selection using Feature ranking techniques

The high dimensionality of microarray datasets represents a serious challenge because not all genes are relevant to the atrophy. To identify the most relevant atrophy-related genes among all of the genes in the dataset. Machine learning techniques have been suggested to address this problem. The wrapper feature selection method is used in order to select more reliable genes than those identified by the evaluation methods, presented in the previous chapter, because it considers the interactions between genes and also avoids over-fitting problems (Saeys et al, 2007). In this investigation, three different evaluation methods and a single wrapper feature selection are used. The evaluation methods and wrapper feature selection method are briefly described as below:

4.2.3 Wrapper feature selection

A wrapper feature selection method is an induction algorithm. It selects the best subset of features according to its predictive power using a supervised classifier (Kohavi and John, 1997; Cadenas and Martínez, 2013). Random subset feature selection (RSFS) (Räsänen et al, 2013) is chosen to deal with the candidate genes that are identified using different evaluation methods and the algorithm includes the following steps:

K= number of subsets

For I= n : K

STEP 1 selection

• Randomly select subset n_i of M features f_x by uniform distribution

STEP 2 classification

• Subset n_i is classified using the K-NN classifier

STEP 3 update relevance

• The relevance r_x of all used features f_x is updated based on

$$r_x \leftarrow \ r_x + C_i - \mathrm{E}\{\mathrm{C}\} \qquad (1)$$

Where $\mathrm{C_i}$ is the value of the criterion function for the present iteration i and $\mathrm{E\{c\}}$ is the expected value of the criterion function.

STEP 4 goodness of features

$\mathrm{F_1} \in [\mathrm{n_1\ n_5\ n_9}]$ performance [50%, 40%, 40%]

$\mathrm{F_2} \in [\mathrm{n_3\ n_4\ n_6}]$ performance [60%, 70%, 30%]

$\mathrm{F_3} \in [\mathrm{n_2\ n_7\ n_8}]$ performance [90%, 90%, 90%]

$\mathrm{F_i}$ is feature, $\mathrm{n_i}$ is feature pool,

End for

4.2.4 Classification

The selected subset of genes is tested for its generalization power using supervised classification. The K-nearest neighbor (KNN) classifier and support vector machine SVM are used to evaluate the classification performance. The classification performance was assessed using the “Leave-One-Out Cross Validation” (LOOCV).

4.2.5 Proposed System

Figure 4.1 illustrates the proposed multi-filter single wrapper system (MFSWS) which is inspired by the fact that combining multiple models can improve the generalization of the system. The dataset is divided into testing and training sets based on leave-one-out cross validation (LOOCV) for example if the total number of samples is 20 samples so in the first step 19 samples are used for training and one sample for testing this will be repeated 20 times. Each time the multi-filter single-wrapper system (MFSWS) is applied, which includes three different evaluation methods, namely the t-test, the ROC-AUC and the entropy as well as random subset feature selection (RSFS) is added to the framework as shown in figure 4.1. All genes were ranked based on the evaluation method criterion. From these sorted genes a unique subset of genes is selected based on majority voting, as shown in Table 4.1 then the candidate genes are filtered using the wrapper feature selection method. Later, the performance of the final subset of genes produced by the wrapper feature selection method are determined using two types of classifier, K-NN and SVM.

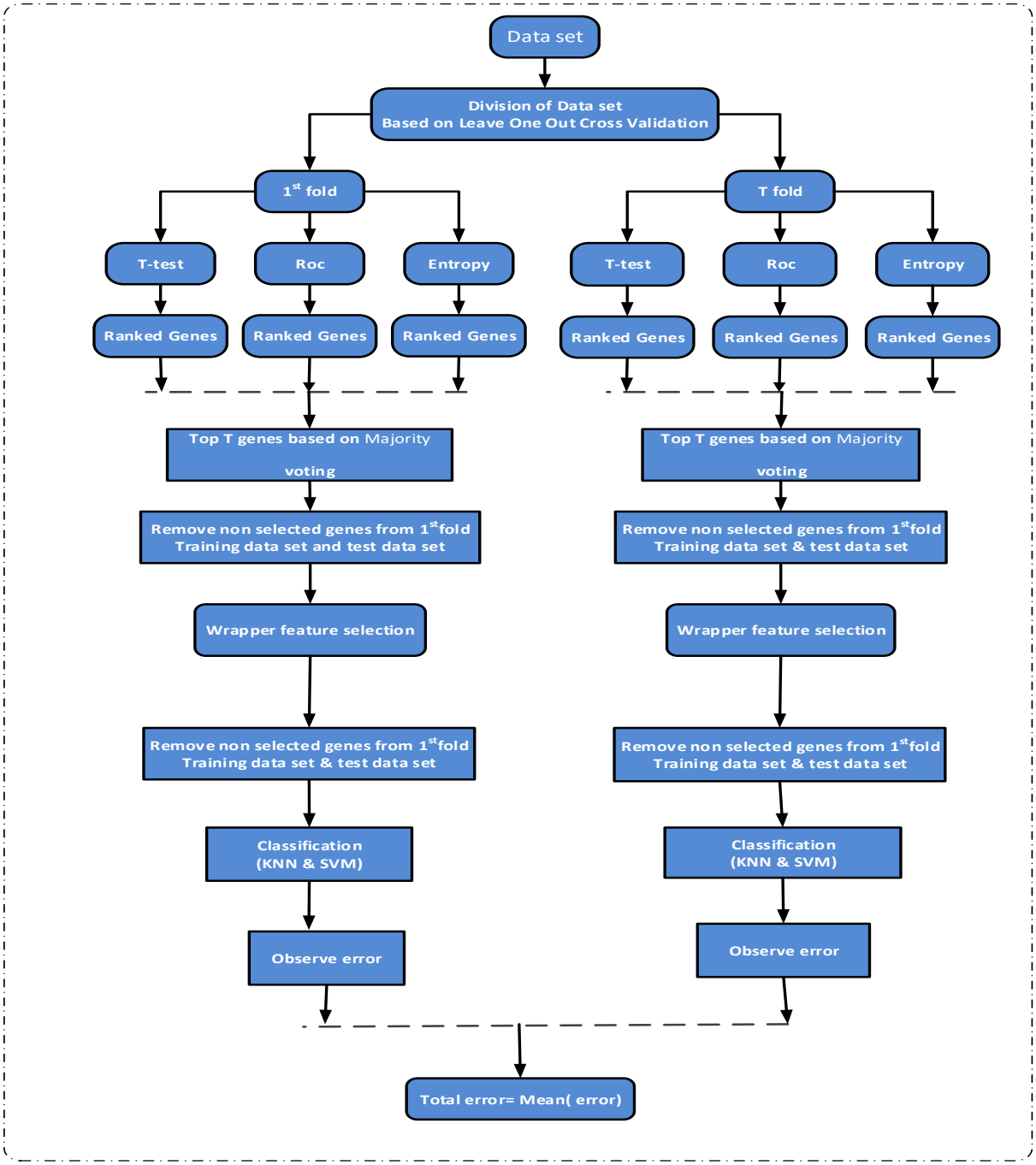

Figure 4.1: Proposed Multi-Filter System single wrapper system (MFSWS).

4.3 Results and discussion

4.3.1 Study A

A. Case study 1: young males versus old males

A total of 11 male samples (7 young and 4 old) has been used in this study. In this case the true positives represent the correctly classified young males, true negatives v correctly classified old males, false positives represent the incorrectly classified young males and false negatives represents incorrectly classified old males. Table 4.2 compares the performance of the proposed MFSWS with that obtained from using the first 20 genes identified by Liu et al (2013). It is observed that the best performance is obtained using the 1NN and the 3NN classifier at 100%, and the improvement is mainly due to high specificity. This improvement starts from the second feature as shown in Figure 4.2. The MFS is only able to achieve 90% while genes obtained by Liu et al, (2013) achieved 63% only. Further analysis has revealed that out of 20 genes, there are no common genes in both systems. A subset of new genes have been identified and they can play significant role in muscle atrophy in different ages. For young and old men. Some of the new and common genes are shown in Table 4.2. Comparing MFS with Liu et al (2013) study, it is observed that the MFSWS achieved the best performance which approximately 100% compared to others due to new important features were added to the subset of features, the significant improvement was 10% compared to MFS which was 90% as shown in figure 4.2. This is due to the selection of some new genes that have significantly contributed to an improvement in the performance of the system.

Table 4.2 Performance comparison between MFS, Liu et al (2013) and MFSWS (young males versus old males)

Classifier	Accuracy			Sensitivity			Specificity		
	Proposed MFS	Liu et al, (2013)	Proposed MFSWS	Proposed MFS	Liu et al, (2013)	Proposed MFSWS	Proposed MFS	Liu et al, (2013)	Proposed MFSWS
1NN	81%	81%	100%	71%	75%	100%	100%	100%	100%
3NN	90%	90%	100%	85%	85%	100%	100%	100%	100%
SVM	72%	72%	**81%**	62%	62%	85%	75%	75%	90%

Table 4.3 Confusion matrix: young male versus old male

6	0
1	4

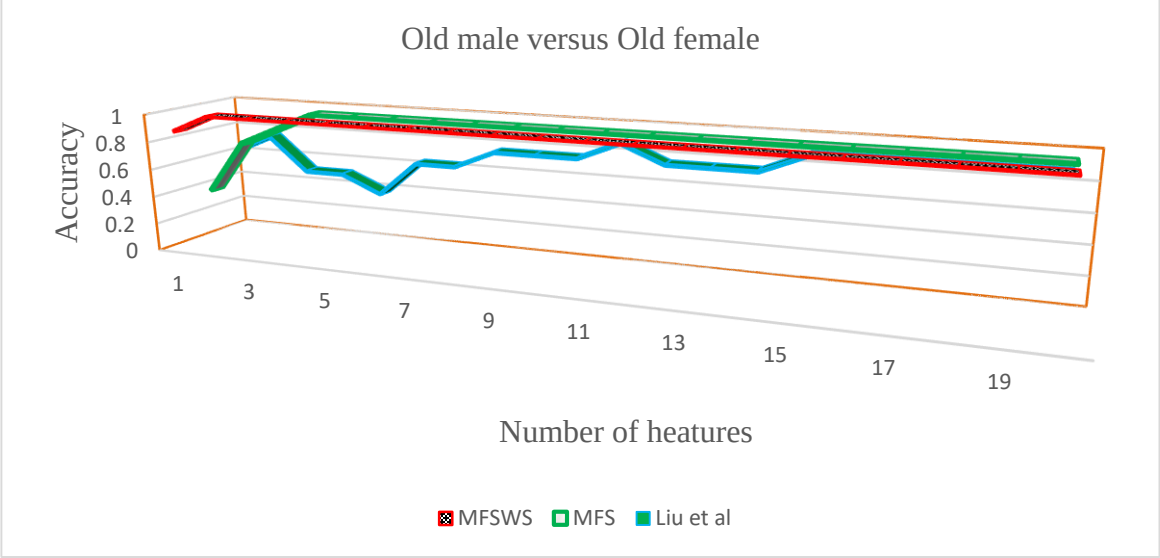

Figure. 4.2. Comparison of the performance change against varying number of genes for MFS, Liu et al and MFSWS (old male versus young male)

Table 4.4: new genes selected by the proposed system MFSWS (young male versus old male).

GENE SYMBOL	GENE DESCRIPTION
PAX8	PAIRED BOX 8
HSPA6	HEAT SHOCK 70KDA PROTEIN 6
RBBP6	ADENYLATE KINASE DOMAIN CONTAINING 1
MFAP3	NUCLEOREDOXIN-LIKE 2

Other analyses have been conducted on this subset of genes, including p-value and fold change (FC. DAVID Bioinformatics Resources 6.8 software (https://david.ncifcrf.gov/) is used to provide a functional interpretation of large lists of genes derived from genomic studies, such as in a microarray. Based on statistical results P-value, FC and the gene directions for the first 20 genes are differentially expressed which 9 genes were up-regulated and 11 down-regulated in addition the gene pathway based on DAVID reflected three biological themes: mitochondrial structure and function, immune function and transcription. Some of obtained genes consider sarcopenia-related genes in old men especially those are reflected mitochondrial function and it were down-regulated. This means that the sarcopenia is very low in old men.

Table 4.5 shows the MFSWS genes with brief descriptions. This subset of genes identified by MFSWS could play a significant role in explaining the impact of age on muscle atrophy between young and old females.

Table 4.5: P-values and FC of genes in male with brief descriptions

P-value	**Fold change FC**	**Gene Symbol**	**Gene Discretion**
0.000104	-1.415	TAB3	TGF-beta activated kinase
0.006975	1.367128	MAPK1	Mitogen-activated protein kinase 1
0.012024	-1.15277	ATP6V1E2	ATPase, H+ transporting, lysosomal 31kDa, V1 subunit E2
0.006207	-1.35658	CORO6	Coronin 6
2.90E-05	1.654688	MFAP3	Microfibrillar-associated protein 3
0.042443	1.081152	C4orf33	Chromosome 4 open reading frame

B. Case study 2: old male versus old female

This case study involves 8 adults (4 old men versus 4 old women). In this case the true positives represent the correctly classified old males, true negatives represent correctly classified old females, false positives represent the incorrectly classified old males and false negative represents incorrectly classified in old females. The aim of this study was to examine gene expression levels among male and female old people. According to table 4.6, MFSWS and MFS achieved the best performance which 100% due to the smaller number of genes.

Table 4.6: Classification performance: old men versus old women.

Classifier	Accuracy			Sensitivity			Specificity		
	Proposed MFS	Liu et al, (2013)	Proposed MFSWS	Proposed MFS	Liu et al, (2013)	Proposed MFSWS	Proposed MFS	Liu et al, (2013)	Proposed MFSWS
1NN	100%	0.75	100%	100%	50%	100%	100%	100%	100%
3NN	100%	0.75	100%	100%	50%	100%	100%	100%	100%
SVM	75%	62%	75%	75%	75%	75%	75%	50%	75%

Table 4.7 Confusion matrix : old male versus old female

4	0
0	4

According to the figure 4.3 it observed that the MFSWS had higher performance starting from the first feature compared to MFS which started after the fourth feature as shown in. It is observed that genes selected using MFSWS had a classification accuracy of 100% using both 1NN and 3NN with high sensitivity and specificity.

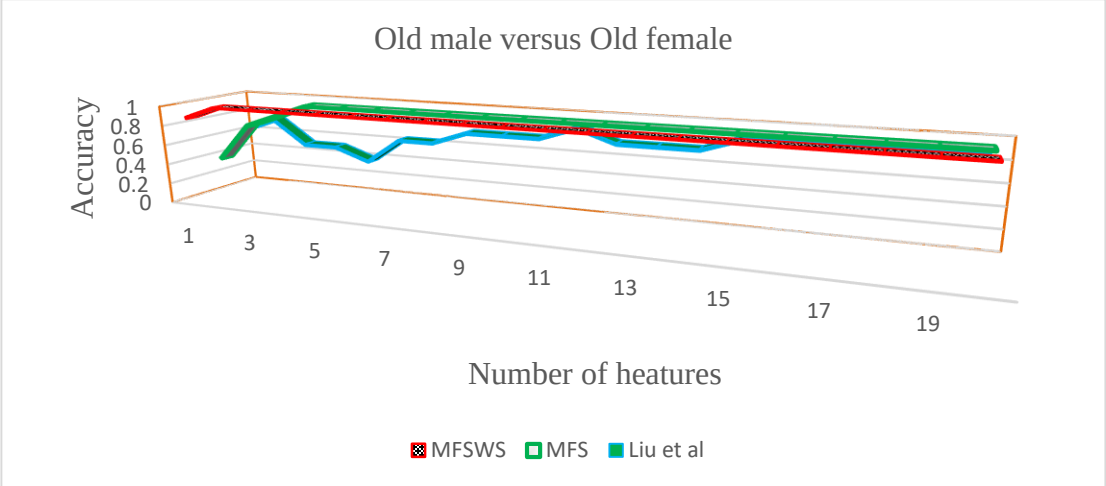

Figure. 4.3. Graph showing performance of the system using varying number of genes (old male versus old female).

P-value and fold change (FC) tests have been found for this subset of genes. Based on the P-value ($P<0.05$) the results indicated that 4 genes were up-regulated and 16 down-regulated as shown in table 4.8. They reflected three biological themes including mitochondrial structure and function, gene transcription and translation, Cytoskeleton. Most of these consider sarcopenia-related genes because they show more than 1.5 fold with p-value <0.05 female compared to their value counterparts in young. Also few of them refer to presence of sarcopenia in men.

Table 4.8: P-values and FC of genes in adults with brief descriptions (old male and old female)

P-value	**Fold change FC**	**Gene Symbol**	**Gene Discretion**
1.23E-05	-10.0692	EIF1AY	eukaryotic translation initiation factor 1A, Y-linked initiation factor 1A, Y-linked
0.030005194	-2.443018858	AQP4	aquaporin 4
0.042921543	-2.3817	ABRA	actin-binding Rho activating protein
0.021555225	-2.05808	CITED2	Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain 2
0.03763042	-2.041988866	RPS23	ribosomal protein S23

C. Case study 3: young females versus old females

Another objective of this investigation was to compare basal level gene expression between old women and young women. In this case the true positives represent the correctly classified young females, true negatives represent correctly classified old females, false positives represent the incorrectly classified young females and false negatives represent incorrectly classified old females.. This case study involved 8 adults (4 young female versus 4 old male). MFSWS gave a classification accuracy of 100% using both 1NN and 3NN with high Sensitivity and specificity. While using SVM is able to achieve only 72% however MFSWS is outperform among all as shown in table 4.9.

Table 4.9 Classification performance: young females versus old females.

classifier	**Accuracy**			**Sensitivity**			**Specificity**		
	Proposed MFS	**Liu et al, (2013)**	**Proposed MFSWS**	**Proposed MFS**	**Liu et al, (2013)**	**Proposed MFSWS**	**Proposed MFS**	**Liu et al, (2013)**	**Proposed MFSWS**
1NN	100%	81%	100%	100%	75%	100%	100%	100%	100%
3NN	0.72	0.72	100%	85%	85%	100%	50%	50%	100%
SVM	63%	54%	72%	70%	60%	85%	50%	50%	50%

Table 4.10 Confusion matrix : old male versus old female

7	0
0	4

Based on the figure 4.4. In the beginning the first feature is achieved classification accuracy more than 80%, while the classification accuracy based on MFS was less than 75%. However

the MFSWS and the MFS have the same classification accuracy of 100%starting from gene number 11.

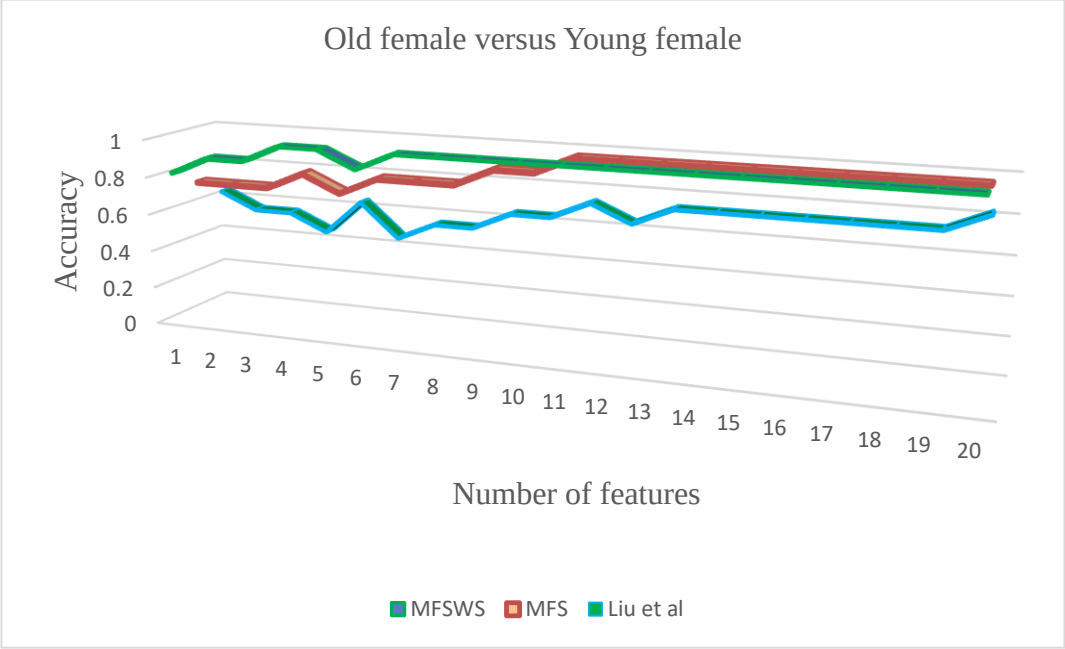

Figure. 4.4. Graph showing performance of the system using varying number of genes (old female versus young female).

These genes were subjected to further analyses, including p-value and fold change (FC). In this study gene expression was considered significant if the finding relative expression showed a P-value < 0.05, and the results indicated that 13 genes were up-regulated and 7 down-regulated and they reflected three biological themes including gene transcription, mitochondrial function and inflammatory response and immune function Table 4.11 shows the MFSWS genes with brief descriptions. Based on the p-value < 0.05 and FC > 1.5 moreover the direction was down-regulated in pathway of mitochondrial function these genes consider as sarcopenia-related genes in women. Our finding support two studies De Magalhães et al, 2009 and Zahn et al, 2006.

Table 4.11: P-values and FC of genes in female with brief descriptions.

P-value	**Fold change FC**	**Gene Symbol**	**Gene Discretion**
0.010757	2.4	SHISA4	Shisa homolog 4 (Xenopus laevis)
0.033737	2.4	MALAT1	Metastasis associated lung adenocarcinoma transcript 1
0.026499	2.5	NEDD1	Neural precursor cell expressed, developmentally
0.004881	-4.3	OGDH	Methionine adenosyl transferase II, alpha
0.005099	-3.4	CXCR2	Chemokine (C-X-C motif) receptor 2
0.192838	-3.2	MGP	Matrix Gla protein

4.3.2 **Study B**

The dataset used in this study includes 54,623 genes, and is publicly available at the Gene Expression Omnibus (GEO) dataset NCBI (2013).The total number of subjects were 36 healthy young men and women of different ages, 19 females (8 young, 11 old) and 17 males (7young, 10 old).

A. Case study 1: young males versus old males

Seventeen male samples were involved in this investigation (7 young and 10 old). In this case the true positives represent the correctly classified young males, true negatives represent

correctly classified old males, false positives represent the incorrectly classified young males and false negatives represent incorrectly classified old males. Table 4.12 shows a comparison of performance among MFSWS, MFS. It is observed that the proposed MFSWS achieved higher accuracy using INN and the 3NN which approximately 100%, especially in the first 20 genes. But it able to achieve 88% using the SVM

Table 4.12 Classification performance: Young males versus old males

classifier	**Accuracy**			**Sensitivity**			**Specificity**		
	Proposed MFS	**Liu et al, (2013)**	**Proposed**	**Proposed MFS**	**Liu et al, (2013)**	**Proposed**	**Proposed MFS**	**Liu et al, (2013)**	**Proposed**
1NN	88%	82%	**100%**	85%	71%	100%	90%	90%	100
3NN	88%	82%	**100%**	85%	71%	100%	90%	90%	100
SVM	82%	76%	88%	71%	72%	85%	90%	80%	90%

Table 4.13 Confusion matrix : young males versus old males

7	0
0	10

The MFSWS gives high classification accuracy of 100% at all times compared to MFS, and this improvement starts from the second feature. The highest accuracy was due to the smaller number of genes that were chosen.

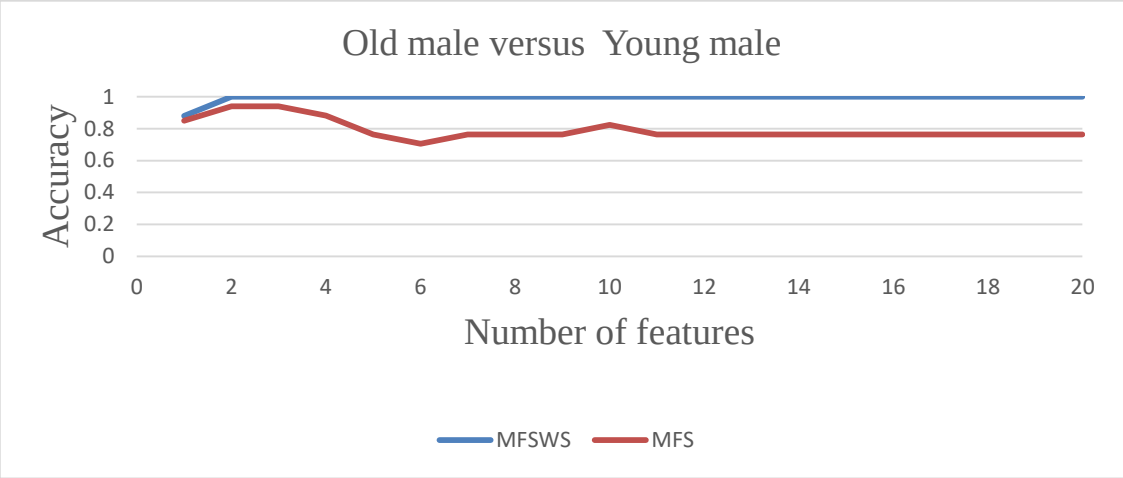

Figure. 4.5. Performance of the MFEWS using varying number of genes (old male versus young male).

Then, P-value and fold change (FC) are used to determine how significantly up- or down-regulated. The results revealed that 8 genes were up-regulated and 12 down-regulated and these genes reflected four biological themes, including mitochondrial structure and function, immune function, inflammatory response and cytoskeleton. There are some genes consider sarcopenia-related genes especially those have low p-value and FC > 1.5 such as CD33 as shown in table 4.14. This means that the probability of presence of sarcopenia is very low in old men because the FC levels show slightly change compared to young men.

Table 4.14: P-values and FC of genes in men adults

P-value	**Fold change**	**Gene Symbol**	**Gene Discretion**
0.03525	-1.57	CD33	CD33 molecule
0.02591	-1.48	ARHGAP26	Rho GTPase activating protein
0.016403	1.06	DDX3X	DEAD (Asp-Glu-Ala-Asp) box polypeptide 3,

B. Case Study 2: young females versus old females

This investigation includes 19 female individuals (8 young and 11 old). In this case the true positives represent the correctly classified old males, true negatives represent correctly classified old females, false positives represent the incorrectly classified old males and false negatives represent incorrectly classified old females. The comparison among the three systems is shown in Table 4.15..

Table 4.15 Young females versus old females

classifier	Accuracy			Sensitivity			Specificity		
	Proposed MFS	Liu et al,	Proposed	Proposed MFS	Liu et al, (2013)	Proposed	Proposed MFS	Liu et al,	Proposed MFSWS
1NN	94%	75%	84%	100%	75%	75%	90%	70%	90%
3NN	82%	68%	84%	100%	62%	75%	100%	70%	90%
SVM	75%	63%	75%	75%	63%	75%	70%	65%	70%

Table 4.16 Confusion matrix : young female versus old female

8	0
0	11

The figure 4.6 shows that the classification accuracy using both systems were fluctuated but in most cases MFS was better than MFSWS this due to there are some genes were selected by MFSWS have effected on the classification accuracy.

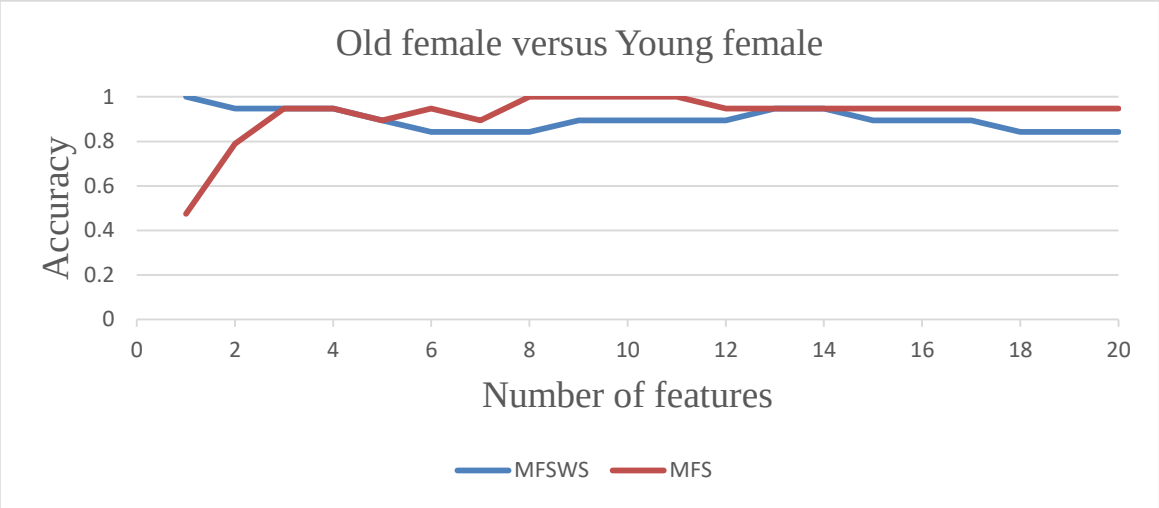

Fig. 4.6. Performance of the system using varying number of genes (old female versus young female)

The selected genes underwent further analyses including p-value and fold change (FC). In this study gene expression was considered significant if the relative expression showed a P-value < 0.05, and the results indicated that 5 genes were up-regulated and 15 were down-regulated as shown in table 4.17, and they reflected four biological themes including gene transcription, mitochondrial function, cytoskeleton response and immune function, Some of obtained genes show high classification accuracy also p-value < 0.05, FC > 1.5, in addition pathway in mitochondrial function such as PP2R and HRK in table 4.14 which indicate to muscle atrophy.

Table 4.17: P-values and FC of genes in adults with brief descriptions.

P-value	**Fold change FC**	**Gene Symbol**	**Gene Discretion**
0.011635	-2.72	PP2R	Protein phosphatase 2, regulatory
0.041451	-1.12	GCLC	Glutamate-cysteine ligase, catalytic subunit
0.034426	-2.77	HRK	Harakiri, BCL2 interacting protein (contains only BH3
0.010457	-5.25	MYO19	Myosin XIX
0.008631	-1.65	NEFL	Neurofilament, light polypeptide

C. Case Study 3: old males versus old females

This investigation includes 21 individuals samples (10 men and 11 women). In this case the true positives represent the correctly classified old males, true negatives represent correctly classified old females, false positives represent the incorrectly classified old males and false negatives represent incorrectly classified old females. Table 4.18 illustrates the performance comparison for 20 features of MFSWS, MFS and Raue et al. 2013.

Table 4.18 Classification performance: old males versus old females

classifier	Accuracy			Sensitivity			Specificity		
	Proposed MFS	Liu et al, (2013)	Proposed MFSWS	Proposed	Liu et al, (2013)	Proposed MFSWS	Proposed MFS	Liu et al, (2013)	Proposed MFSWS
1NN	95%	75%	100%	90%	75%	100%	100%	70%	100%
3NN	80%	66%	100%	90%	0.62	100%	75%	75%	100%
SVM	80%	76%	80%	70%	90%	90%	75%	82%	75%

Table 4.19 Confusion matrix : old male versus old female

10	0
0	11

Figure 4.7 shows that the MFSWS have the best classification accuracy of 100% this due to the smaller number of genes in this comparison, MFS is fail to achieve the same classification accuracy of 100% except after the gene number 4 which mean that there are impotent genes are involved to MSF subset of genes.

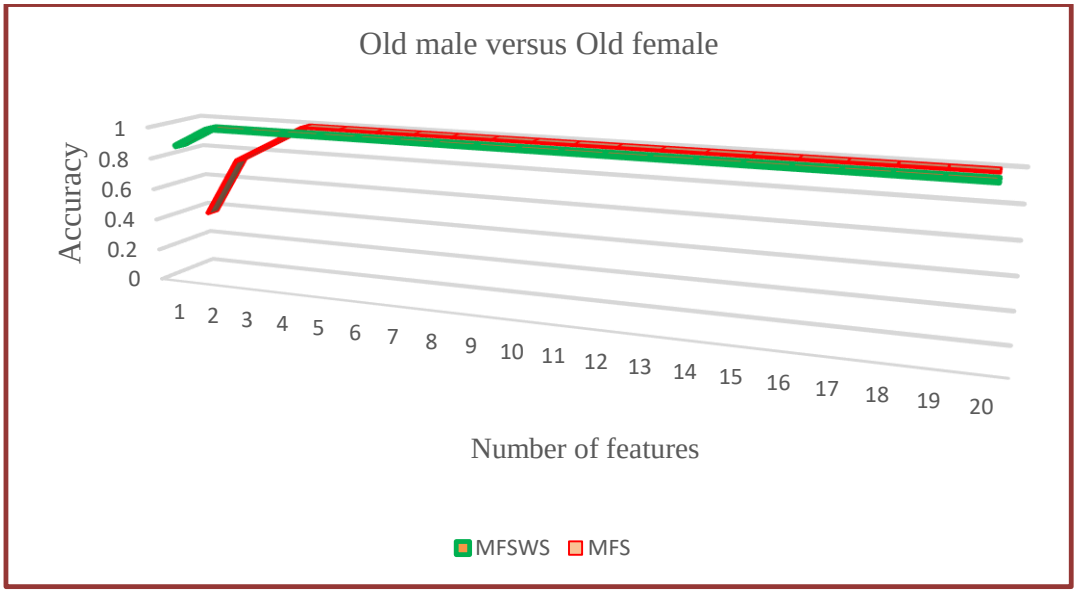

Figure. 4.7. Performance of the system using varying number of genes (old male versus old female)

According to the table 4.18 the MFSWS outperforms the other systems 100 % with a accuracy for the first 20 features using 1NN and 3NN classifiers. Figure 4.7 shows the performance of the first 20 genes selected by MFSWS and MFS, and it can be observed that the MFSWS was able to produce high classification accuracy approximately 100% for the 20 genes compared to MFS. Further analysis including P-value and fold change (FC) applied on these genes. Of those with $P<0.05$ the results show that there are 3 genes which were up-regulated and 17 down-regulated as show in table 4.20. Based on Biological interoperation these genes reflected three biological themes, including mitochondrial structure and function, gene transcription and extracellular. Most of these genes were down-regulated in female and pathway was mitochondrial function especially in old women, which refer to the presence of muscle loss in women human.

Table 4.20: P-values and FC of genes in adults with brief descriptions.

P-value	Fold change FC	Gene Symbol	Gene Discretion
0.029164	4.47	HTRA2	HtrA serine peptidase 2
0.011914	-2.93	PREPL	prolyl endopeptidase-like
0.000646	-3.24	SERTAD2	SERTA domain containing 2
0.0280209	-3.31	PRKACB	protein kinase, cAMP-dependent, catalytic, beta
0.035519	3.09	ALPK1	Alpha-kinase 1

4.4 Conclusion

In this investigation, a multi-filter single wrapper system (MFSWS) has been proposed to select significant subsets of genes for males and females using skeletal muscle related microarray datasets. Firstly genes are sorted using three different evaluation methods (t-test, Entropy and ROC) and then wrapper feature selection has been used proposed to reduce the dimensionality of the subset of genes provided by three filters in each fold in order to select the most important genes. The proposed system is evaluated using publicly available microarray datasets for gene expression in skeletal muscle tissue. Then, important genes are acquired using a combination of multi evaluation methods and wrapper feature selection to improve the generalization of the system. The results indicate that the proposed system yields the best classification performance when compared with similar number of genes identified (20 genes) in previous studies Raue et al, (2012) and Liu et al, (2013) for example as shown in figure 4.7.

This chapter presented a multi-filter single wrapper model (MFSWS), and the evaluation results indicate that the proposed model achieves high performance when applied to different

datasets. The results show that KNN achieved high performance more than SVM in several cases. In terms of biological interpretation based on the DAVID software, the genes selected by the proposed MFSWS had low p-value with down-regulated genes in mitochondrial function for both older women and men. Notably that the aging-associated transcriptional changes in women muscle more than men. Our finding supported by De Magalhães et al, (2009) and Zahn et al, (2006). The limitation of this proposed framework was is the candidate genes were based on single feature selection method RSFS. Therefore using different single wrapper feature selection methods will yield to different new genes. Using a combination of some these feature selection methods will solve this problem.

Therefore the next chapter will discuss the new proposed framework multi-filter multi-wrapper system (MFMWS) based on wrapper feature selection method RSFS. Which designed to identify more reliable sarcopenia-related genes with high classification performance through multi wrapper feature selection techniques rather than single-wrapper feature selection. The system is evaluated using 3 different microarray datasets.

Chapter 5

Multi-feature selection based approach for analysis of microarray data for skeletal muscle mass loss

5.1 Introduction

Although the previous MFSWS system was able to identify age-related genes with high accuracy for each case of study, the final subset of selected genes was based on one type of feature selection. Different types of feature selection methods provide different subsets of genes despite achieving the same training accuracy, therefore it doubtful whether the genes selected by one specific model are true biomarkers. . In order to overcome the drawback of MFSWS, we proposed a new multi-filter multi wrapper system (MFMWS) which is based on majority voting using a combination of different feature selection methods. The use of multiple feature selection methods is intended to enhance the reliability of the classification Therefore the work in this chapter is an extension to the work discussed in chapter 4 in which a subset of genes was identified using a multi-filter single wrapper system (MFSWS).

The main contribution in this chapter is to introduce the multi-filter multi wrapper system (MFMWS) which is able to identify distinct subsets of sarcopenia-related genes with high accuracy. Experiments are conducted on three different datasets and the results

for our system are compared with those for existing systems All datasets are divided into the following case studies.

- young men versus old men
- old men versus old women
- young women versus old women

5.2 Material and proposed method

5.2.1 Microarray gene expression data sets

Three microarray datasets comprising the gene expression of skeletal muscle tissue are used. The datasets are publicly available in the Gene Expression Omnibus (GEO). Dataset 1 and dataset 2 were described in the previous two chapters while the third dataset is used to compare the proposed system MFMWS with MFSWS. Dataset 3 has included a total of 55 healthy males and females in various ages, 31 young (16 male, 15 female) and 24 old (12 male and 12 female).

5.2.2 Genes subset selection using Feature ranking techniques

In this investigation, the same evaluation methods used in previous two chapters, and two different wrapper feature selection including random subset feature selection (RSFS) which is used and described in the previous chapter and SVM-Recursive Feature Elimination (SVM-RFE)

5.2.3 Classification

The final subset of genes was determined for its generalization related to classification using SVM and KNN classifiers, and their performance of classification was measured based on leave-one-out cross validation (LOOCV).

5.2.4 Proposed system

the framework of the proposed MFMWS, which is inspired by a combination of multiple models in order to improve the generalization ability of the system and to select sarcopenia-related genes. Datasets were normalised due to the high variation between gene values and two different feature selection methods are used: random subset feature selection RSFS (Räsänen, and Pohjalainen 2013) and SVM-Recursive Feature Elimination (SVM-RFE) (Guyon et al, 2002). Each evaluation method serves as a filter term in order to rerank the data features, and the goal is to sort genes according to criteria specified in the evaluation methods. Each evaluation method has its own characteristics and there is usually a fair proportion of overlapping and non-overlapping genes among the lists of genes selected by the three different filters. An N unique subset of genes is obtained based on majority voting for the sorted genes, as illustrated in Table 5.1. Let's assume that there is a total of 10 genes and the aim is to select the top 5 genes. Genes 9 and 10 are selected by all of the feature ranking techniques, and so these are considered the most important genes. Genes 1, 4 and 5 are selected twice and thus are also considered important genes by the system. It should be noted that, due to majority voting genes 2, 3 and 6 are not selected by the system. Then three different wrapper feature selection methods are applied in order to select the important genes. Each method has own strategy to select genes, and an N unique subset of genes are obtained based on majority voting for the sorted genes as in the filter method. Finally, KNN (k=1 and 3) and SVM are applied to the last subset of genes obtained by the wrapper methods in order to check the predictive

performance as shown in figure 5.1. The last subset of genes are then evaluated by performing leave-one-out cross-validation (LOOCV).

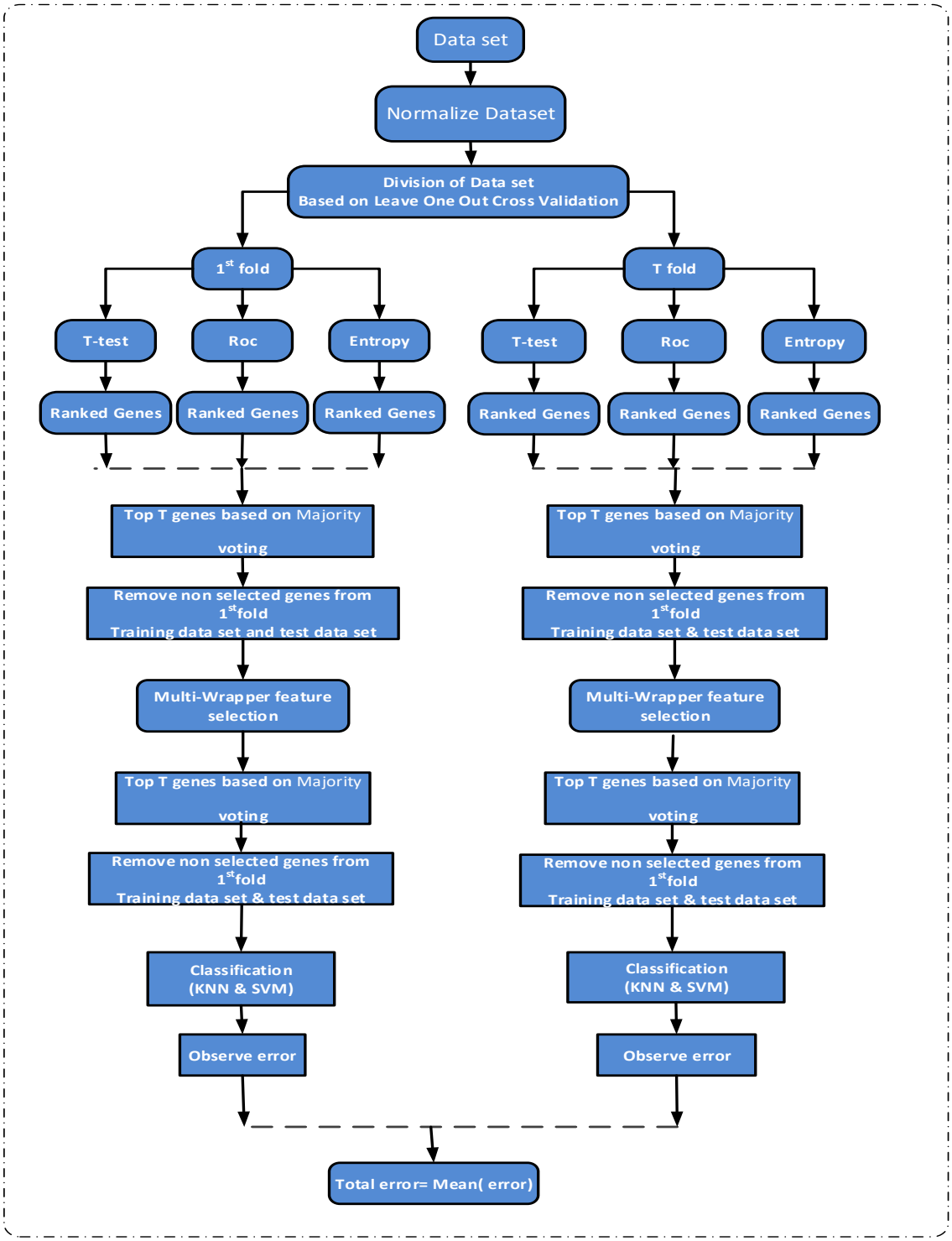

Figure5.1 Proposed Multi-Filter System (MFMWS).

Table 5.1 majority voting to select important genes.

Ranking	Top Ranked Genes
t-test	1, 4, 6, 9, 10
ROC-AUC	1, 2, 5, 9, 10
Entropy	3, 4, 5, 9, 10
Majority Voting	1, 4, 5, 9, 10

5.2.5 Feature selection

A. Random subset feature selection (RSFS)

The feature selection is performed using a new variant of random subset sampling methods (Räsänen et al, 2013)with k-nearest neighbors (kNN) as a classifier (Räsänen et al, 2013).it has been described in previous chapter.

B. Support vector machine recursive feature elimination (SVM-RFE)

This is used to extract an optimal subset of features from the original features. The SVM-RFE is an advanced version of support vector machine recursive feature elimination SVM-RFE presented by Guyon et al (2002). SVM-RFE is an approach for gene selection. Which is considered to be one of the most effective methods for selecting informative genes. It is often used to identify relationships within gene expression datasets.

5.2.6 Functional annotation of top genes

To determine the relationship between biological processes and the impact of age on the obtained genes. differentially functional annotation tools in the Database for Annotation, Visualization, and Integrated Discovery (DAVID) software are used. These tools are publically available () and allow researchers to examine large lists of genes (Dennis et al., 2003). DAVID assists in the interpretation of genome-scale datasets by facilitating the transition from data collection to biological meaning. While researchers are beginning to appreciate the statistical rigors required for the analysis of genome-scale datasets, a rate-limiting step in knowledge growth occurs at the transition from statistical significance to biological discovery. This kind of analysis can discover some common biological themes that are existing in a set of target genes, which lead to providing an investigator with additional clues for the follow up experiments. Gusev. (2008). Has performed pathway enrichment analysis using DAVID software. The aim was to address the problem of identifying major biological processes and signalling pathways that are collectively targeted by co-expressed miRNAs in cancer cells. Results show that there are 19 pathways were found that were enriched with microRNA target genes, some of them were over-expressed in colon cancer. Huang et al, (2014) used the DAVID software and the aim was to identify genes that play significant roles in regulating the pathogenesis of Acute respiratory distress syndrome (ARDS), and to determine how miRNAs contribute to the regulation of these genes. Results revealed that the over expressed genes were involved in two functional clusters more than 33% of them were involved in biological processes such as cellular homeostasis and regulation of apoptosis and the down-regulated genes were involved in two functional clusters, most of them down-regulated genes were enriched in the functional groups of acetylation and ion binding.

5.3 Results and discussion

In this section, the performance of the multi-filter multi-wrapper system (MFMWS) is evaluated using three microarray datasets. The results for the proposed system are then compared with those for MFSWS and MFS in which genes are selected based on three case studies (young males versus old males; young females versus old females; old males versus old females). In order to conduct fair comparison, the same number of genes are selected by the MFMWS and compared with the genes identified by MFSWS and MFS. The evaluation matrix used in this study are classification accuracy, Sensitivity and Specificity.

5.3.1 Dataset 1 In this investigation, microarray dataset is used which is publicly available at the Gene Expression Omnibus (GEO) dataset GDS5218 NCBI (2012).

A. Age-based differential gene expression in male

In this case study there are 11 individuals (4 old and 7 young) and all of them are male. In this case the true positives represent correctly classified young male, true negatives represent correctly classified old male, false positives represent incorrectly classified young male and false negatives represent incorrectly classified old male. Table 5.2 shows the performance of classification of genes identified by MFMWS compared to genes that are identified by MFS and MFSWS. The MFMWS and MFSWS achieved the highest classification accuracy using the 3NN and SVM classifiers which give approximately 100% classification accuracy, while MFS is able to achieve only 90% of classification accuracy.

Table 5.2 Classification performance: old men versus young men

classifier	Accuracy			Sensitivity			Specificity		
	MFMWS	***MFSWS***	***MFS***	***MFMWS***	***MFSWS***	***MFS***	**MFMWS**	***MFSWS***	***MFS***
3NN	100%	100%	90%	100%	100%	80%	100%	100%	100%
SVM	100%	100%	90%	100%	100%	85%	100%	100%	100%

Table 5.3 Confusion matrix : young men versus old men

7	0
0	4

Based on the figure 5.2 it is observed that the MFMWS has improved the accuracy from the first feature and is able to achieve classification accuracy of 100% also in the MFSWS the improvement starts from the second feature while the MFS is fluctuated and the best classification accuracy was approximately 90%. The highest classification accuracy is due to the small number of genes that are used in this experiment.

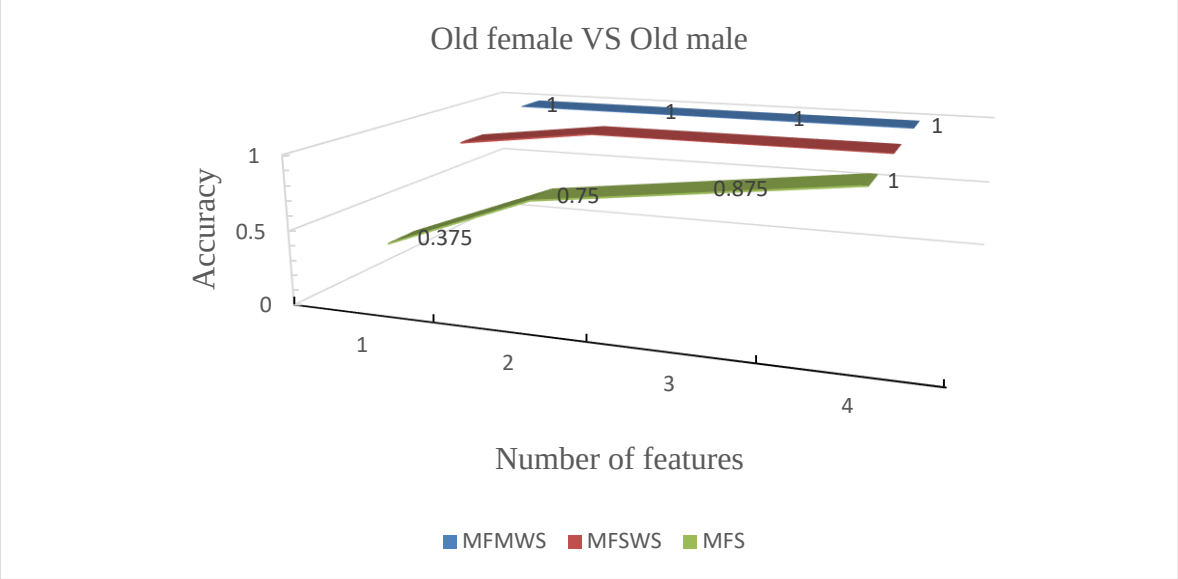

Figure. 5.2. Performance of the system using varying numbers of genes (old male versus young male)

Further analyses have been conducted on the discovered subset of genes such as using p-value and fold change (FC). In this study, a gene is considered important when the relative expression shows a P-value < 0.05. These genes represent two biological themes, including mitochondrial structure and function, gene transcription and translation. Three of these genes were down-regulated and the pathway reflects mitochondrial function which refer.to muscle atrophy in men. Table 5.4 shows some of the genes with brief descriptions. Therefore it can be said that the MFMWS has more accurate subset of sarcopenia-related genes than the other systems.

Table 5.4 P-values and FC of genes in adults with brief descriptions

P-value	Fold change	Gene Symbol	Gene Description
0.015686	-2.42179	ANKFN1	Ankyrin-repeat and fibronectin type III domain containing 1
0.027208	1.34038	JPH1	junctophilin 1
0.028765	-3.17448	THRSP	Thyroid hormone responsive
0.012024	-1.15277	GIMAP1	GTPase, IMAP family member 1

B. Age-based differential gene expression in females

In this study, the objective was to examine basal level gene expression in old and young females. In this case, true positives represent the correctly classified young females, true negatives represent the correctly classified old females, false positives represent the incorrectly classified young females and false negatives represent incorrectly classified old females.

This case study includes 11 adults (4 old females versus 7 old females). Table 5.5 shows the results of the comparison between the classification performances obtained for old and young females for the first 6 genes obtained by MFMWS compared to the 6 genes identified by MFS and MFSWS. The best performance of 100% is achieved by MFMWS using the 3NN and SVM classifiers. Based on figure 5.3, it is observed that the MFMWS outperforms on the other systems with 100%. This excellent classification accuracy is due to the smaller number of genes which just 4 genes in this comparison.

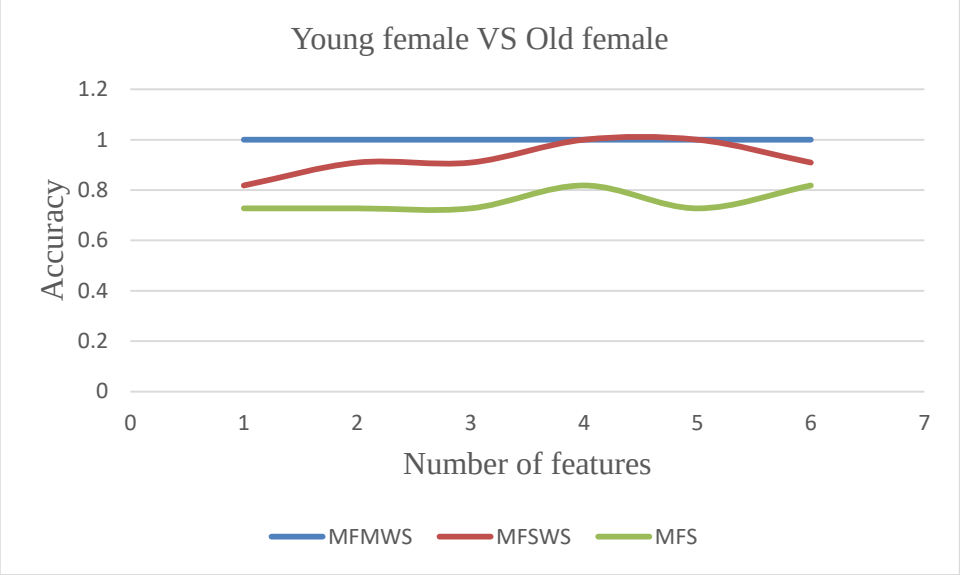

Figure. 5.3. Performance of the system using varying number of genes (young female versus old female).

Table 5.5 Classification performance: young women versus old women

classifier	**Accuracy**			**Sensitivity**			**Specificity**		
	MFMWS	***MFSWS***	***MFS***	***MFMWS***	***MFSWS***	***MFS***	**MFMWS**	***MFSWS***	***MFS***
3NN	100%	90%	81%	100%	100%	85%	100%	90%	80%
SVM	100%	90%	81%	100%	100%	90%	100%	90%	80%

Table 5.6 Confusion matrix : old male versus old female

7	0
0	4

Further analyses of these genes was conducted using the p-value and fold change (FC). In this study gene expression was considered significant if the relative expression showed a P-value < 0.5. The results showed that the 6 genes were down-regulated and they consistently represented mitochondrial structure and

function as shown in table 5.7. Therefore these genes were considered sarcopenia-related genes because they have very low p-values, FC was very high also the pathway was mitochondrial in down-regulated direction. This means that this genes associated to sarcopenia in females.

Table 5.7 P-value and FC for some genes with briefly descriptions.

P-value	**Fold change FC**	**Gene Symbol**	**Gene Description**
0.909695	-6.05	CAP1	CAP, adenylate cyclase-associated protein 1 (yeast)
0.026582	-4.52	ADD1	Adducin 1 (alpha)
0.051486	-7.66	ATP5D	ATP synthase, H+ transporting, mitochondrial F1 complex, delta subunit
0.009064	-8.75	MTRF1L	Mitochondrial translational release factor 1-like

C. Sex-based differential gene expression in older adults

In this case study there are 8 individuals (4 old males and 4 old females). In this case the true positives represent the correctly classified old males, true negatives represent correctly classified old females, false positives represent the incorrectly classified in old males and false negatives represent incorrectly classified in old females.

Table 5.8 Classification performance: old males versus old females

classifier	Accuracy			Sensitivity			Specificity		
	MFMWS	***MFSWS***	***MFS***	***MFMWS***	***MFSWS***	***MFS***	**MFMWS**	***MFSWS***	***MFS***
3NN	100%	100%	100%	100%	100%	100%	100%	100%	100%
SVM	100%	100%	100%	100%	100%	100%	100%	100%	100%

Table 5.9 Confusion matrix : old male versus old female

4	0
0	4

Table 5.8 shows the classification performance of discovered the genes by MFMWS compared to the 75 genes that are identified by MFS and MFSWS. Three systems MFMWS, MFSWS and MFS using the 3NN and SVM classifiers were able to achieve classification accuracy approximately 100%. However based on Figure 5.4 it observed that the optimal subset obtained by MFMWS because all 5 genes starting from the first gene until the fifth genes are able to achieve a high or similar accuracy with 100%.

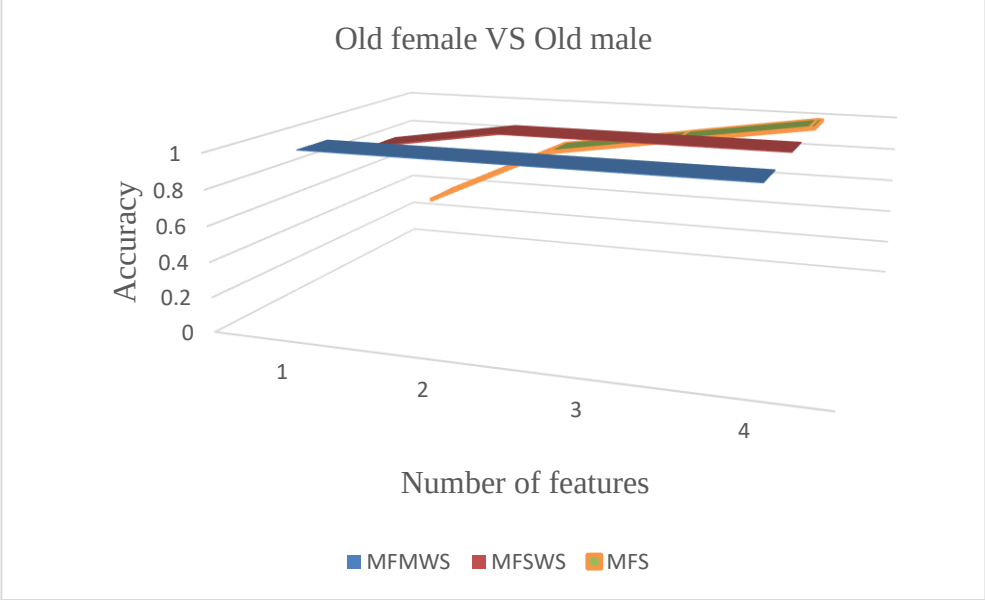

Figure. 5.4. Performance of the system using varying number of genes (old male versus old female).Based on the biological analysis as shown in table 5.10, these genes has the pathway reflected the mitochondrial in down-regulated direction, and compered these genes to their counterparts in young. there is no common genes between male and female in this subset of genes, it observe that three of them related to female while only one gene related to male and also this indicate that prevalence of sarcopenia in female more than male.

Table 5.10 P-value and FC for some genes with briefly descriptions

P-value	**Fold change FC**	**Gene Symbol**	**Biological themes**
0.00123	-10.06	EIF1AY	Eukaryotic translation initiation factor 1A, Y-linked
0.00605597	-2.81	MGP	Matrix Gla protein
0.21529747	-2.45	SYNPO	Synaptopodin
0.030005194	-2.44	AQP4	Aquaporin 4

5.3.2 Dataset 2 This dataset consists of 54,623 genes, and is publicly available at the Gene Expression Omnibus (GEO) dataset http://www.ncbi.nlm.nih.gov. The total number of subjects were 36 healthy young men and women of different ages, 19 females (8 young, 11 old) and 17 males (7 young, 10 old).

A. Age-based differential gene expression in male

This investigation includes 17 males (7 young and 10 old). In this case the true positives represent correctly classified in young male, true negatives represented by correctly classified old male, false positives represent the incorrectly classified young male and false negatives represents incorrectly classified old male. MFMWS was applied using nearest neighbor classifier KNN and the support victor machine (SVM). Comparing MFMWS, MFSWS and MFS, table 5.11 shows the results of classification performance based on three systems. MFMWS and MFSWS are able to achieve the highest performance with 100%. While the mfs is able to achieve just 90

figure 5.5 it is observed that the MFMWS is able to achieve the classification accuracy of 100% ta all features and has the highest or similar accuracy compared to the other systems this due to it involve new genes. This highest accuracy is due to the smaller number of genes.

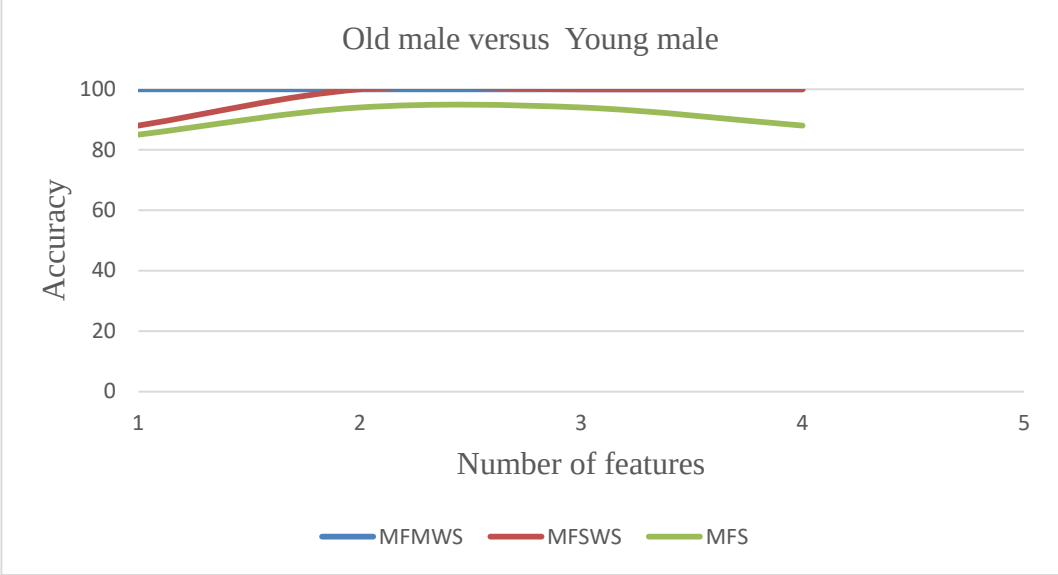

Figure 5.5: performance of the system using varying number of genes (old male versus young male)

Table 5.11 Classification performance: old males versus young males

classifier	Accuracy			Sensitivity			Specificity		
	MFMWS	*MFSWS*	*MFS*	*MFMWS*	*MFSWS*	*MFS*	MFMWS	*MFSWS*	*MFS*
3NN	100%	100%	90%	100%	100%	100%	100%	100%	80%
SVM	100%	100%	90%	100%	100%	100%	100%	100%	80%

Table 5.12 Confusion matrix : young men versus old men

7	0
0	10

Based on the biological analysis. The results revealed that these genes were down-regulated, and they reflecting three biological themes including mitochondrial

function and inflammatory response as shown in table 5.10. this indicate that these genes are sarcopenia-related genes in males.

Table 5.10: P-values and FC of genes in Adults (P-value <0.05)

P-value	Fold change FC	Gene Symbol	Gene Description
0.001156	-2.12	CD24	CD24 molecule
0.048906	-1.3	CXADR	Coxsackie virus and adenovirus receptor
0.062677	2.3	PEX13	peroxisomal biogenesis factor 13
0.026844	-1.55	NAT8L	N-Acetyltransferase 8-like (GCN5-related, putative)
0.560374	-2.2	BOK	BCL2-related ovarian

B. Age-based differential gene expression in female

This investigation includes 19 individuals (8 young females and 11 old females). In this case the true positives represent the correctly classified young females, true negatives represent correctly classified old females, false positives represent the incorrectly classified young females and false negatives represent incorrectly classified old females. The MFSWS was applied using the K-nearest neighbor (K-NN) classifier and support victor machine (SVM). Compared the first 7 genes that are identified by MFMWS, MFSWS and MFS. It observed that MFMWS able to achieve the best classification performance of 100% accuracy, this due to the smaller number of genes in this comparison while MFSWS is only able to achieve 90% and MFS achieved 84% as shown in table 5.11. This improvement mainly due to the high Sensitivity and Specificity.

Table 5.11: Classification performance of MFMWS, MFSWS and MFS

classifier	Accuracy			Sensitivity			Specificity		
	MFMWS	***MFSWS***	***MFS***	***MFMWS***	***MFSWS***	***MFS***	**MFMWS**	***MFSWS***	***MFS***
3NN	100%	89%	84%	100%	90%	80%	100%	90%	80%
SVM	100%	90%	84%	100%	90%	80%	100%	90%	80%

Table 5.12 Confusion matrix : young female versus old female

8	0
0	11

However MFMWS classification accuracy was constant with approximately 100% at all times compared to other systems, this due to the smaller number of genes in this comparison which are 7 genes. as shown in figure 5.6. This means that there are important genes involved in this subset of genes compared to others obtained by MFSWS and MFS.

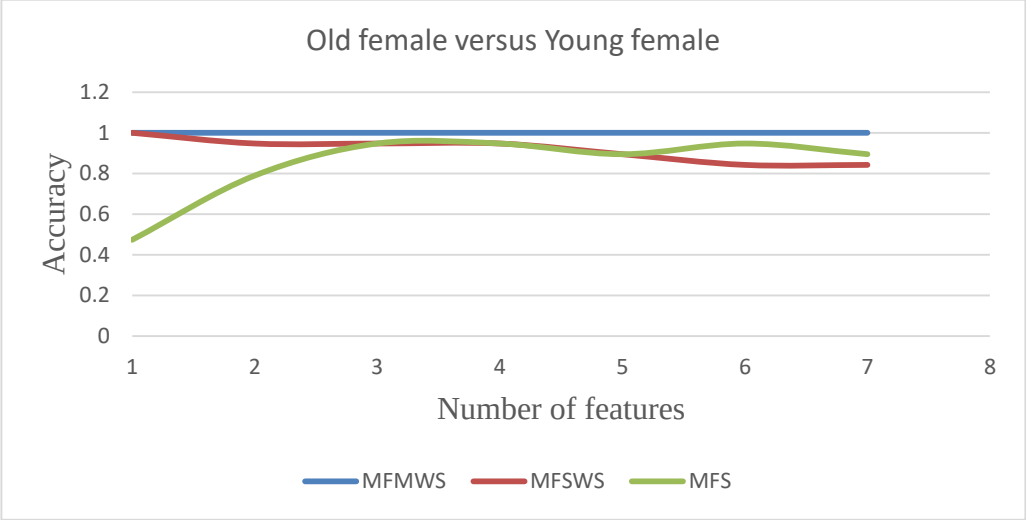

Figure 5.6: Performance of the system using varying numbers of genes (old females versus young females).

According to the biological analysis. The results revealed that MFMWS genes were down-regulated they reflected one biological theme including mitochondrial function. And also these genes were in high FC and very low p-values as shown in Table 5.13. Therefore these genes considered as sarcopenia-related genes in females were reflected mitochondrial function which refer to the presence of sarcopenia in female.

Table 5.13: P-values and FC of genes in adults

P-value	Fold change (FC)	Gene Symbol	Gene Description
0.041451	-2.12	GCLC	Glutamate-cysteine ligase, catalytic subunit
0.034426	-2.77	HRK	Harakiri, BCL2 interacting protein (contains only BH3
0.034193	-3.13	BMP1	Bone morphogenetic protein 1
0.002824	-2.30	APOE	Apolipoprotein E
0.009699	-8.78	ATAD2	ATPase family, AAA domain containing 2
0.028449	-3.21	CCR2	Chemokine (C-C motif)
0.00015	-1.75	LEP	Leptin

C. Sex-Based Differential Gene Expression in Older Adults

This investigation includes 21 adults (10 males and 11 females).). In this case the true positive represented by the correctly classified in old males, true negative represented by correctly classified in old females , false positive represents the incorrectly classified in old males and false negative represents incorrectly classified in old females. The MFSWS was applied using K-NN and SVM. Comparing the first 5 genes identified by the MFMWS, MFSWS and MFS the best performance was achieved all systems using 4 genes 100% as shown in table 5.14.,

Table 5.14: Classification performance: Old males versus Old females

classifier	Accuracy			Sensitivity			Specificity		
	MFMWS	***MFSWS***	***MFS***	***MFMWS***	***MFSWS***	***MFS***	***MFMWS***	***MFSWS***	***MFS***
3NN	100%	100%	100%	100%	100%	100%	100%	100%	100%
SVM	100%	100%	100%	100%	100%	100%	100%	100%	100%

Table 5.15 Confusion matrix : Old males versus Old females

10	0
0	11

According to the figure 5.7, it is observed that only MFS has not achieve the best performance using 3 or 3 genes.

Figure 5.7 shows that MFMWS is outperform to the other systems and the the MFMWS classification accuracy is 100%, this due to the smaller number of genes in this comparison.

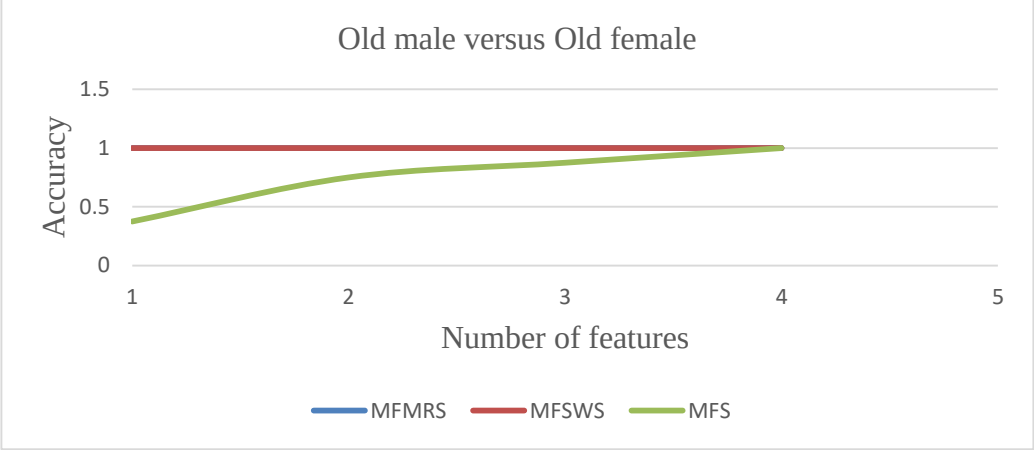

Figure 5.7: performance of the system using varying number of genes (old female versus old male)

Biological interpretation revealed that the MFMWS selected genes were down-regulated, and reflected one biological theme, mitochondrial function. Most of these genes have lower p-values and higher FC as show in table 5.16, which indicate to presence of muscle atrophy in both female and males. Also results revealed that there are 3 common genes between males and females.

Table 16 show MFSWS genes with brief descriptions

P-value	**Fold change FC**	**Gene Symbol**	**Gene Description**
0.0037723	-4.62	BCL2A1	BCL2-related protein A1
0.0012898	-9.75	DDX3X	DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, X-linked
0.0002981	-7.12	POLA1	Polymerase (DNA directed), alpha 1, catalytic subunit
0.0061763	-2.5	IQGAP3	IQ motif containing GTPase activating protein 3

5.3.3 Dataset 3: The total number of subjects was 55 healthy males and females of various ages, young (16 male, 15 female) and old (12 male and 12 female).

A. Age-based differential gene expression in male

In this case study there are 28 individuals (16 young and 12 old) all of whom are male. In thi s case the true positives represent correctly classified young male, true negatives represent correctly classified old male, false positives represent the incorrectly classified young male and false negatives represent incorrectly classified old male. Comparing young males with old males, Table 5.16 shows the classification performance of the genes that are identified by the MFMWS compared to the genes identified by the MFSWS. It is clearly that MFMWS achieved the best performance with approximately 100%, while MFSWS achieved only 96%.

Table 5.16 Classification performance: old males versus young males

Classifier	**Accuracy**		**Sensitivity**		**Specificity**	
	MFMWS	**MFSWS**	***MFMWS***	***MFSWS***	**MFMWS**	**MFSWS**
1NN	100%	96%	100%	95%	100%	100%
SVM	100%	96%	100%	95%	100%	100%

Table 5.17 Confusion matrix : young males versus old males

16	0
0	12

Based on the figure 5.8. It observed that MFMWS achieved classification performance better than MFSWS with approximately 100%, this due to the smaller number of genes in this comparison.

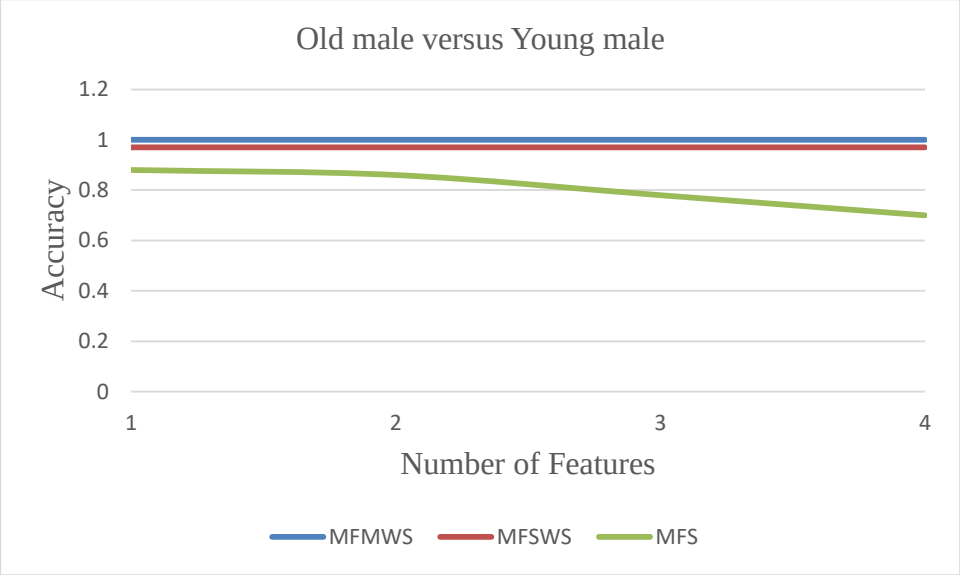

Figure. 5.8. Performance of the system using varying number of genes (old male versus young male).

Biological interpretation show that the obtained genes were consistently represented two overriding biological theme mitochondrial function, gene transcription. In addition these genes were in down-regulated directions with lower p-values and reasonable FC > 1.5 as shown in table 5.18. it can be said that some of these genes is effected due to the mitochondrial dysfunction which leads to muscle atrophy. Therefore these genes consider sarcopenia-related genes in males.

Table 5.18 P-values and FC of genes in males

P-value	Fold change (FC)	Gene Symbol	Gene Description
0.00179627	1.7	PAX8	Paired box 8
0.00693422	-1.9	DDR1	Discoidin domain receptor tyrosine kinase 1
0.00124052	-1.7	UNC13C	Unc-13 homolog C (C. elegans)
0.081856589	1.1	CYP2E1	Cytochrome P450, family 2, subfamily E, polypeptide 1

B. Age-based differential gene expression in female

In this study. This case study includes 27 adults (17 young females versus 12 old females). The true positives represent the correctly classified young females, true negatives represent correctly classified old females, false positive represent the incorrectly classified young females and false negatives represent incorrectly classified old females.

The results of the comparison between old and young females revealed the classification performance of the genes obtained by the MFMWS are outperforms with approximately 100% compared to those identified by MFSWS as shown in table 5.19.

Table 5.19 young females versus old females

Classifier	Accuracy		Sensitivity		Specificity	
	MFMWS	**MFSWS**	***MFMWS***	**MFSWS**	**MFMWS**	**MFSWS**
3NN	100%	96%	100%	95%	100%	100%
SVM	100%	96%	100%	96%	100%	100%

Table 5.20 Confusion matrix : young females versus old females

17	0
0	12

Figure 9.5 shows the classification performance for both systems using several genes, and it is clearly that MFMWS has achieved the best classification accuracy of 100% at all genes, this due to the smaller number of genes in this comparison.

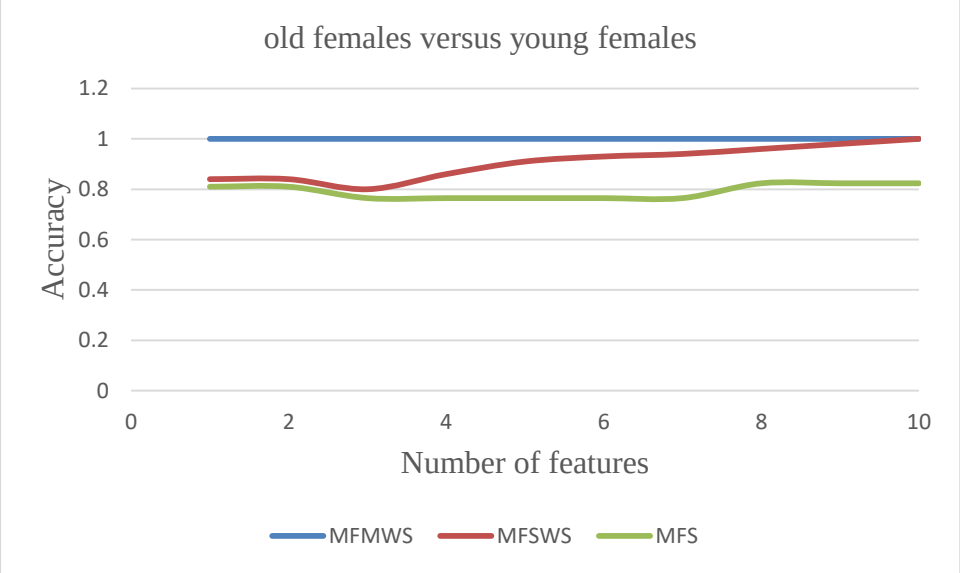

Figure. 5.9. Performance of the system using varying numbers of genes (old female versus young female).

Further analyses was conducted of these genes using p-value and fold change (FC) tests. The results showed that these genes have very low p-values and higher FC as show in table 5.21. Moreover the direction was in down-regulated. And they consistently represented three overriding biological theme including mitochondrial function. Therefore this subset of genes consider sarcopenia-related gens in females.

Table 5.21 P-values and FC of some of genes in female with brief descriptions.

P-value	Fold change FC	Gene Symbol	Gene Description
0.0001452	-8.9	MCL1	Myeloid cell leukemia sequence 1 (BCL2-related)
.00052121	-8.3	NOL3	Nucleolar protein 3 (apoptosis repressor with CARD domain)
0.0000457	-7.3	PTCD3	Pentatricopeptide repeat domain 3
0.00166145	-6.2	SPATA18	Spermatogenesis associated 18
0.0008452	-10.2	TNRC6A	Trinucleotide repeat containing 6A
0.00006215	-8.4	LOC100507237	Transcribed locus
0.00134365	-11.7	CHPT1	Choline phosphotransferase 1

C. Sex-based differential gene expression in older adults

In this case of study there are12 individuals (12 old males and 12 old females). In this case the true positives represent the correctly classified old males, true negatives represent correctly classified old females, false positives represent the incorrectly classified old males and false negative represent incorrectly classified old females. In comparison of old females and old males Table 5.22 shows the classification performance of the genes

discovered by the MFMWS compared to those identified by the MFSWS. Results indicate that MFMWS achieve the classification performance better than MFSWS using the 3NN and SVM classifiers with accuracy at 100%.

Table 5.22 old males versus old females

Classifier	**Accuracy**		**Sensitivity**		**Specificity**	
	MFMWS	**MFSWS**	***MFMWS***	**MFSWS**	**MFMWS**	**MFSWS**
3NN	100%	96%	100%	92%	100%	100%
SVM	100%	96%	100%	93%	100%	100%

Table 5.23 Confusion matrix : old male versus old female

12	0
0	12

According to figure 5.10, it observed that the classification Performance achieved by the MFMWS has high performance in the first 4 genes compared to the MFSWS. This means that the subset of genes which obtained by MFMWS is more reliable than counterparts in MFSWS.

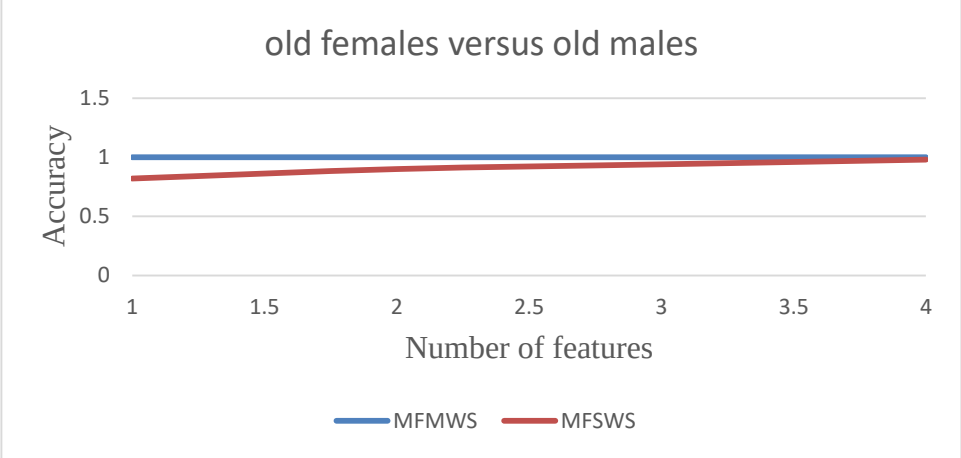

Figure. 5.10. Graph showing performance of the system using varying number of genes (old female versus old male).

The P-value and fold change (FC) tests have been applied to this subset of genes, and the results indicated that these have very low p-value and high FC > 1.5 as shown in table 5.24. In addition these genes were down-regulated they consistently represented three overriding biological theme mitochondrial structure and function. Therefore these genes are sarcopenia-related genes in both males and females.

Table 5.24 P-values and FC of genes in adults human with brief descriptions

P-value	Fold change FC	Gene Symbol	Gene Description
0.000215	-4.12	AKAP1	A kinase (PRKA) anchor protein 1
0.001542	-9.23	CD24	CD24 molecule
0.655312	-4.86	GAPT	GRB2-binding adaptor protein, transmembrane
0.00012	-9.25	FAM105B	Family with sequence similarity 105, member B

5.4 Conclusion

In the current study, a combination of machine learning techniques such as feature selection methods including RSFS and SVM-RFE based on the multi-filter multi-wrapper system (MFMWS), statistical methods of P-value and fold change (FC) and functional analyses using DAVID software were proposed. The aim was to identify different and common sarcopenia-related genes in males and females based on human skeletal muscle. This framework is designed to select a significant atrophy-related genes for males and females using skeletal muscle. Firstly genes are sorted using three different ranking techniques (t-test, Entropy and ROC-AUC) and then multi wrapper feature selection based on random subset feature selection and SVM- recessive feature elimination SVM-REF have been investigated to further reduce the dimensionality or the subset of genes. The proposed system is then evaluated on publicly available microarray datasets of the gene expression of skeletal muscle tissue. The final set of genes in each case of study shows high performance of classification in very small P-value and in term of biology most of genes are associated with older females and old males suggest that, they are more predisposed to muscle degeneration.. in some cases results revealed that there are common sarcopenia-related genes between males and females. Further pathway analysis showed that the biomarkers are associated mitochondrial dysfunction which are consistent with previous literature. So this means that the final list of genes comprises a promising set of sarcopenia biomarkers. The results show that this new approach multi-filter multi-wrapper system (MFMWS) succeeded in biomarker discovery associated with sarcopenia.

Chapter 6

Conclusion and future work

This chapter summarises the contributions of this thesis, draws conclusions from this work and discusses potential future work. Section 6.1 presents a thesis summary and section 6.2 presents the summary of the contributions. Challenges are discussed in section 6.4, section 6.3 presents difficulties while section 6.5 discusses future work

6.1 Summary of the contributions

The importance of the proposed approach is summarised in the following points. Firstly, it enables selection of significant genes associated with age and related to skeletal muscle mass loss. Secondly, the proposed system achieves a high level of accuracy in relation to the obtained genes. Lastly, the multi-wrapper methodology used in the proposed approach leads to an improvement in the system's ability to select more reliable age-related genes without affecting performance accuracy.

The combination of a multi-filter multi-wrapper system, statistical methods and functional analyses makes our approach a valuable contribution towards discovering important biomarkers associated with age related skeletal muscle mass loss in both males and females.

In relation to the first contribution, the aim was to develop framework which is able to overcome the shortcoming of using single evaluation in order to select most important genes related to muscle atrophy. The proposed multi-filter system has been designed based on three different evaluation methods, including t-test, entropy and ROC-AUC. Implemented on

different datasets, the results indicate that the proposed system achieves considerable efficiency in relation to selecting age-related genes compared to the exciting systems. Moreover, the results show that the MFS has the ability to select subsets of genes with high classification performance compared to previous studies, such as Liu et al. (2012) and Raue et al. (2013).

The second contribution is related to the multi-filter single wrapper system. This system is proposed to improve the performance accuracy of selecting the most reliable genes associated with age. The MFSWS was designed and applied on different microarray datasets. The selected genes underwent further statistical tests such as P-value and fold change. The evaluation results of the presented approach show that the MFSWS genes attained a high level of accuracy compared to existing studies which used the same microarray datasets.

The third contribution in order to find the distinct subset of genes related to muscle atrophy across both genders. A novel framework has been proposed and a new algorithm has been designed and implemented on three different datasets. In addition, the final list of genes underwent further analysis based on functional analysis. The results indicate that the MFMWS achieves a high level of accuracy compared with previously proposed systems and it has the ability to discover distinct subsets of genes compared to existing works.

6.2 Limitations

Although there has been a number of studies conducted in this area of selecting age-related genes, such as Liu et al. (2013) and Raue et al. (2012), the main issue related to these studies is that important genes are identified using whole training data. This can lead to poor generalization because one of the fundamental goals of machine learning is to generalize beyond samples in training data. In addition, microarray data normally includes a large

number of genes from a low number of samples [7]. However, not all genes in microarray data can help improve the classification accuracy because some of these genes are irrelevant or redundant.

In order to remove the irrelevant and redundant genes and at the same time identify a distinct subset of genes with high accuracy that can be used as a biomarker to assess disease risk [6], the first possible solution is to reduce the size of the dataset using methods such as feature reduction and filter methods that can be used to select genes based on their discriminative powers, such as t-test, ROC, entropy, t-statistics (TS), f-test and signal-to-noise ratio (SNR). Some of these previous methods have been shown to be effective in measuring the discriminative power of genes. Therefore, the presented framework adopts a multi-filter system based on majority voting among three evaluation methods, namely t-test, ROC and entropy, to reduce the size of the dataset with high classification accuracy.

Secondly, although gene selection using evaluation methods is fast and simple, it has some shortcomings, as follows:

1 – Evaluation methods ignore the interaction with the classifier. They are often univariate or low-variate. This means that each feature is considered separately, thereby ignoring feature dependencies, which may lead to worse classification performance when compared to other types of feature selection techniques. Therefore, the wrapper feature selection approach is essential to overcome the above shortcomings of evaluation methods because it consider the interaction between features.

Thirdly, each wrapper feature selection method can lead to different results. For example, if different wrapper methods are applied on the same subset of genes, they may all identify different subsets of genes with different levels of accuracy. Therefore, the proposed

framework multi-filter multi-wrapper system has help improve the performance accuracy and discover distinct subsets of genes associated with age.

6.3 Future research directions

Our gene selection approach is be able to discover very important subsets of genes associated with age, which may contribute towards the preservation of muscle atrophy and avoid strength decrease in the elderly. However, this reduction in muscle size is not fully illustrated by muscle mass loss alone.

- **Suggestions**

We therefore suggest another assumption to consider muscle quality and quantity. Although our proposed approach is able to discover very important subsets of genes, these genes consider as a target of future research with the aim of understanding aging biomarker mechanisms. It may provide assistance to the development of emerging disciplines related to computational biology for the elderly, leading to the provision of an overview about sarcopenia in the elderly.

- **Recommendations**

We recommend that more investigation into the molecular levels of aging is undertaken, as this could lead to a greater understanding of the aging process Also using other features selection such as Maximum relevance & minimum redundancy (mRMR), RELIEF and GA in addition using other evaluation methods such as Wilcox, Bhattacharyya and f-score moreover using other classifies such as Decision tree, Linear regression, Naïve Bayes, Neural networks, Logistic regression, Perceptron. Using all of them provide a more accurate insight on the impact of age on gene function.

6.4 List of publications

1 – Dreder, A., M. A. Tahir, H. Seker and M. N. Anwar (2016). "Majority voting approach for the identification of differentially expressed genes to understand gender- related skeletal muscle aging." Computer science, Engineering and Information system CCSEIT 6(1): 237-244.

2- Dreder, A., M. Tahir, H. Seker and N. Anwar (2016). "Discovering differences in gender-related skeletal muscle aging through the majority voting-based identification of differently expressed genes." International Journal on Bioinformatics & Biosciences 6(2): 1-14.

References

Abeel, T., T. Helleputte, Y. Van de Peer, P. Dupont and Y. Saeys (2010). "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods." Bioinformatics 26(3): 392-398.

Acharya, U. R., E. Ng, L. W. J. Eugene, K. P. Noronha, L. C. Min, K. P. Nayak and S. V. Bhandary (2015). "Decision support system for the glaucoma using Gabor transformation." Biomedical Signal Processing and Control 15: 18-26.

Alon, U., N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack and A. J. Levine (1999). "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays." Proceedings of the National Academy of Sciences 96(12): 6745-6750.

Altidor, W., T. M. Khoshgoftaar, J. Van Hulse and A. Napolitano (2011). Ensemble feature ranking methods for data intensive computing applications. Handbook of data intensive computing, Springer: 349-376.

Andy P. Field. Analysis of Variance (ANOVA), pages 33–36. SAGE Publications, Inc., 1 edition, 2007

Antonov, A., M. Krestyaninova, R. Knight, I. Rodchenkov, G. Melino and N. Barlev (2014). "PPISURV: a novel bioinformatics tool for uncovering the hidden role of specific genes in cancer survival outcome." Oncogene 33(13): 1621-1628.

Armstrong, S. A., J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub and S. J. Korsmeyer (2002). "MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia." Nature genetics 30(1): 41-47.

Babu, M. M. (2004). "Introduction to microarray data analysis." Computational genomics: Theory and application: 225-249.

Bauer, J. M., Verlaan, S., Bautmans, I., Brandt, K., Donini, L. M., Maggio, M., ... & Ceda, G. P. (2015). Effects of a vitamin D and leucine-enriched whey protein nutritional supplement on measures of sarcopenia in older adults, the PROVIDE study: a randomized, double-blind, placebo-controlled trial. Journal of the American Medical Directors Association, 16(9), 740-747.

Bell, K. E., von Allmen, M. T., Devries, M. C., & Phillips, S. M. (2016). Muscle disuse as a pivotal problem in sarcopenia-related muscle loss and dysfunction. J Frailty Aging, 5(1), 33-41

Blanco, R., P. Larrañaga, I. Inza and B. Sierra (2004). "Gene selection for cancer classification using wrapper approaches." International Journal of Pattern Recognition and Artificial Intelligence 18(08): 1373-1390.

Blum, A. L. and P. Langley (1997). "Selection of relevant features and examples in machine learning." Artificial intelligence 97(1): 245-271.

Bodine, S. C., E. Latres, S. Baumhueter, V. K.-M. Lai, L. Nunez, B. A. Clarke, W. T. Poueymirou, F. J. Panaro, E. Na and K. Dharmarajan (2001). "Identification of ubiquitin ligases required for skeletal muscle atrophy." Science 294(5547): 1704-1708.

Box, J. F. (1987). "Guinness, Gosset, Fisher, and small samples." Statistical science: 45-52.

Breiman, L. (2001). "Random forests." Machine learning 45(1): 5-32.

Breiman, L., J. Friedman, C. J. Stone and R. A. Olshen (1984). Classification and regression trees, CRC press.

Bresell, A. (2008). "Characterization of protein families, sequence patterns, and functional annotations in large data sets." Thesis

Brown, M. and E. M. Hasser (1996). "Complexity of age-related change in skeletal muscle." The Journals of Gerontology Series A: Biological Sciences and Medical Sciences 51(2): B117-B123.

Brown, M. P., W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares and D. Haussler (2000). "Knowledge-based analysis of microarray gene expression data by using support vector machines." Proceedings of the National Academy of Sciences 97(1): 262-267.

Cadenas, J. M., M. C. Garrido and R. MartíNez (2013). "Feature subset selection filter–wrapper based on low quality data." Expert systems with applications 40(16): 6241-6252.

Cai, D., J. D. Frantz, N. E. Tawa, P. A. Melendez, B.-C. Oh, H. G. Lidov, P.-O. Hasselgren, W. R. Frontera, J. Lee and D. J. Glass (2004). "IKKβ/NF-κB activation causes severe muscle wasting in mice." Cell 119(2): 285-298.

Cai, D., M. Yuan, D. F. Frantz, P. A. Melendez, L. Hansen, J. Lee and S. E. Shoelson (2005). "Local and systemic insulin resistance resulting from hepatic activation of IKK-β and NF-κB." Nature medicine 11(2): 183-190.

Calle, E. E., M. J. Thun, J. M. Petrelli, C. Rodriguez and C. W. Heath Jr (1999). "Body-mass index and mortality in a prospective cohort of US adults." New England Journal of Medicine 341(15): 1097-1105.

Calvanese, V., E. Lara, A. Kahn and M. F. Fraga (2009). "The role of epigenetics in aging and age-related diseases." Ageing research reviews 8(4): 268-276.

Carter, H. N., C. C. Chen and D. A. Hood (2015). "Mitochondria, muscle health, and exercise with advancing age." Physiology 30(3): 208-223.

Cawley, G. C. and N. L. Talbot (2003). "Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers." Pattern Recognition 36(11): 2585-2592.

Cesari, M., F. Landi, B. Vellas, R. Bernabei and E. Marzetti (2014). "Sarcopenia and physical frailty: two sides of the same coin." Pathophysiological Mechanisms of Sarcopenia in Aging and in Muscular Dystrophy: A Translational Approach. 6:192.

Chabi, B., P. J. Adhihetty, V. Ljubicic and D. A. Hood (2005). "How is mitochondrial biogenesis affected in mitochondrial disease?" Medicine and science in sports and exercise 37(12): 2102.

Chen, X.W. and Wasikowski, M., 2008, August. Fast: a roc-based feature selection metric for small samples and imbalanced data classification problems. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 124-132). ACM

Cheng, T., P. Lin, L. Huang, Y. Wu, S. Jin, C. Liu and Q. Xia (2016). "Genome-Wide Analysis of Host Responses to Four Different Types of Microorganisms in Bombyx Mori (Lepidoptera: Bombycidae)." Journal of Insect Science 16(1): 69.

Chesley, A., J. MacDougall, M. Tarnopolsky, S. Atkinson and K. Smith (1992). "Changes in human muscle protein synthesis after resistance exercise." Journal of applied physiology 73(4): 1383-1388.

Cho, S.-B. and H.-H. Won (2003). Machine learning in DNA microarray analysis for cancer classification. Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics 2003-Volume 19, Australian Computer Society, Inc. (pp. 189-198)

Chow, M., E. Moler and I. Mian (2001). "Identifying marker genes in transcription profiling data using a mixture of feature relevance experts." Physiological genomics 5(2): 99-111.

Colak, S. and C. Isik (2003). Feature subset selection for blood pressure classification using orthogonal forward selection. Bioengineering Conference, 2003 IEEE 29th Annual, Proceedings of, IEEE. pp. 122-123)

Cotter, S. F., K. Kreutz-Delgado and B. D. Rao (2001). "Backward sequential elimination for sparse vector subset selection." Signal Processing 81(9): 1849-1864.

Dalma-Weiszhausz, D. D., J. Warrington, E. Y. Tanimoto and C. G. Miyada (2006). "[1] The Affymetrix GeneChip® Platform: An Overview." Methods in enzymology 410: 3-28.

D'Antona, G., M. A. Pellegrino, R. Adami, R. Rossi, C. N. Carlizzi, M. Canepari, B. Saltin and R. Bottinelli (2003). "The effect of ageing and immobilization on structure and function of human skeletal muscle fibres." The Journal of physiology 552(2): 499-511.

Day, K., L. L. Waite, A. Thalacker-Mercer, A. West, M. M. Bamman, J. D. Brooks, R. M. Myers and D. Absher (2013). "Differential DNA methylation with age displays both common and dynamic features across human tissues that are influenced by CpG landscape." Genome biology 14(9): 1.

De Magalhães, J. P., J. Curado and G. M. Church (2009). "Meta-analysis of age-related gene expression profiles identifies common signatures of aging." Bioinformatics 25(7): 875-881.

Dennis, R. A., B. Przybyla, C. Gurley, P. M. Kortebein, P. Simpson, D. H. Sullivan and C. A. Peterson (2008). "Aging alters gene expression of growth and remodeling factors in human

skeletal muscle both at rest and in response to acute resistance exercise." Physiological genomics 32(3): 393-400.

Deschenes, M. R. (2004). "Effects of aging on muscle fibre type and size." Sports Medicine 34(12): 809-824.

Ding, C. and H. Peng (2005). "Minimum redundancy feature selection from microarray gene expression data." Journal of bioinformatics and computational biology 3(02): 185-205.

Doherty, T. J. (2001). "The influence of aging and sex on skeletal muscle mass and strength." Current Opinion in Clinical Nutrition & Metabolic Care 4(6): 503-508.

Doherty, T. J., A. A. Vandervoort, A. W. Taylor and W. F. Brown (1993). "Effects of motor unit losses on strength in older men and women." Journal of Applied Physiology 74(2): 868-874.

Doncaster, C. P. and A. J. Davey (2007). Analysis of variance and covariance: how to choose and construct models for the life sciences, Cambridge University Press.

Dos Santos, E. M. and H. M. Gomes (2002). A comparative study of polynomial kernel SVM applied to appearance-based object recognition. Pattern Recognition with Support Vector Machines, Springer: 408-418.

Dreyer, H. C., S. Fujita, J. G. Cadenas, D. L. Chinkes, E. Volpi and B. B. Rasmussen (2006). "Resistance exercise increases AMPK activity and reduces 4E-BP1 phosphorylation and protein synthesis in human skeletal muscle." The Journal of physiology 576(2): 613-624.

Drummond, M. J., J. J. McCarthy, M. Sinha, H. M. Spratt, E. Volpi, K. A. Esser and B. B. Rasmussen (2011). "Aging and microRNA expression in human skeletal muscle: a microarray and bioinformatics analysis." Physiological genomics 43(10): 595-603.

Dubowitz, V. and A. E. Pearse (1960). "A comparative histochemical study of oxidative enzyme and phosphorylase activity in skeletal muscle." Histochemistry and Cell Biology 2(2): 105-117.

Dudoit, S., J. Fridlyand and T. P. Speed (2002). "Comparison of discrimination methods for the classification of tumors using gene expression data." Journal of the American statistical association 97(457): 77-87.

Durham, W. J., S. L. Miller, C. W. Yeckel, D. L. Chinkes, K. D. Tipton, B. B. Rasmussen and R. R. Wolfe (2004). "Leg glucose and protein metabolism during an acute bout of resistance exercise in humans." Journal of Applied Physiology 97(4): 1379-1386.

Enright, M. C., N. P. Day, C. E. Davies, S. J. Peacock and B. G. Spratt (2000). "Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones ofStaphylococcus aureus." Journal of clinical microbiology 38(3): 1008-1015.

Farooqi, M. and K. Raza (2012). "A Comprehensive Study of CRM through data mining Techniques." arXiv preprint arXiv:1205.1126.

Fawcett, T. (2006). "An introduction to ROC analysis." Pattern recognition letters 27(8): 861-874.

Flynn, M., G. Nolph, A. S. Baker, W. Martin and G. Krause (1989). "Total body potassium in aging humans: a longitudinal study." The American Journal of Clinical Nutrition 50(4): 713-717.

Furey, T. S., N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer and D. Haussler (2000). "Support vector machine classification and validation of cancer tissue samples using microarray expression data." Bioinformatics 16(10): 906-914.

Gauthier, G. F. and H. A. Padykula (1966). "CYTOLOGICAL STUDIES OF FIBER TYPES IN SKELETAL MUSCLE A Comparative Study of the Mammalian Diaphragm." The Journal of cell biology 28(2): 333-354.

George, G. and V. C. Raj (2011). "Review on feature selection techniques and the impact of SVM for cancer classification using gene expression profile." arXiv preprint arXiv:1109.1062.

Gheorghe, M., M. Snoeck, M. Emmerich, T. Bäck, J. J. Goeman and V. Raz (2014). "Major aging-associated RNA expressions change at two distinct age-positions." BMC genomics 15(1): 132.

Giresi, P. G., E. J. Stevenson, J. Theilhaber, A. Koncarevic, J. Parkington, R. A. Fielding and S. C. Kandarian (2005). "Identification of a molecular signature of sarcopenia." Physiological genomics 21(2): 253-263.

Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing and M. A. Caligiuri (1999). "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." science 286(5439): 531-537.

Goni, R., P. García and S. Foissac (2009). "The qPCR data statistical analysis." Integromics White Paper: 1-9.

Goodpaster, B. H., S. W. Park, T. B. Harris, S. B. Kritchevsky, M. Nevitt, A. V. Schwartz, E. M. Simonsick, F. A. Tylavsky, M. Visser and A. B. Newman (2006). "The loss of skeletal muscle strength, mass, and quality in older adults: the health, aging and body composition study." The Journals of Gerontology Series A: Biological Sciences and Medical Sciences 61(10): 1059-1064.

Gunavathi, C. and K. Premalatha (2014). "Performance analysis of genetic algorithm with KNN and SVM for Feature Selection in Tumor Classification." International Journal of Computer, Electrical, Automation, Control and Information Engineering 8(8): 1490-1497.

Gusev, Y. (2008). "Computational methods for analysis of cellular functions and pathways collectively targeted by differentially expressed microRNA." Methods **44**(1): 61-72.

Guyon, I. and A. Elisseeff (2003). "An introduction to variable and feature selection." Journal of machine learning research 3(Mar): 1157-1182.

Guyon, I., J. Weston, S. Barnhill and V. Vapnik (2002). "Gene selection for cancer classification using support vector machines." Machine learning 46(1-3): 389-422.

Han, J., J. Pei and M. Kamber (2011). Data mining: concepts and techniques, Elsevier.

Hangelbroek, R. W., P. Fazelzadeh, M. Tieland, M. V. Boekschoten, G. J. Hooiveld, J. P.

Duynhoven, J. A. Timmons, L. B. Verdijk, L. C. Groot and L. J. Loon (2016). "Expression of protocadherin gamma in skeletal muscle tissue is associated with age and muscle weakness." Journal of cachexia, sarcopenia and muscle. 5 (2016): 604-614.

Hanson, J. and H. E. Huxley (1953). "Structural basis of the cross-striations in muscle." Nature 172: 530-532.

Harber, M. P., A. R. Konopka, B. Jemiolo, S. W. Trappe, T. A. Trappe and P. T. Reidy (2010). "Muscle protein synthesis and gene expression during recovery from aerobic exercise in the fasted and fed states." American Journal of Physiology-Regulatory, Integrative and Comparative Physiology 299(5): R1254-R1262.

Haury, A.-C., P. Gestraud and J.-P. Vert (2011). "The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures." PloS one 6(12): e28210.

Hiesinger, P. R. and B. A. Hassan (2005). "Genetics in the age of systems biology." Cell 123(7): 1173-1174.

Hsu, C.-W. and C.-J. Lin (2002). "A comparison of methods for multiclass support vector machines." IEEE transactions on Neural Networks 13(2): 415-425.

Hu, H., J. Li, H. Wang and G. Daggard (2006). Combined gene selection methods for microarray data analysis. International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, vol. 4251, Springer. pp. 976-983

Huang, J., W. Wei, J. Zhang, G. Liu, G. R. Bignell, M. R. Stratton, P. A. Futreal, R. Wooster, K. W. Jones and M. H. Shapero (2004). "Whole genome DNA copy number changes identified by high density oligonucleotide arrays." Human genomics 1(4): 1.

Huang, J., Y. Cai and X. Xu (2007). "A hybrid genetic algorithm for feature selection wrapper based on mutual information." Pattern Recognition Letters 28(13): 1825-1844.

Hughes, V. A., W. R. Frontera, R. Roubenoff, W. J. Evans and M. A. F. Singh (2002). "Longitudinal changes in body composition in older men and women: role of body weight change and physical activity." The American journal of clinical nutrition 76(2): 473-481.

Huang, D. W., B. T. Sherman, Q. Tan, J. Kir, D. Liu, D. Bryant, Y. Guo, R. Stephens, M. W. Baseler and H. C. Lane (2007). "DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists." Nucleic acids research 35(suppl 2): W169-W175.

Huang, C., X. Xiao, N. R. Chintagari, M. Breshears, Y. Wang and L. Liu (2014). "MicroRNA and mRNA expression profiling in rat acute respiratory distress syndrome." BMC medical genomics **7**(1): 46.

Inza, I., P. Larrañaga, R. Blanco and A. J. Cerrolaza (2004). "Filter versus wrapper gene selection approaches in DNA microarray domains." Artificial intelligence in medicine 31(2): 91-103.

Jakobsdottir, J., M. B. Gorin, Y. P. Conley, R. E. Ferrell and D. E. Weeks (2009). "Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers." PLoS Genet 5(2): e1000337.

Jaksik, R., M. Iwanaszko, J. Rzeszowska-Wolny and M. Kimmel (2015). "Microarray experiments and factors which affect their reliability." Biology direct 10(1): 46.

Janssen, I. (2011). "The epidemiology of sarcopenia." Clinics in geriatric medicine 27(3): 355-363.

Janssen, I., S. B. Heymsfield, Z. Wang and R. Ross (2000). "Skeletal muscle mass and distribution in 468 men and women aged 18–88 yr." Journal of applied physiology 89(1): 81-88.

Jirapech-Umpai, T. and S. Aitken (2005). "Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes." BMC bioinformatics 6(1): 1.

JOACHIMS, T. (2002). Learning to Classify Text Using Support Vector Machines: Methods, Theory, and Algorithms. Kluwer Academic, Boston, 655-664

Kalyani, R. R., M. Corriere and L. Ferrucci (2014). "Age-related and disease-related muscle loss: the effect of diabetes, obesity, and other diseases." The lancet Diabetes & endocrinology 2(10): 819-829.

Kashyap, H., H. A. Ahmed, N. Hoque, S. Roy and D. K. Bhattacharyya (2015). "Big Data Analytics in Bioinformatics: A Machine Learning Perspective." arXiv preprint arXiv:1506.05101.

Keller, K. and M. Engelhardt (2013). "Strength and muscle mass loss with aging process. Age and strength loss." Muscles, ligaments and tendons journal 3(4): 346.

Khamseh, M. E., M. Malek, R. Aghili and Z. Emami (2011). "Sarcopenia and diabetes: pathogenesis and consequences." The British Journal of Diabetes & Vascular Disease 11(5): 230-234.

Klitgaard, H., M. Mantoni, S. Schiaffino, S. Ausoni, L. Gorza, C. Laurent-Winter, P. Schnohr and B. Saltin (1990). "Function, morphology and protein expression of ageing skeletal muscle: a cross-sectional study of elderly men with different training backgrounds." Acta Physiologica Scandinavica 140(1): 41-54.

Kodratoff, Y. (2014). Introduction to machine learning, Morgan Kaufmann.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Ijcai. vol. 14, no. 2, pp. 1137-1145

Kohavi, R. and G. H. John (1997). "Wrappers for feature subset selection." Artificial intelligence 97(1): 273-324.

Kononenko, I. (1994). Estimating attributes: analysis and extensions of RELIEF. European conference on machine learning, Springer. ECML-94. pp. 171–182

Krishnapuram, B., L. Carin and A. Hartemink (2004). "1 Gene expression analysis: Joint feature selection and classifier design." Kernel Methods in Computational Biology: 299-317.

Landers, K. A., G. R. Hunter, C. J. Wetzstein, M. M. Bamman and R. L. Weinsier (2001). "The interrelationship among muscle mass, strength, and the ability to perform physical tasks of daily living in younger and older women." The Journals of Gerontology Series A: Biological Sciences and Medical Sciences 56(10): B443-B448.

Lazar, C., J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. de Schaetzen, R. Duque, H. Bersini and A. Nowe (2012). "A survey on filter techniques for feature selection in gene expression microarray analysis." IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 9(4): 1106-1119.

Lenoir, T. and E. Giannella (2006). "The emergence and diffusion of DNA microarray technology." Journal of biomedical discovery and collaboration 1(1): 1

Li, L., T. A. Darden, C. Weingberg, A. Levine and L. G. Pedersen (2001). "Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method." Combinatorial chemistry & high throughput screening 4(8): 727-739.

Li, S., E. J. Harner and D. A. Adjeroh (2011). "Random KNN feature selection-a fast and stable alternative to Random Forests." BMC bioinformatics 12(1): 1.

Liu, D., M. A. Sartor, G. A. Nader, L. Gutmann, M. K. Treutelaar, E. E. Pistilli, H. B. IglayReger, C. F. Burant, E. P. Hoffman and P. M. Gordon (2010). "Skeletal muscle gene expression in response to resistance exercise: sex specific regulation." BMC genomics 11(1): 659.

Liu, D., M. A. Sartor, G. A.Nader, E. E. Pistilli, L. Tanton, C. Lilly, L. Gutmann, H. B. IglayReger, P. S. Visich, E. P. Hoffman and P. M. Gordon (2013). Microarray Analysis Reveals Novel Features of the Muscle Aging Process in Men and Women. Journals of Gerontology: Biological Sciences. 68(9): 1035–1044.

Liu, H., J. Li and L. Wong (2004). Selection of patient samples and genes for outcome prediction. Computational Systems Bioinformatics Conference, 2004. CSB 2004. Proceedings. 2004 IEEE, IEEE, pp. 382-392

Losonczy, K. G., T. B. Harris, J. Cornoni-Huntley, E. M. Simonsick, R. B. Wallace, N. R. Cook, A. M. Ostfeld and D. G. Blazer (1995). "Does weight loss from middle age to old age explain the inverse weight mortality relation in old age?" American Journal of Epidemiology 141(4): 312-321.

Lu, Q. and R. C. Elston (2008). "Using the optimal receiver operating characteristic curve to design a predictive genetic test, exemplified with type 2 diabetes." The American Journal of Human Genetics 82(3): 641-651.

Luscombe, N. M., D. Greenbaum and M. Gerstein (2001). "What is bioinformatics? An introduction and overview." Yearbook of Medical Informatics 1: 83-99.

Lynch, N., E. Metter, R. Lindle, J. Fozard, J. Tobin, T. Roy, J. Fleg and B. Hurley (1999). "Muscle quality. I. Age-associated differences between arm and leg muscle groups." Journal of Applied Physiology 86(1): 188-194.

MacDougall, J. D., M. J. Gibala, M. A. Tarnopolsky, J. R. MacDonald, S. A. Interisano and K. E. Yarasheski (1995). "The time course for elevated muscle protein synthesis following heavy resistance exercise." Canadian Journal of applied physiology 20(4): 480-486.

Masiero, E., L. Agatea, C. Mammucari, B. Blaauw, E. Loro, M. Komatsu, D. Metzger, C. Reggiani, S. Schiaffino and M. Sandri (2009). "Autophagy is required to maintain muscle mass." Cell metabolism 10(6): 507-515.

Melov, S., M. A. Tarnopolsky, K. Beckman, K. Felkey and A. Hubbard (2007). "Resistance exercise reverses aging in human skeletal muscle." PLoS One 2(5): e465.

Metter, E. J., N. Lynch, R. Conwit, R. Lindle, J. Tobin and B. Hurley (1999). "Muscle quality and age: cross-sectional and longitudinal comparisons." The Journals of Gerontology Series A: Biological Sciences and Medical Sciences 54(5): B207-B218.

Mishra, D. and B. Sahu (2011). "Feature selection for cancer classification: a signal-to-noise ratio approach." International Journal of Scientific & Engineering Research 2(4): 1-7.

Mitchell, T. M. (1997). Machine learning. WCB, McGraw-Hill Boston, MA.

Mootha, V. K., C. M. Lindgren, K.-F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstråle and E. Laurila (2003). "PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes." Nature genetics 34(3): 267-273.

Murphy, K. P. (2012). Machine learning: a probabilistic perspective, MIT press.

Newman, A. B., V. Kupelian, M. Visser, E. M. Simonsick, B. H. Goodpaster, S. B. Kritchevsky, F. A. Tylavsky, S. M. Rubin and T. B. Harris (2006). "Strength, but not muscle mass, is associated with mortality in the health, aging and body composition study cohort." The Journals of Gerontology Series A: Biological Sciences and Medical Sciences 61(1): 72-77.

NCBI (2013).' GEO Dataset Browser' Available at: https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4858

NCBI (2012).' GEO Dataset Browser' Available at: https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS5216

NCBI (2012).' GEO Dataset Browser' Available at: https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS5218

Novaković, J. (2016). Toward optimal feature selection using ranking methods and classification algorithms. Yugoslav Journal of Operations Research, 21(1).

Nyberg, K. (2011). "Document Classification Using Machine Learning and Ontologies." Aalto University, School of Science, Master's Thesis.

Ogborn, D. I., B. R. McKay, J. D. Crane, A. Safdar, M. Akhtar, G. Parise and M. A. Tarnopolsky (2015). "Effects of age and unaccustomed resistance exercise on mitochondrial transcript and protein abundance in skeletal muscle of men." American Journal of Physiology-Regulatory, Integrative and Comparative Physiology 308(8): R734-R741

Ooi, C. and P. Tan (2003). "Genetic algorithms applied to multi-class prediction for the analysis of gene expression data." Bioinformatics 19(1): 37-44.

Ozer, H. G., J. D. Parvin and K. Huang (2012). "DFI: gene feature discovery in RNA-seq experiments from multiple sources." BMC genomics 13(8): 1.

Oyelade, O., J. Soyemi, I. Isewon and O. O. Obembe (2015). "Bioinformatics, healthcare informatics and analytics: an imperative for improved healthcare system." International Journal of Applied Information Systems 8(5): 1-6.

Park, S. W., B. H. Goodpaster, E. S. Strotmeyer, N. de Rekeneire, T. B. Harris, A. V. Schwartz, F. A. Tylavsky and A. B. Newman (2006). "Decreased muscle strength and quality in older adults with type 2 diabetes The Health, Aging, and Body Composition Study." Diabetes 55(6): 1813-1818.

Patti, M. E., A. J. Butte, S. Crunkhorn, K. Cusi, R. Berria, S. Kashyap, Y. Miyazaki, I. Kohane, M. Costello and R. Saccone (2003). "Coordinated reduction of genes of oxidative metabolism in humans with insulin resistance and diabetes: Potential role of PGC1 and NRF1." Proceedings of the National Academy of Sciences 100(14): 8466-8471.

Peng, S., Q. Xu, X. B. Ling, X. Peng, W. Du and L. Chen (2003). "Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines." FEBS letters 555(2): 358-362.

Powers, S. K., M. P. Wiggs, J. A. Duarte, A. M. Zergeroglu and H. A. Demirel (2012). "Mitochondrial signaling contributes to disuse muscle atrophy." American Journal of Physiology-Endocrinology and Metabolism 303(1): E31-E39.

Pranckeviciene, A. and A. Bunevicius (2015). "Depression screening in patients with brain tumors: a review." 4:71-78.

Pranckeviciene, E. (2015). Bioinformatics Tools for the Analysis of Gene-Phenotype Relationships Coupled with a Next Generation ChIP-Sequencing Data Analysis Pipeline, Université d'Ottawa/University of Ottawa. Thesis

Pudil, P., J. Novovičová and J. Kittler (1994). "Floating search methods in feature selection." Pattern recognition letters 15(11): 1119-1125.

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann.

Quinlan, J.R., 1996. Improved use of continuous attributes in C4.5. J. Artificial Intell. Res. 4 (1), 77–90.

Rakotomamonjy, A. (2003). "Variable selection using SVM-based criteria." Journal of machine learning research 3(Mar): 1357-1370.

Ramaswamy, S., P. Tamayo, R. Rifkin, S. Mukherjee, C.-H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe and J. P. Mesirov (2001). "Multiclass cancer diagnosis using tumor gene expression signatures." Proceedings of the National Academy of Sciences 98(26): 15149-15154.

Rantanen, T., K. Masaki, D. Foley, G. Izmirlian, L. White and J. Guralnik (1998). "Grip strength changes over 27 yr in Japanese-American men." Journal of Applied Physiology 85(6): 2047-2053.

Räsänen, O. and J. Pohjalainen (2013). Random subset feature selection in automatic recognition of developmental disorders, affective states, and level of conflict from speech. Interspeech. pp. 210-214. 2013

Raue, U., D. Slivka, B. Jemiolo, C. Hollon and S. Trappe (2007). "Proteolytic gene expression differs at rest and after resistance exercise between young and old women." The Journals of Gerontology Series A: Biological Sciences and Medical Sciences 62(12): 1407-1412.

Raza, K. (2012). "Application of data mining in bioinformatics." arXiv preprint arXiv:1205.1125 Vol 1 No 2, 114-118

Reunanen, J. (2003). "Overfitting in making comparisons between variable selection methods." Journal of Machine Learning Research 3(Mar): 1371-1382.

Rivas, D. A., S. J. Lessard, N. P. Rice, M. S. Lustgarten, K. So, L. J. Goodyear, L. D. Parnell and R. A. Fielding (2014). "Diminished skeletal muscle microRNA expression with aging is associated with attenuated muscle plasticity and inhibition of IGF-1 signaling." The FASEB Journal 28(9): 4133-4147.

Rosenberg, I. H. (1989). "Summary comments." The American journal of clinical nutrition 50(5): 1231-1233.

Rosenberg, I. H. (1989). "Summary comments." The American journal of clinical nutrition 50(5): 1231-1233.

Roth, S. M., R. E. Ferrell, D. G. Peters, E. J. Metter, B. F. Hurley and M. A. Rogers (2002). "Influence of age, sex, and strength training on human muscle gene expression determined by microarray." Physiological genomics 10(3): 181-190.

Ruiz, R., J. C. Riquelme and J. S. Aguilar-Ruiz (2006). "Incremental wrapper-based gene selection from microarray data for cancer classification." Pattern Recognition 39(12): 2383-2392.

Russo, S. Bandinelli, B. Bartali, C. Cavazzini, A. Di Iorio, A. M. Corsi, T. Rantanen, J. M. Guralnik and L. Ferrucci (2003). "Age-associated changes in skeletal muscles and their effect on mobility: an operational diagnosis of sarcopenia." Journal of applied physiology 95(5): 1851-1860.

Sacheck, J. M., J.-P. K. Hyatt, A. Raffaello, R. T. Jagoe, R. R. Roy, V. R. Edgerton, S. H. Lecker and A. L. Goldberg (2007). "Rapid disuse and denervation atrophy involve transcriptional changes similar to those of muscle wasting during systemic diseases." The FASEB Journal 21(1): 140-155.

Saeys, Y., T. Abeel and Y. Van de Peer (2008). Robust feature selection using ensemble feature selection techniques. Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer. pp. 313–325

Şahan, S., K. Polat, H. Kodaz and S. Güneş (2007). "A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis." Computers in Biology and Medicine 37(3): 415-423.

Sánchez, A. and M. C. R. de Villa (2008). "A tutorial review of microarray data analysis." Bioinfotmatics Tutorial, Universitat de Barcelona. 1–55

Sandri, M. (2010). "Autophagy in skeletal muscle." FEBS letters 584(7): 1411-1416.

Sandri, M., C. Sandri, A. Gilbert, C. Skurk, E. Calabria, A. Picard, K. Walsh, S. Schiaffino, S. H. Lecker and A. L. Goldberg (2004). "Foxo transcription factors induce the

atrophy-related ubiquitin ligase atrogin-1 and cause skeletal muscle atrophy." Cell 117(3): 399-412.

Sasik, R., C. Woelk and J. Corbeil (2004). "Microarray truths and consequences." Journal of molecular endocrinology 33(1): 1-9.

SCHIAFFINO, S. & REGGIANI, C. 2011. Fiber types in mammalian skeletal muscles. Physiol Rev, 91, 1447-531.

Schiaffino, S. and C. Reggiani (2011). "Fiber types in mammalian skeletal muscles." Physiological reviews 91(4): 1447-1531.

SCHOLKOPF, B., Tsuda, K. and Vert, J.-P. (2004). Kernel Methods in Computational Biology. MIT Press, Cambridge, MA

Shafique, U. and H. Qaiser (2014). "A comparative study of data mining process models (KDD, CRISP-DM and SEMMA)." Int. J. Innov. Sci. Res 12(1): 217-222.

Sharma, R., R. B. Pachori and U. R. Acharya (2015). "An integrated index for the identification of focal electroencephalogram signals using discrete wavelet transform and entropy measures." Entropy 17(8): 5218-5240.

Shipp, M. A., K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich and G. S. Pinkus (2002). "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning." Nature medicine 8(1): 68-74.

Sifakis, E. G., I. Valavanis, O. Papadodima and A. A. Chatziioannou (2013). Identifying gender independent biomarkers responsible for human muscle aging using microarray data. Bioinformatics and Bioengineering (BIBE), 2013 IEEE 13th International Conference on, IEEE. 10 (pp. 1-5)

Singh, D., P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico and J. P. Richie (2002). "Gene expression correlates of clinical prostate cancer behavior." Cancer cell 1(2): 203-209..

Su, A. I., J. B. Welsh, L. M. Sapinoso, S. G. Kern, P. Dimitrov, H. Lapp, P. G. Schultz, S. M. Powell, C. A. Moskaluk and H. F. Frierson (2001). "Molecular classification of human carcinomas by use of gene expression signatures." Cancer research 61(20): 7388-7393.

Su, Y., T. Murali, V. Pavlovic, M. Schaffer and S. Kasif (2003). "RankGene: identification of diagnostic genes based on expression data." Bioinformatics 19(12): 1578-1579.

Tarca, A. L., R. Romero and S. Draghici (2006). "Analysis of microarray experiments of gene expression profiling." American journal of obstetrics and gynecology 195(2): 373-388.

Terrell, G. R. and D. W. Scott (1992). "Variable kernel density estimation." The Annals of Statistics: 1236-1265.

Tessler, L. A. and R. D. Mitra (2011). "Sensitive single-molecule protein quantification and protein complex detection in a microarray format." Proteomics 11(24): 4731-4735.

Thalacker-Mercer, A. E., L. J. Dell'Italia, X. Cui, J. M. Cross and M. M. Bamman (2010). "Differential genomic responses in old vs. young humans despite similar levels of modest muscle damage after resistance loading." Physiological genomics 40(3): 141-149.

Theilhaber, J., T. Connolly, S. Roman-Roman, S. Bushnell, A. Jackson, K. Call, T. Garcia and R. Baron (2002). "Finding genes in the C2C12 osteogenic pathway by k-nearest-neighbor classification of expression data." Genome research 12(1): 165-176.

Thomas, D. R. (2003). "The relationship between functional status and inflammatory disease in older adults." The Journals of Gerontology Series A: Biological Sciences and Medical Sciences 58(11): M995-M998.

Thomas, D. R. (2005). "Weight loss in older adults." Reviews in Endocrine and Metabolic Disorders 6(2): 129-136.

Thomas, D. R. (2007). "Loss of skeletal muscle mass in aging: examining the relationship of starvation, sarcopenia and cachexia." Clinical nutrition 26(4): 389-399.

Thouleimat, N., D. Hernandez-Lobato and P. Dupont (2010). Variance Estimators for t-Test Ranking Influence the Stability and Predictive Performance of Microarray Gene Signatures. European Conference on Computational Biology.

Tsai, M.-C. and Y. Park (2007)"Cancer classification using Machine Learning Technique on Microarray Data." Machine Learning

Tzankoff, S. P. and A. H. Norris (1978). "Longitudinal changes in basal metabolism in man." Journal of Applied Physiology 45(4): 536-539.

van der Net, J. B., A. C. J. Janssens, J. C. Defesche, J. J. Kastelein, E. J. Sijbrands and E. W. Steyerberg (2009). "Usefulness of genetic polymorphisms and conventional risk factors to predict coronary heart disease in patients with familial hypercholesterolemia." The American journal of cardiology 103(3): 375-380.

Van't Veer, L. J., H. Dai, M. J. Van De Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton and A. T. Witteveen (2002). "Gene expression profiling predicts clinical outcome of breast cancer." nature 415(6871): 530-536.

Vermeulen, A., J. Kaufman and V. Giagulli (1996). "Influence of some biological indexes on sex hormone-binding globulin and androgen levels in aging or obese males." The Journal of Clinical Endocrinology & Metabolism 81(5): 1821-1826.

Visser, M., M. Pahor, F. Tylavsky, S. B. Kritchevsky, J. A. Cauley, A. B. Newman, B. A. Blunt and T. B. Harris (2003). "One-and two-year change in body composition as measured by DXA in a population-based cohort of older men and women." Journal of applied physiology 94(6): 2368-2374.

Volpi, E., M. Sheffield-Moore, B. B. Rasmussen and R. R. Wolfe (2001). "Basal muscle amino acid kinetics and protein synthesis in healthy young and older men." Jama 286(10): 1206-1212.

Walrand, S., C. Guillet, J. Salles, N. Cano and Y. Boirie (2011). "Physiopathological mechanism of sarcopenia." Clinics in geriatric medicine 27(3): 365-385.

Wang, Q. A. (2008). "Probability distribution and entropy as a measure of uncertainty." Journal of Physics A: Mathematical and Theoretical 41(6): 065004.

Wang, H., C. Ma and L. Zhou (2009). A brief review of machine learning and its application. 2009 International Conference on Information Engineering and Computer Science, IEEE. 19 (pp. 1-4)

Wang, Y., F. S. Makedon, J. C. Ford and J. Pearlman (2005). "HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data." Bioinformatics 21(8): 1530-1537.

Welle, S., A. I. Brooks, J. M. Delehanty, N. Needler and C. A. Thornton (2003). "Gene expression profile of aging in human muscle." Physiological genomics 14(2): 149-159.

Welle, S., R. Tawil and C. A. Thornton (2008). "Sex-related differences in gene expression in human skeletal muscle." PloS one 3(1): e1385.

West, M., C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. A. Olson, J. R. Marks and J. R. Nevins (2001). "Predicting the clinical status of human breast cancer by using gene expression profiles." Proceedings of the National Academy of Sciences 98(20): 11462-11467..

Weston, J., Elisseff, A., Schoelkopf, A.B., Tipping, M., 2003. Use of the zero norm with linear models and kernel methods. J. Machine Learn. Res. 2003 (3), 1439–1461.

Whitley, E. and J. Ball (2002). "Statistics review 3: hypothesis testing and P values." Critical Care 6(3): 1.

Wilson, V. L., R. Smith, S. Ma and R. Cutler (1987). "Genomic 5-methyldeoxycytidine decreases with age." Journal of Biological Chemistry 262(21): 9948-9951.

Witten, D. and R. Tibshirani (2007). "A comparison of fold-change and the t-statistic for microarray data analysis." Analysis 17. 1776 (2007): 58-85

Wohlgemuth, S. E., A. Y. Seo, E. Marzetti, H. A. Lees and C. Leeuwenburgh (2010). "Skeletal muscle autophagy and apoptosis during aging: effects of calorie restriction and life-long exercise." Experimental gerontology 45(2): 138-148.

Wray, N. R., J. Yang, M. E. Goddard and P. M. Visscher (2010). "The genetic interpretation of area under the ROC curve in genomic profiling." PLoS Genet 6(2): e1000864.

Wu, X., V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu and S. Y. Philip (2008). "Top 10 algorithms in data mining." Knowledge and information systems 14(1): 1-37.

Wu, Y., K. Ianakiev and V. Govindaraju (2002). "Improved k-nearest neighbor classification." Pattern recognition 35(10): 2311-2318.

Xia, Q., D. Cheng, J. Duan, G. Wang, T. Cheng, X. Zha, C. Liu, P. Zhao, F. Dai and Z. Zhang (2007). "Microarray-based gene expression profiles in multiple tissues of the domesticated silkworm, Bombyx mori." Genome biology 8(8): 1.

Xing, E. P. and R. M. Karp (2001). "CLIFF: clustering of high-dimensional microarray data

via iterative feature filtering using normalized cuts." Bioinformatics 17(suppl 1): S306-S315.

Xing, E. P., M. I. Jordan and R. M. Karp (2001). Feature selection for high-dimensional genomic microarray data. ICML, Citeseer. vol. 1, pp. 601-608

Xu, Y., V. Olman and D. Xu (2002). "Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees." Bioinformatics 18(4): 536-545.

Yan, X. (2013). "Weighted K-Nearest Neighbor Classification Algorithm Based on Genetic Algorithm." Indonesian Journal of Electrical Engineering and Computer Science 11(10): 6173-6178.

Yeoh, E.-J., M. E. Ross, S. A. Shurtleff, W. K. Williams, D. Patel, R. Mahfouz, F. G. Behm, S. C. Raimondi, M. V. Relling and A. Patel (2002). "Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling." Cancer cell 1(2): 133-143.

Yeoh, E.-J., M. E. Ross, S. A. Shurtleff, W. K. Williams, D. Patel, R. Mahfouz, F. G. Behm, S. C. Raimondi, M. V. Relling and A. Patel (2002). "Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling." Cancer cell 1(2): 133-143.

Yu, L. and H. Liu (2004). Redundancy based feature selection for microarray data. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 10 ACM. pp. 737-742

Yun, G. A., K. S. LeunG, P. M. F. Siu, J. H. Qin, S. K. H. Chow, L. Qin and C. Y. Li (2015). "Muscle mass, structural and functional investigations of senescence-accelerated mouse P8 (SAMP8)." Experimental animals 64(4): 425-433.

Zacharewicz, E., P. Della Gatta, J. Reynolds, A. Garnham, T. Crowley, A. P. Russell and S. Lamon (2014). "Identification of microRNAs linked to regulators of muscle protein synthesis and regeneration in young and old skeletal muscle." PloS one 9(12): e114009.

Zahn, J. M., R. Sonu, H. Vogel, E. Crane, K. Mazan-Mamczarz, R. Rabkin, R. W. Davis, K. G. Becker, A. B. Owen and S. K. Kim (2006). "Transcriptional profiling of aging in human muscle reveals a common aging signature." PLoS Genet 2(7): e115.