Northumbria Research Link

PARETO OPTIMAL-BASED FEATURE SELECTION FRAMEWORK FOR BIOMARKER IDENTIFICATION

OMER FARUK OGUTCEN

FACULTY OF ENGINEERING ENVIRONMENT

NORTHUMBRIA UNIVERSITY

THIS DISSERTATION IS SUBMITTED FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

October 2019

DECLARATION

I declare that the work contained in this thesis has not been submitted for any other award and that it is all my own work. 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. Approval has been sought and granted by the University Ethics Committee on February 2015.

I declare that the Word Count of this Thesis is 28671 words.

Omer Faruk Ogutcen

19.10.2020

TABLE OF CONTENTS

DECLARATION 2

TABLE OF CONTENTS 3

LIST OF FIGURES 7

LIST OF TABLES 8

List of ACRONYMS 9

ACKNOWLEDGEMENTS 11

Abstract 13

1 INTRODUCTION 14

1.1 Research Motivation 14

1.2 Rationale of the study 19

1.3 Research Aim and Objectives 20

1.4 Research Methodology 20

1.5 Research Scope 21

1.6 Contribution 21

1.7 Thesis Organisation 22

2 LITERATURE REVIEW 25

2.1 Introduction 25

2.2 Biological Datasets 25

2.2.1 Sequencing Data 25

2.2.2 Microarrays 26

2.3 Bioinformatics 27

2.4 Feature Selection 28

2.4.1 Overfitting 31

2.4.2 Curse of dimensionality 31

2.4.3 Class Imbalance 32

2.4.4 Increase interpretability 33

2.4.5 Ensemble Feature Selection 34
2.4.6 Feature Selection Methods 35
2.5 Pareto Optimal (PO) 39
2.6 Conclusion 42
3 MATERIALS AND METHODS 43
3.1 Introduction 43
3.2 Microarray Data Analysis 43
3.2.1 Gene-Expression Data 43
3.2.2 Imbalanced Data 45
3.2.3 Imputation Methods for Data with Missing Values 46
3.3 Experimental Datasets 47
3.3.1 Colon Cancer 47
3.3.2 DLBCL (Lymphoma) 47
3.3.3 DUKE Breast Cancer 47
3.3.4 CNS (Central Nervous System) Embryonal Tumour 47
3.3.5 Ovarian Cancer 48
3.3.6 DLBCL (Blood) Cancer 48
3.3.7 Brain Cancer Data 48
3.3.8 Lung Cancer Data 48
3.3.9 Prostate Cancer Data 49
3.3.10 Endometrium 49
3.4 Data Validation 49
3.5 ***K***-Fold Cross-Validation 49
3.6 Classification 50
3.6.1 The ***k***-nearest neighbour (***k***-NN) 50
3.6.2 Support Vector Machine (SVM) 50
3.7 Ranking Methods 52

3.7.1 Two sample ***T***-test 52
3.7.2 Entropy 52
3.7.3 Bhattacharyya 53
3.7.4 Receiver Operating Characteristic (ROC) Analysis 53
3.7.5 Wilcoxon 55
3.8 Pareto Optimality (PO) Method 56
Four objectives example for PO 57
3.9 Performance Evaluation 60
3.10 Conclusion 60
4 A PARETO OPTIMAL MULTI-OBJECTIVE FRAMEWORK FOR AGGREGATED DISCOVERY OF DISEASE-RELATED GENES 61
4.1 Introduction 61
4.2 The Problem of Gene Selection 61
4.3 Ensemble Feature Selection Methods 62
4.4 PO-based Multi-Criteria Framework 65
4.5 Integration of the Wrapper Method with the Proposed Framework . 66
4.5.1 Cross-Validated Aggregated Gene Selection using Filter Methods 67
4.5.2 Cross-Validated Aggregated Gene Selection using Filter-Wrapper Combination 71
4.6 Results and Discussion 74
4.7 Conclusion 83
5 A PARETO-OPTIMAL GENE SELECTION FRAMEWORK FOR IMBALANCED DATA 85
5.1 Introduction 85
5.1.1 Performance Metrics for Imbalanced Microarray Gene Expression Data 86
5.1.2 Gene selection in imbalanced microarray datasets 87

5.2 Proposed approach for handling data imbalance 88
5.3 Results and Discussion 90
5.4 Conclusion 93
6 A PARETO-OPTIMAL GENE SELECTION FRAMEWORK WITH MISSING VALUE IMPUTATION 95
6.1 Introduction 95
6.2 Imputation Methods 97
6.2.1 Statistical Imputation for Gene Expression 98
6.2.2 Machine Learning-based Imputation for gene expression 98
6.3 Proposed Framework 99
6.4 Results and Discussion 102
6.5 Conclusion 105
7 CONCLUSION AND FUTURE WORK 107
7.1 Thesis Summary 107
7.2 Summary of Core Contributions 108
7.2.1 Pareto Optimal Framework for Feature Selection on Biological Data Sets 108
7.2.2 Gene Selection on Imbalanced Datasets 109
7.2.3 Gene Selection on Datasets Having Missing Values 109
7.3 Future Research Directions 109
BIBLIOGRAPHY 111

LIST OF FIGURES

FIGURE 2-1 ILLUSTRATION OF THE DNA STRUCTURE (GENOME.GOV, 2020) 26
FIGURE 2-2 CURSE OF DIMENSIONALITY (HAURY, 2013) 32
FIGURE 3-1 ILLUSTRATIVE EXAMPLE OF SVM CLASSIFICATION (LIANG ET AL., 2016) 51
FIGURE 3-2 AN ILLUSTRATIVE EXAMPLE OF AREA UNDER CURVE 54
FIGURE 3-3. DEMONSTRATION OF A MAXIMISATION CASE WITH TWO OBJECTIVES. 57
FIGURE 4-1. FLOWCHART OF THE PROPOSED FRAMEWORK USING PARETO OPTIMALITY ON PAIRED-UP FEATURES OF FILTER METHODS (MODEL-1) 68
FIGURE 4-2 FLOWCHART OF THE PROPOSED FRAMEWORK USING PARETO OPTIMALITY ON PAIRED-UP FEATURES OF THE APPLIED FILTER METHOD (MODEL-2) 69
FIGURE 4-3 FLOWCHART OF THE PROPOSED FRAMEWORK WITHOUT PARETO OPTIMALITY ON PAIRED-UP FEATURES OF THE APPLIED FILTER METHOD (MODEL-3) 70
FIGURE 4-4 FLOWCHART OF THE PROPOSED FRAMEWORK USING PARETO OPTIMALITY ON PAIRED-UP FEATURES OF THE FILTER-WRAPPER COMBINATION 72
FIGURE 4-5 FLOWCHART OF THE PROPOSED FRAMEWORK WITHOUT PARETO OPTIMALITY ON PAIRED-UP FEATURES OF THE APPLIED FILTER-WRAPPER COMBINATION 73
FIGURE 5-1 FLOWCHART OF THE PROPOSED FRAMEWORK USING PARETO OPTIMALITY ON PAIRED-UP FEATURES OF THE FILTER-WRAPPER COMBINATION TO BE USED FOR IMBALANCED DATASETS. 88
FIGURE 6-1 FLOWCHART OF THE PROPOSED FRAMEWORK USING PARETO OPTIMALITY ON PAIRED-UP FEATURES OF FILTER METHODS (MODEL-1) APPLIED FOR THE DATASETS HAVING MISSING VALUES 100

LIST OF TABLES

TABLE 3.1 CONFUSION MATRIX 53

TABLE 3.2 SIMPLIFIED 4 OBJECTIVES AND 6 SAMPLES DATASET EXAMPLE FOR PO 58

TABLE 4.1: PERFORMANCE OF MODEL-1 75

TABLE 4.2: PERFORMANCE OF MODEL-2 (WITH PO) AND MODEL-3 (WITHOUT PO) 75

TABLE 4.3: PERFORMANCE COMPARISON OF MODEL-1 USING LOO CROSS-VALIDATION WITH STATE-OF-THE-ART TECHNIQUES USING THE SAMPLE SUBSETS 76

TABLE 4.4: CLASSIFICATION RESULTS OF COLON BREAST CANCER WITH THE PROPOSED FILTER-WRAPPER COMBINED FRAMEWORK 77

TABLE 4.5: CLASSIFICATION RESULTS OF DLBCL BREAST CANCER WITH THE PROPOSED FILTER-WRAPPER COMBINED FRAMEWORK 78

TABLE 4.6: CLASSIFICATION RESULTS OF DUKE BREAST CANCER WITH THE PROPOSED FILTER-WRAPPER COMBINED FRAMEWORK 79

TABLE 4.7: PERFORMANCE COMPARISON OF ENSEMBLE METHODS APPLIED TO COLON DATASET 80

TABLE 4.8: PERFORMANCE COMPARISON OF ENSEMBLE METHODS APPLIED TO DLBCL DATASET 81

TABLE 4.9: COMPARISON OF ENSEMBLE METHODS APPLIED TO DUKE DATASET 82

TABLE 5.1 CONFUSION MATRIX 86

TABLE 5.2 CHARACTERISTICS OF IMBALANCED GENE EXPRESSION DATASETS 90

TABLE 5.3 MEAN AND STANDARD DEVIATIONS GENERATED FOR THE SUBSET-BALANCED DATASET SIZES AND K-LEVEL OF THE KNN CLASSIFIER AT THE MAXIMUM ACCURACY 91

TABLE 5.4 PERFORMANCE COMPARISONS FOR THE IMBALANCED CNS GENE EXPRESSION DATASET 91

TABLE 5.5 PERFORMANCE COMPARISONS FOR THE IMBALANCED LYMPH GENE EXPRESSION DATASET 92

TABLE 5.6 PERFORMANCE COMPARISONS FOR THE IMBALANCED OVARIAN GENE-EXPRESSION DATASET 92

TABLE 6.1 THE COMPUTATIONAL COMPLEXITY OF MACHINE LEARNING-BASED IMPUTATION METHODS USED IN THE PROPOSED FRAMEWORK 99

TABLE 6.2 CANCER DATASETS WITH MISSING VALUES 103

TABLE 6.3 ACCURACY PERFORMANCE COMPARISON OF VARIOUS IMPUTATION METHODS ON GENE EXPRESSION DATASETS USING PO-BASED FEATURE SELECTION WITH KNN CLASSIFIER. 103

TABLE 6.4 ACCURACY PERFORMANCE COMPARISON OF VARIOUS IMPUTATION METHODS ON GENE EXPRESSION DATASETS USING PO-BASED FEATURE SELECTION WITH SVM CLASSIFIER. 104

List of ACRONYMS

AUC	Area Under Curve
BPCA	Bayesian Principal Component Analysis
CFS	Correlation-based Feature Selection
CNS	Central Nervous System
DLBCL	Diffuse Large B-cell Lymphoma
DNA	Deoxyribonucleic Acid
FN	False Negative
FP	False Positive
FPR	False Positive Rate
KNN	K- Nearest Neighbour
LLS	Local Least Squares
LOO-CV	Leave-One-Out Cross-Validation
MI	Mutual Information
MVs	Missing Values
PCA	Principal Components Analysis
PO	Pareto Optimal
RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristic
RSFS	Random Subset Feature Selection
SFFS	Sequential Floating Forward Selection
SFS	Sequential Forward Selection
SNR	Signal-to-Noise Ratio
SVM	Support Vector Machine
SVM-RFE	Support Vector Machine-based Recursive Feature Elimination
TN	True Negative
TNR	True Negative Rate
TP	True Positive
TPR	True Positive Rate
WKNN	Weighted K-Nearest Neighbours

To my guide

***

There is no emigration for the virtuous; also, there is no homeland for the ignorant.

[Süfyan-ı Sevri]

***

“The requisites of being a student are humiliation and being in away from home, or else all is for nought”

[Shams Tabrizi]

***

ACKNOWLEDGEMENTS

First and foremost, I sincerely wish to express my gratitude to my supervisor, Dr Ammar Belatreche, for giving me immense encouragement and guidance in the completion of this thesis. I am also thankful to my ex-co-supervisors, Dr Huseyin Seker and Dr Muhammed Atif Tahir, for their direction, guidance and support.

I would like to thank my colleagues Baqar Abbas Rizvi, Daniel Organisciak for their invaluable collaborations, comments, and contributions.

Especially thankful to my family members for all their encouragement, support and love during my entire venture in the Northumbria.

I would like to acknowledge Northumbria University, for various support to carry out this research.

PUBLICATIONS

Ogutcen, O.F., Gormez, Z., Tahir, M.A. and Seker, H., 2016. An aggregated cross-validation framework for computational discovery of disease-associative genes. In *XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016* (pp. 489-494). Springer, Cham.

Ogutcen, O.F., Belatreche, A. and Seker, H., 2018, September. A Multi-objective Pareto-Optimal Wrapper Based Framework for Cancer-Related Gene Selection. In *Proceedings of SAI Intelligent Systems Conference* (pp. 353-364). Springer, Cham.

Ogutcen, O.F., Belatreche, A. and Seker, H. A Pareto-optimal based gene selection framework for hybrid multi-criteria biomarker discovery (submitted to Journal of Bioinformatics and Computational Biology)

Ogutcen, O.F., Belatreche, A. and Seker, H. Balanced learning while using original Imbalanced data (to be submitted to 2020 IEEE EMBC - Conference of Engineering in Medicine and Biology, July 20-24, 2020, Montréal, Québec, Canada)

Abstract

Numerous computational techniques have been applied to identify the vital features of gene expression datasets in aiming to increase the efficiency of biomedical applications. The classification of microarray data samples is an important task to correctly recognise diseases by identifying small but clinically meaningful genes. However, identification of disease representative genes or biomarkers in high dimensional microarray gene-expression datasets remains a challenging task. This thesis investigates the viability of Pareto optimisation in identifying relevant subsets of biomarkers in high-dimensional microarray datasets. A robust Pareto Optimal based feature selection framework for biomarker discovery is then proposed.

First, a two-stage feature selection approach using ensemble filter methods and Pareto Optimality is proposed. The integration of the multi-objective approach employing Pareto Optimality starts with well-known filter methods applied to various microarray gene-expression datasets. Although filter methods provide ranked lists of features, they do not give information about optimum subsets of features, which are namely genes in this study. To address this limitation, the Pareto Optimality is incorporated along with filter methods. The robustness of the proposed framework is successfully demonstrated on several well-known microarray gene expression datasets and it is shown to achieve comparable or up to 100% predictive accuracy with comparatively fewer features. Better performance results are obtained in comparison with other approaches, which are single-objective approaches. Furthermore, cross-validation and k-fold approaches are integrated into the framework, which can enhance the over-fitting problem and the gene selection process is subsequently more accurate under various conditions.

Then the proposed framework is developed in several phases. The Sequential Forward Selection method (SFS) is first used to represent wrapper techniques, and the developed Pareto Optimality based framework is applied multiple times and tested on different data types. Given the nature of most real-life data, imbalanced classes are examined using the proposed framework. The classifier achieves high performance at a similar level of different cases using the proposed Pareto Optimal based feature selection framework, which has a novel structure for imbalanced classes. Comparable or better gene subset sizes are obtained using the proposed framework. Finally, handling missing data within the proposed framework is investigated and it is demonstrated that different data imputation methods can also help in the effective integration of various feature selection methods.

1 INTRODUCTION

1.1 Research Motivation

Microarray datasets have been increasingly used during the last two decades, and this creates an intriguing challenge for machine learning and bioinformatics. Microarray data is comprised of gene expression information gained from tissue or cell samples. It uses gene expression values for disease diagnosis or the classification of specific types of tumour (Bolón-canedo *et al.*, 2014). Mostly, while there is a minimal number of samples (usually less than 100) for training and testing, the number of features generally found in original data ranges between 6000 and 60,000, and hence a gene expression dataset is considered as a type of a big data (Bolón-Canedo, Sánchez-Maroño and Alonso-Betanzos, 2015). Microarray technology provides biologists with a valuable tool for measuring thousands of gene expression levels in a single experiment (Saeys *et al.*, 2007). The sheer amount of gene expressions obtained in various microarray data analyses can be used in classification tasks (Jirapech-Umpai and Aitken, 2005).

It is difficult for machine learning methods to deal with high numbers of input features (Bolón-Canedo *et al.*, 2013) and high dimensional big data often creates a problem for researchers when working with them (Khoshgoftaar *et al.*, 2013). Therefore, feature selection algorithms have become indispensable components of the learning process and face to many input features during the descriptive analysis of the research. However, the performance of feature selection methods can be adversely affected from high numbers of input features (e.g., Information Gain and ReliefF). In addition, the classification process can be conducted without feature selection and training and validation performances in terms of accuracy can be reasonably good. However, testing the performance of the classification model often far from meeting expectations, and many studies have shown that feature selection has a significant effect on efficiency of machine learning methods (Jirapech-Umpai and Aitken, 2005). It is reasonable to expect that having so many features with limited samples create a high likelihood of finding "false positives" in discovering relevant genes and in building reliable predictive models. While more features might be expected, at least theoretically, to provide more discriminatory ability, they, in reality,

contribute to a significant slow-down of the classifier training process and also result in overfitting the training data due to the presence of redundant features which impacts negatively on the learning process (Yu and Liu, 2004). Several studies have shown that microarray data often contain redundant genes that are irrelevant to accurately classify different classes of the microarray dataset (Bolón-canedo *et al.*, 2014).

Machine learning methods have been utilised to discover hidden structures in biological data and biomarkers. For this purpose, reducing dimensionality for the analysis of the data, selecting a subset of the features, classifying the data, clustering, and estimating new situations are the primary applications (Saeys *et al.*, 2007). Most machine learning algorithms for feature selection demand a high number of samples to effectively produce less but relevant descriptors (Mandoiu and Zelikovsky, 2008; Wu *et al.*, 2008; Bolón-Canedo and Alonso-Betanzos, 2019). Therefore, feature selection methods as a common and efficient strategy, aim to decrease the number of dimensions of the dataset. In gene expression studies, gene selection is a de-facto method to identify the most relevant genes by deleting irrelevant and redundant genes (V. Bolon-Canedo *et al.*, 2015). Feature selection refers mainly to gene selection, in such studies.

Feature selection has been broadly studied in the literature, particularly for classification (Saeys *et al.*, 2007; Sun *et al.*, 2015; Bolón-Canedo and Alonso-Betanzos, 2019). Recently, with the proliferation of high dimensional datasets, big data analysis has become a widely studied research area in engineering as well as other fields including the physical, biological and biomedical sciences (Bolón-Canedo, Sánchez-Maroño and Alonso-Betanzos, 2014). For instance, the analysis of microarray gene expressions is an important process in the biological sciences where thousands of genes simultaneously are expressed in relation to diseases. The information or knowledge obtained from microarray data is valuable and could be big and further analysis is required to be performed for subsequent prospective actions (Bolón-canedo *et al.*, 2014; Veronica Bolon-Canedo *et al.*, 2015). One such action is the gene selection before the classification takes place. The robustness of feature selection is a crucial issue in classification and mostly obtained by the validation methods (Khoshgoftaar *et al.*, 2013; Gerlein *et al.*, 2016). Recent studies have shown that the efficiency

of the classification is highly affected by the feature selection methods that were used (Yang and Mao, 2011; Seijo-Pardo *et al.*, 2015).

In well-known low dimensional problems, the number of observations is much higher than the number of genes. This means that the dataset contains sufficient information for the training process in order to learn the distribution of patterns (Yang and Mao, 2011). Many existing pattern recognition algorithms can provide satisfactory results with this kind of low dimensional data. However, gene-expression datasets have high dimensionality and a small structure (HDSS) that has attracted considerable attention from the machine learning perspective (Yang *et al.*, 2015). Higher numbers of features and low number of samples creates a considerable challenge for pattern recognition algorithms. This situation is called the "curse of dimensionality" (Bellman, 1961) where the search space for a subset increases exponentially (Malley, Dasgupta and Moore, 2013). Thousands of features make it impossible to form the candidate solutions using an exhaustive search. Another problem is the small numbers of samples for the algorithms. These make the search even more challenging in establishing a subset of relevant features (Winkler, Affenzeller and Wagner, 2007).

In the analysis of biological data, selecting the subset of the identified problem-related features (e.g. genes in microarrays) can be used for cancer treatment studies, biomarker discovery, drug discovery and similar research, so it is an essential and critical process. Nowadays, microarray datasets have become the most common source for the analysis of gene expression (Veronica Bolon-Canedo *et al.*, 2015). Microarray datasets have existed for two decades (Taub, Deleo and Thompson, 1983), and they contain huge amounts of biomarker information (Makałowski, Jakalski and Makałowska, 2014). Biomarker detection is a difficult problem, especially when the number of genes is significant, and thus the search space grows exponentially (Ahmed, Zhang and Peng, 2015). Biomarker selection is a challenging process due to the high dimensional nature of post-genomic data such as microarray gene expression (Uslan, 2015; Peng, Li and Liu, 2017). Subsequently, biological datasets contain many features, only a small group of them is associated with the problem (Saeys *et al.*, 2007). Gene information is collected from many patients

especially cancer patients. Researchers collect it to compare with the data of healthy people whether to predict, diagnose, or monitor disease. Machine learning techniques became indispensable tools, and researchers are highly benefiting from many machine-learning methods to identify useful information from the massive amount of raw data they have. Biomarkers are beneficial for each stage of patient care. (Michiels, Koscielny and Hill, 2005; Yang and Mao, 2011; Bolon-Canedo *et al.*, 2017).

Therefore, it is of great importance to select a subset of representative biomarkers. The aims of the biomarker selection can be summarised as follows: (i) to reveal features associated with a problem identified in biological data such as genetic networks, structures and mechanisms, which are involved in the onset and progression of the disease, and (ii) to improve the accuracy and interpretability of the predictive model used to extract the features.

However, small sample sizes and high numbers of features in biological data are the main difficulties encountered in these studies. Thus, many feature selection algorithms have been proposed to overcome these problems. The disadvantage of having many solution methods is that, when different selection methods are applied to the same dataset, different biomarkers (genes) or subsets of biomarkers may be selected. At the same time, the selection process must be independent of sample variation in the dataset (Alpaydin, 2010). The same feature selection algorithm can result in the selection of different features even when applied to two sets of data that are very similar to each other. In many related studies, it has been shown that feature selection and ranking are related to dataset variation and selection methods (Michiels, Koscielny and Hill, 2005; Yang and Mao, 2011). These studies show that it is unreliable to use only one learning set per feature or only one method for feature selection.

Despite the expectation that a feature selection method would eliminate irrelevant variables resulting in good classification performance, classifiers often tackle the contradiction between bias and variance in datasets. The bias-variance trade-off is an important concept to be considered during the design of the model. The bias indicates a degree of adaptation of the model to the training set, but this is not a measure of generalisation. However, variance represents the

variability in the model performance that occurs during the class prediction of data samples (Alpaydin, 2010).

Feature selection, also called biomarker or gene selection in gene expression studies, can be utilised by applying the filter, wrapper or embedded approaches (Saeys *et al.*, 2007). In filter methods, information-based evolution relies on the overall features of the training data and the gene selection is regarded as a pre-processing stage which can be considered independently from the induction algorithm (Bolón-canedo *et al.*, 2014).

Existing studies show that it is unreliable to use only one learning set per feature, only one method for feature selection, and ranking for disease-associated genes. The current trend is based on ensemble methodologies to gain better results (Bolón-Canedo and Alonso-Betanzos, 2019).

This continuing research trend tackles many challenges in microarray data. It shows the generic characteristic of big data even with the small sample sizes that present fundamental difficulties. Class imbalance, missing data, overlapping between classes, non-linearity, and such as genes extracted under different distributions (Bolón-Canedo, Sánchez-Maroño and Alonso-Betanzos, 2015). Typically, microarray datasets have fewer than 100 samples, and thousands of features make the training process of a prediction model more laborious. Less information and more criteria for decisions aggravate the problem. In the literature, several studies show that most gene-expression values in a DNA(deoxyribonucleic acid) microarray experiment are not related to objective class so there is no benefit to classification accuracy (Hira and Gillies, 2015). This type of information may cause a reduction in the performance of classifiers such as the C4.5 decision tree classifier as the decision tree method is faced with numerous redundant features that are completely unrelated with the response classes of the study. (Bolón-Canedo, Sánchez-Maroño and Alonso-Betanzos, 2015). Likewise, instance-based learners such as the *k*-nearest neighbour (*k*-NN) are strongly affected by redundant features. Higher numbers of irrelevant features exponentially increases the demand for training instances to produce a predetermined level of accuracy performance (Alpaydin, 2010). The irrelevant features slow down the learning process of these methods and

decrease their performance (KaltenbachHans-Michael, 2013). Adding redundant features, even those related to the concept under research, rapidly decrease prediction performance. These circumstances create the necessity for minimising the negative effects by establishing a robust framework for the analysis. The robustness of the established framework often obtained using the validation approaches (e.g., cross-validation). Thus, the established framework avoids negative effects, such as the problem of overfitting. Furthermore, the models used must deal with classes that are imbalanced, and with samples extracted under different conditions in both training and test datasets, thus stimulating research into utilising the analysis of microarray data (Veronica Bolon-Canedo *et al.*, 2015).

This study aims to create a new framework combining a selection method, which is not influenced by sample variation in the dataset with the multi-purpose optimisation Pareto Optimal (PO) approach.

1.2 Rationale of the study

Most of the existing related studies rely on an individual feature selection technique with single feature ranking. The singular evaluation provides a set of redundant attributes which often slow down the learning process and degrade the classification accuracy (Bolón-Canedo and Alonso-Betanzos, 2019).

Important aspects to consider are class imbalanced datasets (Fernandez, Garcia and Herrera, 2011) and missing datasets. In imbalanced datasets, a majority class generally has a greater effect while selecting features. This challenge remains an open research issue in handling high dimensional datasets with noticeably imbalanced class distributions (Maldonado, Weber and Famili, 2014). However, most of the techniques applied certain limitations and yield different subsets of attributes. Therefore, it is considered that there is still room for improvement in the feature selection process, especially in the big data era.

Several types of feature selection methods perform a random selection process, where different feature subsets are selected in every run from the same dataset and all subsets are used in classification such as SVM (Maldonado, Weber and Famili, 2014) or K-means (Wu *et al.*, 2008). In these cases, the results of

classification can be unstable, and considerable computational resources could be needed during the classification process.

One of the challenging aspects of big data is missing data, where microarray datasets contain missing information about genes (Troyanskaya *et al.*, 2001). The major research challenge here is how to fit the most suitable model induction algorithm for missing data. Current methods and approaches can only recover data according to other successful samples (Souto, Jaskowiak and Costa, 2015). Identifying the most valuable features is necessary for research in this area (Muresan *et al.*, 2015).

This research study, therefore, aims to address the aforementioned problems when handling big microarray data with the use of computational methods. The specific objectives of this research are discussed in the following section.

1.3 Research Aim and Objectives

This research aims to investigate the viability of Pareto optimisation in identifying relevant subsets of biomarkers in high-dimensional microarray datasets. The primary objectives are as follows:

- To propose and develop a PO based predictive gene selection framework for biomarker discovery
- Integrate different feature selection and cross-validation approaches
- To handle the problem of imbalanced data distribution for gene selection within the proposed framework
- To evaluate the contribution of different data imputation methods on the proposed framework performance when there are missing values in the data
- To evaluate the classification and generalisation performance of the proposed framework on well-known microarray gene expression datasets

1.4 Research Methodology

Different microarray datasets are applied to test the unique subsets and robustness of the proposed framework. The framework contains multiple stages for multi-criteria decision cases. Its adjustable structure is analysed with different parameters and altered structures.

Given the nature of non-streaming data, the cross-validation method may be the solution to validate the dataset. Variations in the dataset are applied to prove the robustness of the model. When variables change, the PO-based framework is consistent in achieving a high level of accuracy. This consistency makes the framework more reliable, especially for further biological studies. The role of PO is to eliminate irrelevant and redundant features and to generate a common gene subset for each dataset. The gene selection performance is evaluated by classification. Relevant information about performance results is shown in tables.

1.5 Research Scope

This research focuses primarily on gene selection problems and is limited to microarray data case studies. Microarray datasets are suitable for the analysis of big data applications. The aim is to not only use common gene expression datasets but also to determine the applicability of PO on imbalanced and missing datasets. Besides, these datasets contain valuable information about various cancer diseases, which are used as case studies.

Most of the imbalanced datasets have binary values, so the present research is limited to imbalanced binary classification problems. Two of comparison datasets contains multi-class samples. We applied one vs all classification to match up binary datasets classes as healthy vs all other disease classes.

1.6 Contribution

In this thesis, a generalised computational framework is proposed that could employ any feature selection method and classifier model to identify representative subsets of genes in high-dimensional gene expression datasets. The main contributions of this thesis are summarised as follows:

1. A computational framework is developed based on Pareto Optimization for the purpose of identifying subsets of disease representative genes. PO based framework is designed to be independent of any feature selection methods and classifier models. The stability and robustness of the proposed system are validated using different feature selection methods. Due to its properties, such a framework can be easily adapted

with any feature selection method that may suffer from randomisation problems.

2. The proposed PO based framework is applied to imbalanced datasets. The gene expression datasets due to its nature are very high dimensional datasets. The proposed framework has not only addressed the high dimensionality with the use of feature selection methods, but the issues related to the imbalanced datasets are also addressed with the utilization of multi-criteria approach that considers the imbalanced dataset as a number of balanced sub-datasets.

3. The proposed PO based framework is applied to datasets having missing values. Real-world datasets commonly contain missing values and elimination of those values from datasets cause loss of information (Troyanskaya *et al.*, 2001). Missing values is a common problem in gene expression datasets (Souto, Jaskowiak and Costa, 2015). Simply discarding samples having missing values might result in valuable information being lost (Silva and Perera, 2017). This study has also investigated how the proposed PO-based framework selects features/genes when missing values are replaced using well-known statistical imputation methods. The classification performance of the proposed method is assessed for each imputation method. Moreover, these imputation methods are compared to each other in order to identify which ones offer better results.

1.7 Thesis Organisation

Chapter 1, Introduction: This chapter introduces the problems of dealing with high dimensional data, shows the significance of feature selection, and outlines the research motivation, aim and contribution. An outline of the thesis structure is presented.

Chapter 2, Literature Review: This chapter presents background information about feature selection with big data, followed by a review of studies of gene selection in microarray datasets. It then describes relevant studies, and the progress made on gene-expression data classification approaches used in bioinformatics and gene selection methods used to decrease dimensionality to

give for improvements in classification. It also discusses the main challenges involved such as the curse of dimensionality and overfitting. Finally, the PO method is discussed and the research methodology and the research hypothesis for the present investigation are derived from the findings of the literature review.

Chapter 3, Materials and Methods: This chapter explains the methods applied and microarray datasets used.

Chapter 4 An Aggregated Framework for Biomarker Discovery of Disease-Related Genes: The improvements found in comparison with the literature for all datasets are presented. How the framework was implemented to overcome the challenges of dealing with high dimensional data is discussed. The evaluation metrics used to analyse and compare the effectiveness of supervised feature selection methods previously applied are considered. The Cancer datasets exploited in this research to evaluate the performance of proposed frameworks are also described.

Chapter 5 Pareto-Optimal Feature Selection Framework for Imbalanced Data: This chapter discusses the challenges posed by imbalanced datasets and the shortcomings of existing imbalanced data approaches. The proposed PO-based framework for imbalanced datasets is introduced, which is constructed from the proposed first framework. Finally, the results of the application of the proposed framework compared to previous gene selection methods used in imbalanced data studies of the Central Nervous System (CNS), ovarian, and lymphoma are presented.

Chapter 6 PO Feature Selection Framework Application for Imputation Methods Comparison: Identifies research gaps in the area of the gene selection and imputation methods for missing values data. Imputation methods are compared to explore their effects on feature selection in missing values. Finally, experimental results are presented to show the effectiveness of imputation methods with feature selection for the five cancer datasets utilised to investigate performance.

Chapter 7 Conclusion and Future Work: This chapter presents a discussion of the application and evaluation of the proposed framework for the different data

set types. The potential contribution of the framework to feature selection with different data types is discussed in greater detail. This chapter also concludes the thesis and outlines future research directions.

2 LITERATURE REVIEW

2.1 Introduction

This chapter reviews previous studies of gene selection. The significance of the field is discussed first, followed by a review of dimensionality reduction. Then, a review of gene selection and PO is presented followed by a discussion of existing feature selection methods for gene subset selection and classification problems. As Pareto Optimisation forms an important element of the proposed a gene selection framework, a detailed section on Pareto optimality is presented.

2.2 Biological Datasets

Biological datasets can be grouped into sequencing and microarrays data. Strings are pure textual data. Microarrays are numerical expressions that show the expression level of genes (Wang, Miller and Clarke, 2008).

2.2.1 Sequencing Data

DNA (Deoxyribonucleic acid) is a nucleic acid that contains the genetic information necessary for the biological functions of living creatures (Figure 2-1). DNA particles (genes) are responsible for making proteins that are the basic building blocks for the living cell (Chuang *et al.*, 2011). Genes that act as templates for protein production turn into RNA (Ribonucleic acid) sequences with properties similar to DNA. RNA sequences allow the formation of protein sequences (Schnattinger *et al.*, 2013).

DNA is the most basic text-based biological sequence that carries genetic information and has an alphabet with 4 letters (A, C, G, T). Words of the desired length can be produced with these letters, which are called Adenine, Cytosine, Guanine and Thymine (Veronica Bolon-Canedo *et al.*, 2015). Researchers can encode enough words with this 4-letter DNA alphabet, just like the words created with two letters (dot and dash) in the same Morse alphabet and 2 letters in the computer language (0 and 1).

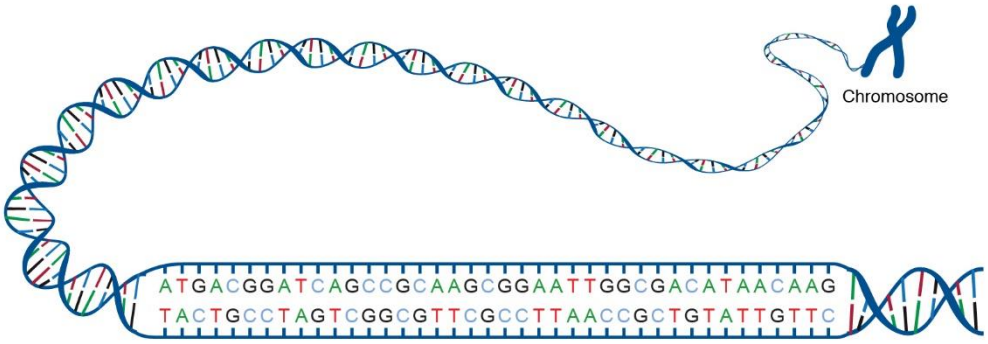

Figure 2-1 Illustration of the DNA structure (genome.gov, 2020)

2.2.2 Microarrays

Genes, which are a sequence of DNA, transform first into RNA and then into protein and take part in all biological functions in the living cell. Each gene has different roles in biological life functions. If a researcher wants to find out which genes are involved in the biological events of a cell, they can make a decision by looking at the amount of proteins that are products of that gene in the cell. This process can be defined as gene expression which is the exact indicator of protein level of the cell. The frequency with which a gene is transformed into DNA→RNA→Protein is referred to as the expression level of that gene. The increase (up) of the expression level is interpreted as the activation of a gene, and the decrease (down) as the suppression of that gene (Bolón-Canedo, Sánchez-Maroño and Alonso-Betanzos, 2016).

The expression level of genes can vary from cell to cell. Even for a particular cell type, the expression of a gene can vary depending on internal and external factors. For example, a gene that is highly expressed in a muscle cell may not be expressed at all in a liver cell. On the other hand, it happens that genes can be differentially expressed in the same cell when the internal and external factors are changed (Ochs, Casagrande and Davuluri, 2010). For example, a normally under-expressed gene may be more expressed during cell division. The level of gene expression may vary according to environmental factors other than the internal factors of an organism. For example, a gene that is expressed

at the usual level in a plant can be expressed more or less based on the adaptation of the plant to the environment in times of drought.

Microarray technology is used to determine the expression level of genes. With the application of the microarray technique, changes in the expression levels of thousands of genes can be detected simultaneously. The opportunity to compare many genes at the same time has enabled this technique to be used widely for different purposes (V. Bolon-Canedo *et al.*, 2015).

Generating gene expression data with microarray technology consists of many processes. The process of data creation is beyond the scope of this thesis. Therefore, the data used in this study are microarray data ready for analysis past all wet laboratory, image processing and quality evaluation processes. The data used is in a matrix format, where each row represents a gene, each column a sample.

2.3 Bioinformatics

The proliferation of data produced by systems biology has required extensive attention to sophisticated machine-learning tools and techniques, and this field is often described as bioinformatics (Cesario and Marcus, 2011). Nowadays, numerous biological datasets are publicly available (Williams *et al.*, 2010). Over the previous decade, databases have doubled in size every 15 month (Bolón-canedo *et al.*, 2014). The doubling period has now dramatically decreased and enormous amounts of data are being created.

Because of this growth in data, computational approaches to process and analyse large datasets have become essential for biological studies (Bolón-Canedo *et al.*, 2013). Recent advancements of computational power allow the efficient analysis of large quantities of data and the discovery of thoroughly complex structures in nature. *"Bioinformatics, is often defined as the application of computational techniques to understand and organise the information associated with biological macromolecules"* (Luscombe, Greenbaum and Gerstein, 2001). Bioinformatics has three main purposes. Firstly, it allows enormous volumes of data to be organised so that new entries can be added systematically. Secondly, tools and resources can be developed which support the analysis of data. Thirdly, to apply these tools on the

aforementioned data and discover biologically meaningful information (Luscombe, Greenbaum and Gerstein, 2001).

Bioinformatics is a vital area of application in understanding biomarker detection; for example, in determining disease-associated genes (Mount, 2004). A systematic approach should be used to improve the effectiveness of biomarkers in medical usage in the diagnostic and therapeutic environment (Eisenhaber, 2008).

Bioinformatics tools are necessary for drug development and have become more productive owing to their systematic approach for the elucidation of disease-associated genes and cancer drugs as well as the investigation of defective proteins as targets for drug discovery (Alpaydin, 2010).

One of the sources of bioinformatics is gene expression datasets. Generally, they contain a large number of features, i.e. high dimensionality, a low amount of samples, i.e. a small dataset size. In medical studies, researchers extensively study gene expression datasets for analysing cancer-related cell lines and observing differentiation in them. (Luscombe, Greenbaum and Gerstein, 2001). Research in cancer biology has been significantly progressed due to the use of experimental methodologies and the revolution in genomics and bioinformatics that has produced enormous amounts of biological data (Niu *et al.*, 2017) . One problem encountered is the conceptual frameworks used to organise high-dimensional microarray datasets in such a way that more progress can be made in the understanding of diseases (Ochs, Casagrande and Davuluri, 2010). It is clear that research studies in bioinformatics can incorporate experimental results in wet laboratories through computational analysis and modelling. These studies can provide valuable insights into understanding the cause of diseases and help design novel and innovative therapeutic procedures.

2.4 Feature Selection

In recent years, owing to the internet and increased digital data storage capacities, many of high dimensional datasets have become publicly available. In this situation, it is a challenging for machine learning methods to be able to deal with high dimensional datasets (Mahajan, Abhishek and Singh, 2016). At present, the dimensionality of any archive datasets is significantly increased,

and different repositories store this data. The methods used for reducing the number of dimensions are essential in improving the efficiency of the computational models in order to tackle the issue of the high numbers of input features. (Bolón-Canedo, Sánchez-Maroño and Alonso-Betanzos, 2016).

The task of dimensionality reduction is to identify a function $F: \mathbb{R}^d \rightarrow \mathbb{R}^k$, to reduce the number of features in the dataset from d to k, which suffer minimal accuracy loss when machine-learning models are applied. Dimensionality reduction methods usually exploit the two concepts of feature selection and feature extraction, each of which has its own merits (Zhao and Liu, 2011). Feature extraction methods accomplish a reduction in dimensions of data from the initial set of features. Feature extraction combines the primary variables in order to generate less number of variables. Then, a new set of features that are representative of the dataset yielding more learning performance are created. Various real world applications such as signal processing (Silvério Lopes and Magalh, 2011), image analysis (Samiappan, Prasad and Bruce, 2013) and information retrieval (Li, Li and Liu, 2017) mostly prefer this type of dimensionality reduction approach. In these cases, the interpretation is not as essential as the accurateness of the computational model. Widely used approaches for reducing the dimensionality are principal components analysis (PCA) (Moreau and Tranchevent, 2012) and linear discriminant analysis (Saeys *et al.*, 2007).

On the other hand, feature selection involves a process of finding a subset containing *k* features with the most information relevant to the problem defined in d dimensional data. In this way, irrelevant and redundant features are removed from the subset (Kira and Rendell, 1992; Guyon *et al.*, 2006). This strategy widely is used in data mining applications, such as genetics analysis (Mooney and Wilmot, 2015), sensor data processing(García-Laencina, Sancho-Gómez and Figueiras-Vidal, 2010) and text mining (Steenhoff *et al.*, 2012). Consequently, feature selection retains the existing attributes that are essential for predictive models and extract useful knowledge (Bolón-Canedo, Sánchez-Maroño and Alonso-Betanzos, 2015).

There are many problems encountered with very high dimensional data such as biological data sets where machine-learning methods are used for analysing these datasets and discovering hidden patterns of identifying useful correlations (Nilsson, 2007). However, applying machine learning on these large and high-dimensional datasets is computationally prohibitive and requires powerful and expensive computing facilities. Also, ill-defined features can affect the performance of the employed machine learning methods (Hira and Gillies, 2015). Therefore, the process of selecting smaller feature subsets is one of the most critical steps. Data can only be analysed after noise is removed.

The most important aspects of dimensionality reduction are summarised below as a process before the analysis of the data (Alpaydin, 2010):

a. Most learning algorithms depend on feature size (k) and the number of samples (N). Therefore, to improve efficiency, it is necessary to reduce the number of feature number. This will reduce the complexity of the inference algorithm. It also may help to reduce costs while limiting memory requirements.
b. When irrelevant data is discarded, it will reduce the amount of time required for processing it.
c. Learning algorithms used with small datasets are less influenced by specific factors such as noise and contraindication. In this case, simpler models are more reliable.
d. If the dataset is explained with fewer variables, it will be easier to extract the information because the production process is better understood.
e. When data is visualised in several dimensions without losing information, a visual analysis of the structure of and outliers in the data will be possible.

The process of feature selection mainly involves the identification of the most relevant features and the removal of redundant and irrelevant features to create a subset (Maldonado, Weber and Famili, 2014). The first benefit of feature selection is that it characterises the given problem more accurately and hence minimises the error rate. Without feature selection, ill-defined features can negatively affect the accuracy of classification of big datasets (Abdi, Hosseini

and Rezghi, 2012). (Yu and Liu, 2004) advised that additional effort is required to investigate the methods used on genomic microarray data in order to identify more informative genes and how few variables affect the performance in gene selection. Gene selection remains an open research area and every year the number of publications in those area increases (Bolón-Canedo and Alonso-Betanzos, 2019). If optimal feature are chosen, a better understanding of the underlying problem and more accurate representations of aggregated data can be achieved (Chuang *et al.*, 2011). In bioinformatics, which is the area of focus of this research, feature selection is referred to as gene selection. The central argument for gene selection is that a limited subset of disease-associated genes are required in order to identify reliable biomarkers. The following objectives are pursued when looking for a proper subset of representative genes:

2.4.1 Overfitting

When the dataset is small and has a large number of features, such as the microarray gene expression datasets, the analysis of the dataset can be highly dependent on samples of the training dataset (Wang, Miller and Clarke, 2008). In such a case, the features are not properly selected, classifier learns many irrelevant patterns than the relevant ones, and in turn, classifiers might lack a generalization capability and provide very poor testing performances. Many efforts have been made on analysing the problem of overfitting (Hawkins, 2004) and mitigating its effect on the prediction performance of a classifier (Subramanian and Simon, 2013).

2.4.2 Curse of dimensionality

The concept of curse of dimensionality states to a number of issues that might arise during the organization and analysis of the high-dimensional data (Bellman, 1961). One such issue is that the available number of features far exceed the number of samples. The processing of the data usually relies on identifying parts of data where samples form groups with similar attributes. However, as the number of samples are often sparse and dissimilar from each other in high-dimensional data, data analysis approaches become inefficient and suffers from establishing the required computational model. A symbolic rule is that there should be at least five samples having meaningful values for each of the variables (Dash and Misra, 2017). Classifier performance often increase

when sufficient number of samples are available which are related to a good representative number of features.

However, microarray experiments often lack the availability of sufficient amount of samples while the number of genes is naturally very large. It is often the case in microarray experiments that the number of probes representing the number of genes examined far exceeds the sample size.

Figure 2-2 shows that there is an optimal number of genes maximising the predictive accuracy.

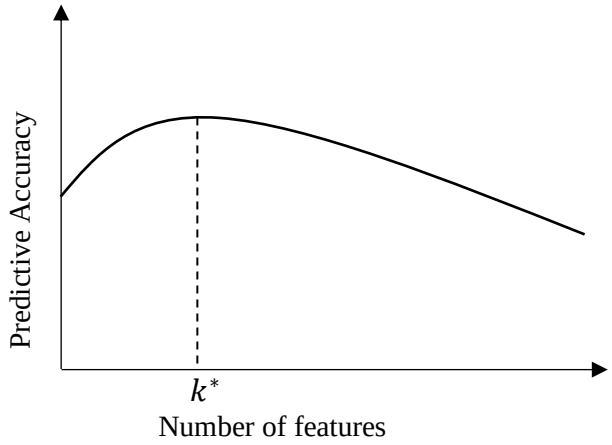

Figure 2-2 Curse of dimensionality (Haury, 2013)

Prediction models are frequently used for gene expression datasets to develop rules that can be applied to accurately classify the results for patients based on their characteristics. Creating microarray datasets are, however, costly to process (Haury, 2013). Such models represent an appropriate tool in the diagnostic process as they provide clinicians with estimates of possibility that people have or will develop a disease or have similar research objectives.

2.4.3 Class Imbalance

The nature of most biomedical datasets is to be imbalanced. For example, there may be more cancer cases than control samples (Shipp *et al.*, 2002). Therefore, applying feature selection to imbalanced datasets needs further study since the features selected may be more representative for the majority class (Fernandez, Garcia and Herrera, 2011). The multi-criteria approach will explore tackling this issue of feature selection from imbalanced datasets. Feature selection from high-dimensional imbalanced datasets will be transformed into feature selection from different relatively small classification sample balanced datasets.

Besides, this skewed class distribution has been found to cause training problems. Imbalanced sample proportion makes it harder to classify rare or minority classes because of training chance is much lower than majority cases. These minority or rare cases are with usually be cancer samples, so that it is crucial to classify them correctly. Commonly held is learning tendency of prediction models is upon majority class. From many data-level several approaches, most of the studies frequently used such as under-sampling, over-sampling. These type of methods causes the loss of valuable data or over-rated results because of virtually created samples. Instead of missing the originality of the data, this study proposes an approach that can equalise the level of class learning and preserve existing data.

2.4.4 Increase interpretability

An interpretable computational model is much easier to understand for humans as compared to ones, which are uninterpretable. Gene selection undoubtedly prevents the complexity of the computational model and ease the difficulty in interpretation. Furthermore, the subset of selected genes can be different from one model to another. Prediction performance of case studies such as the cancer related microarray datasets are often accurate and efficient when appropriate subset of genes was extracted. The subset of selected genes can further be verified in lab environment in order to consider their relation and link with the cancer.

Accepting infinite computational power, in theory no doubt that the optimal set of genes could be selected (Hastie, Tibshirani and Friedman, 2009; Hira and Gillies, 2015; Bolón-Canedo, Sánchez-Maroño and Alonso-Betanzos, 2016),

However, testing each possible subset among all numerous gene subset options is not possible in practice. In such a case, processing necessity of the selection of genes procedure can only happen in nondeterministic polynomial time. Thus, finding a solution for feature selection on microarray datasets become unattainable(Morán-Fernández, Bolón-Canedo and Alonso-Betanzos, 2017).

Simplification techniques are inevitably required for establishing good interpretable computational models. Various feature selection methods are investigated in terms of how they rank genes efficiently and accurately. Each gene is assigned with a score showing its significance. Then these genes are ordered based on their significance scores. The ranking order for gene significance scores is set from highest to lowest. Then the list of scores can be limited according to threshold values to create a suitable subset of genes. There are three types of feature selection methods familiar in the literature are filter, wrapper and embedded methods(Saeys *et al.*, 2007). Related algorithms utilised for gene selection are listed below for these groups. Finally, the PO method is introduced which is used intensively in of the core of this dissertation.

2.4.5 Ensemble Feature Selection

Ensemble feature selection combines various techniques by creating an ensemble learning from values of, for example, average, median, minimum, and maximum (Deb and Raji Reddy, 2003; Niijima and Kuhara, 2006; Uncu *et al.*, 2007; Handl, Kell and Knowles, 2007; Dinu and Popescu, 2008; Soto *et al.*, 2009; Yang and Mao, 2010; Lee, Jung and Shatkay, 2010; Ting, Lin and Huang, 2010; Chuang *et al.*, 2011; Luo *et al.*, 2011; Sabzevari and Abdullah, 2011). Furthermore, more complex models include ‘rank distance categorization’ (Dinu and Popescu, 2008), and ‘resampling and permutation feature importance’ (Yang and Mao, 2010) that downgrade the problem into a single-objective optimisation problem.

The performance of selected features can be calculated according to single or multiple objectives. In recent years, multi-purpose optimisation methods have

also been used for selection in bioinformatics studies (Moosa *et al.*, 2016). Often, two-objectives functions have been used in evaluating the performance of feature subclasses (Zitzler and Thiele, 1999). The first goal is the success in classification. If a subset correctly classifies the classes of the targeted classes such as patient, then this is a representative subset, and the features selected by the method can be said to be related to the disease. The second objective is to optimise the size or model cost of the subset of features. In many studies, one of the aims of researchers has been to minimise error or maximise accuracy. Other goals are to reduce the number of features and the cost of the model simultaneously (Deb and Raji Reddy, 2003; Sabzevari and Abdullah, 2011). As an alternative to reduce single objective only, this study proposes a technique widespread in operations research, which is the PO approach, in a framework, applied to the selection of the subsets of features.

2.4.6 Feature Selection Methods

Despite advances in the field of gene selection, immense challenges remain for gene-expression analysis where data comprise information about tens of thousands of genes(Li, Li and Yin, 2016). As described in Section 2.4.2, this is often referred to as the 'curse of dimensionality'. Moreover, high-dimensional data frequently includes numerous redundant and irrelevant genes. A relevant gene may be strongly correlated to another gene so in the presence of correlated gene it became redundant. Both experimental evidence and theoretical analysis show that data on redundant and irrelevant genes clearly affect the accuracy and speed of learning algorithms and thus it is advised that this data should be removed (Saeys *et al.*, 2007; Anaissi and Kennedy, 2011; Bolón-Canedo, Sánchez-Maroño and Alonso-Betanzos, 2012; Hasnat, 2016). Feature selection techniques can be categorised as either individual evaluation methods or subset evaluation methods (Yu and Liu, 2004). The evaluation considers distinct features by assigning to them weights according to their degree of relevance. This is also known as feature ranking. Techniques of subset evaluation look for the smallest subset of genes that fulfil some measure of importance and are able to eliminate redundant genes as well as irrelevant ones. However, many of the current heuristic search approaches used for the subset evaluation. One particular method for these approaches for feature selection is the SFS. SFS

enables a significant amount of decrease in the search space. However, the search space is still hard to be processed in a reasonable time especially for microarray experiments where the number of genes is huge in amount. Thus, there is a need to consider a different framework for gene selection which prevents the gene redundancy and evaluates genes that are highly relevant (Yu and Liu, 2004).

The three main approaches to feature selection are filter, wrapper, and embedded feature selection methods (Saeys *et al.*, 2007). These methods are discussed in the following sections.

2.4.6.1 Filter Methods

Filter approaches are based on the general characteristics of training data and are used as a pre-processing step with the independence of the induction algorithm. The first merit of these methods is their low computational cost and good generalisation ability (Bolón-Canedo, Sánchez-Maroño and Alonso-Betanzos, 2016). Also, their execution is faster than wrapper and embedded methods. However, they are prone to select large subsets (Saeys *et al.*, 2007). Filter methods include the t-test (Chen *et al.*, 2007), entropy (Nguyen *et al.*, 2015), Bhattacharyya (Haury, 2013), ROC (Nguyen *et al.*, 2015), and Wilcoxon (Haury, 2013; Nguyen *et al.*, 2015) tests. Fisher's ratio and Relief (Yang and Mao, 2010), and correlation-based feature selection (CFS) (Chuang *et al.*, 2011) are used in the post-genome domain due to their simplicity and speed.

These methods aim to achieve a possible subset that gives the highest score according to in limits of the technique itself. In these approaches, genes are often ranked in such a way that is appropriate for the research purpose, such as predictive success, correlation with class label, and so on. However, filter methods are independent of the estimation model selected and are not susceptible to any estimation model at this time. Filter methods are often counted as univariate methods and do not account for associations between features (Saeys *et al.*, 2007).

In the filter approach rather than the processing the search space, variables are selected based on the statics happen to be in the data. Filter methods are initially

preferred for microarray datasets as the feature selection method owing to the fact that they are highly efficient (Moosa *et al.*, 2016).

2.4.6.2 Wrapper Methods

This type of methods processes the whole search space. The criteria function decides the variables to be kept or eliminated during the process in order to maximize relevancy of features. In a wrapper technique, one may use a loss function j and any type of predictor g. The first proposed version of a wrapper method was sequential forward selection, which works by adding features to the subset one by one. The next version is called sequential backward elimination starts with the whole set of features and removes features successively.

Several selection criteria can be used in these algorithms. The final subset size can also be selected by estimating prediction success with cross-validation. Wrappers can be considered to be feature-ranking techniques, which return nested subsets of variables of pre-defined sizes.

Wrappers consist of a learning algorithm and run a prediction algorithm to identify the relative usefulness of subsets of variables. This communication with the classifier tends to yield higher accuracy than filters.

Wrapper methods such as SNR, SVM-RFE (Niijima and Kuhara, 2006; Luo *et al.*, 2011), simulated annealing (Lee, Jung and Shatkay, 2010) and genetic algorithms (Chuang *et al.*, 2011) are methods that depend on the classifier selected. These methods use the success of the chosen classifier as an evaluation constraint in determining the most essential features. Wrapper methods can conduct forward or backward selection. In forward selection, each element is tested individually while the subset is empty to start with. In a forward selection method, new features are added to the subset according to the objective at each step, and the number of elements in the cluster thus increases.

On the other hand, in a sequential backward selection method, a feature is removed in each step and the number of elements is thus reduced. In both cases, the stopping criterion may be an initial k-feature, or selection may continue until there is no additional reduction in error (Alpaydin, 2010). A significant disadvantage of such methods is the high computational cost incurred due to the need for classifier training and evaluation for each feature subset. When a high-

dimensional space data analysis is required, the filter approach appears to be the best alternative in terms of reducing computational cost (Uncu *et al.*, 2007). Embedded approaches are the third type of technique used for in feature selection, where filter and wrapper methods are combined. Correlation-based feature selection and Taguchi-genetic algorithm methods (Chuang *et al.*, 2011) or correlation coefficients and K-nearest neighbours (*k*NN) (Uncu *et al.*, 2007) are examples.

A good example that can be suggested to be used as a wrapper method is the sequential forward selection method. Sequential forward selection (SFS) starts from an empty subset and adds features sequentially without any backward step until no more improvement can be achieved (Whitney, 1971) or when a stopping criterion is reached (Pohjalainen, Räsänen and Kadioglu, 2015). This approach is well known and widely used in practice. Many similar methods have been proposed. For instance, sequential floating forward selection (SFFS) is performed in backward steps while the objective function increases (Pudil, Novovicova and Kittler, 1994) , random subset feature selection (RSFS) randomly adds features from subgroups to a chosen subset (Ho, 1998). SFFS provides a greedier and more flexible search of the dataset, and so checks more subsets. However, Reunanen (Reunanen, 2003) reported that the SFFS algorithm is computationally intensive and more prone to overfitting than SFS. It has been found in a recent comparative study that SFFS results in sizeable search space but fails to determine to a robust feature subset, performing worse than SFS (Pohjalainen, Räsänen and Kadioglu, 2015). RSFS can reduce the local optima problem and provide more diversity. Though, it relies on random values so it may over-fit to noisy data and may give different subsets every time.

Furthermore, Saeys et al. (Saeys *et al.*, 2007) cite the use of simulated annealing, probabilistic hill-climbing and genetic algorithms to overcome the local optima issue based on randomised search. These optimisation techniques make the algorithms less prone to local optima and take into account feature dependencies. On the other hand, randomised wrapper solutions contain similar disadvantages, being more computationally intensive and suffering from a higher risk of overfitting than deterministic algorithms. SFS is selected as the representative wrapper method due to its wide usage and simplicity.

Sequential forward selection is straightforward, and it stops searching when the highest value of an objective function is reached. However, this could be due to local optima in the dataset, but the SFS cannot continue to explore further (Uncu *et al.*, 2007).

For SFS, the objective function $H(S, D, C) = c$ is used to find the best possible subset S of features from the whole dataset, where D denotes the dataset used, C denotes the classification model, c is the criterion function indicating the overall classification performance. SFS starts from an empty dataset, which is sequentially updated by including in each iteration the feature g, which results in the maximal score $H(S, D, C)$. Therefore, the feature set of size = f is given by (Pohjalainen, Räsänen and Kadioglu, 2015)

$$S = S_{f-1} \cup argmax\, H\left(S_{f-1} \cup g, D, C\right) \qquad \textit{Equation 2-1}$$

where g is the feature in the feature set, f.

2.4.6.3 Embedded Methods

This kind of method performs feature selection during the training stage and the method chosen is usually specific to a defined learning machine (Bonilla-Huerta *et al.*, 2016). Consequently, the search for an feature subset is integrated into classifier structure(Li, Meng and Ni, 2008). These methods can identify dependencies at lower computational cost than wrappers. However, they are not suitable for generalisation (Haury, Gestraud and Vert, 2011)s, and commonly they are not preferred in the bioinformatics domain, and therefore are not used in the proposed framework.

2.5 Pareto Optimal (PO)

Given comparable empirical error, it can be said simpler models are more likely to generalise as compared to models that are relatively complex. Inspired from the principle of Ockham's razor, models would be better when designed more straightforward with less number of features (Alpaydin, 2010).

In the real world, problems can involve various conflicting objectives. Thus, the search for solutions to real-life issues can yield multiple options. These options emerge when various criteria are considered at once. Common gene selection objectives are listed as follows (Sabzevari and Abdullah, 2011):

- enhancing the prediction performance of classification models;
- organising less complicated models in terms of the computational resources required and, as a result, providing quicker and more cost-effective modelling; and where
- the purpose is a better understanding of underlying gene-expression data gained by eliminating irrelevant genes and building a more comprehensive dataset.

PO has been applied to various problems such as public transportation (Prakash *et al.*, 2008), sorting/scheduling (Kacem, Hammadi and Borne, 2002; Tavakkoli-Moghaddam, Azarkish and Sadeghnejad-Barkousaraie, 2011) and vehicle routing (Jie and Gao, 2010). In recent decades, the PO approach has been applied to biological features such as genes, biomarkers, and subset selection, and has been successful in selecting the best subclasses from evaluations of the classification performance of many gene subclasses (Fleury *et al.*, 2002; Deb and Raji Reddy, 2003; Chen *et al.*, 2007; Soto *et al.*, 2009; Ting, Lin and Huang, 2010; Sabzevari and Abdullah, 2011). One of the earlier studies used three objective functions of mean slope, slope deviation and valid trajectories to filter genes (Fleury *et al.*, 2002). In another study, two- and three-objective feature selection was performed, taking into account univariate grading constraints such as fold-change, p-value, and selection frequency (Chen *et al.*, 2007). However, the characteristic feature of these studies is that the genes selected were not ranked according to their importance. The Ranking is one of the critical outputs of biomarker selection.

Local learners that use different subsets of data have been created for feature selection to avoid dependency on the diversity of dataset. PO approach is a multi-criteria decision-making method that can conduct feature selection by taking into account the rating values produced by local learners. The vectors created using multiple training sets are used as objective values in the present study.

Feature selection depends on the diversity of the data set, and the difference of the data set variations affects the sorting of the features. Because ranking methods sort them before selecting features, they choose a subset according to

their scores. Besides, ranking methods as local learners, use altered samples of subgroups, they create a different order of importance for features.

Multi-criteria decision methods choose the best solution or most dominant solution sets between the conflicting objectives of real-life problems (Zitzler and Thiele, 1999). Three main multi-objective decision procedures can be found in the literature (Sabzevari and Abdullah, 2011) : aggregation-based strategy, PO-based and non PO-based strategies such as lexicographical methods, and PO-based approaches. Aggregation-based strategy is the widely used approach from three of them for multi-criteria oriented models. This consists of transforming multiple objectives into a single purpose by converting them into a complex scalar function. Thus, the multi-criteria characteristics of the target are abolished. The approach is straightforward and remarkably easy to use. However, it has quite prominent weaknesses, such as predetermining the trade-off between objectives without relying on data content. This independence of target and trade-off leads to the need for intervention according to the intuition of the user, the combining of different units of measurements in one equation and, more importantly, the mixing of incommensurable criteria (Sabzevari and Abdullah, 2011). These drawbacks do not exist when Pareto-based approaches are used.

However, the lexicographical approach can be used to classify different non-comparable objectives utilising ad-hoc parameters determined by an aggregation-based method, it is less desirable as it providing only a single solution to the decision-maker (Sabzevari and Abdullah, 2011)

On the other hand, each objective vector has a set of all elements, and accepts as equivalent value before comparison. The user exerts no influence in predetermining any trade-off or assigning priority to objectives in the PO. The decision-maker receives information on all statistically essential features when a Pareto optimal solution is chosen. The PO gives a meaningful set of possible features which can be analysed further by experts (Zitzler and Thiele, 1999; Sabzevari and Abdullah, 2011).

Sabzevari and Abdullah (2011) narrowed down multi-objective feature selection methods to those that are proposed for gene expression datasets. There

are three different approaches (Aggregation, Lexicographical and Pareto) that can meet multi-objective perspective. Of these, only PO retains solutions as multi-objective outcomes.

In bioinformatics, the expected benefit of discovering disease-related genes is to provide an advantage to molecular biologists in prioritising small numbers of meaningful genes from the thousands of genes in datasets. PO retains all objective-associated features and eliminates all redundant and irrelevant genes without user interaction. This is came from main idea from PO, which is *"when no one can be made better off without making someone else worse off"* (Pardalos and Du, 2008). Gene-expression data contains vital information, which can be gained by various methods, and there may be differences in each sample. Therefore, a molecular biologist should take the final decision concerning each potential disease-related gene. Ultimately, PO retains all potential disease-related genes, which is a more suitable objective than only providing a minimum subset. If PO discovers even one gene in one of the local training sets, it may help in further studies related to the diagnosis of diseases.

2.6 Conclusion

The computational models constructed on gene expression datasets often have a large number of genes and a smaller number of samples. As the models encounter difficulties in processing high dimensionality of the microarray data, feature selection methods were widely used to overcome such difficulties. However, selected genes may differ from one feature selection method to another even though the gene expression dataset remains the same. Therefore, depending on one feature selection method solely for a computational model may miss a feature having a more discrimination affect. Moreover, as each feature selection method may have some pros and cons in their own merit, it may also be hard to choose the appropriate feature selection approach for the computation model. Thus, it is considered that one possible approach that could be used to overcome such limitations is the use of ensemble methods. As compared to use of one method only, ensemble approach might provide more opportunities for finding better set of variables. The multi-objective decision-making method, Pareto Optimality, is adopted as the basis of the proposed gene selection framework PO and is extensively analysed for this purpose. Besides,

classifier performance in real-world gene expression datasets often suffers from issues such as the class imbalance between disease and control samples and also missing gene expression values. However, the literature appears to suggest there is a lack of studies addressing these aspects for gene-expression classifier models. Therefore, class imbalance and missing value problems in microarray experiments are also investigated in this research.

3 MATERIALS AND METHODS

3.1 Introduction

In this chapter, the first type of datasets applied is discussed. Including a description of how it was produced and how it can be useful for this research. Then, ranking methods are discussed followed by a brief description of different imbalanced datasets and imputation methods used for handling missing values. The datasets applied are explained and the main method of Pareto Optimality method is examined in detail.

3.2 Microarray Data Analysis

Microarrays can be grouped as array data in biological datasets. The sequences comprise numeric values. Numerical expressions of micro sequences are express the activation level of genes.

Microarray datasets have existed for two decades and they became challenging for machine-learning methods. It shows the generic characteristic of big data even with small sample sizes that present significant complexity (e.g., class overlaps, imbalanced classes, dataset shift). Establishing a predictive model with usually fewer than 100 samples and thousands of features makes the learning process more laborious. In the literature, several studies show us the most gene-expression values in a DNA microarray experiment are irrelevant in the classification of samples and only a fraction of them are necessary. When a classifier faces numerous number of irrelevant features, its accuracy performance inevitably decreases (Veronica Bolon-Canedo et al., 2015). Likewise, instance-based learners such as K-NN with Euclidean distance metrics are strongly affected by redundant features. A higher number of irrelevant features exponentially increases the demand for training instances to produce a predetermined level of accuracy (Haixiang et al., 2016). Furthermore, the models used have to deal with imbalanced classes, samples extracted under different conditions in both training and test datasets, dataset shift or the presence of outliers (Veronica Bolon-Canedo *et al.*, 2015).

3.2.1 Gene-Expression Data

Genes are sequences of DNA. They create RNA then creates protein and are active in all biological functions in the living cells. Each gene has different roles

in the functions of biological vitality. If a gene is particular of interest from a research point of view, such decisions can be made by looking at the number of proteins that are produced by that gene. This is the definition of gene expression. The frequency of conversion of DNA→RNA→Protein of a gene is called the expression level of that gene. An increase in the level of expression is interpreted as the activation of a gene, and a decrease as the suppression of that gene.

The expression level of genes can vary from cell to cell, as well as from case to case. For instance, a gene expressed in one tissue of our body may differ from another tissue or may not be expressed at all. On the other hand, the same cell can be found different expression level in particular situations. For example, a gene that is usually under-expressed may be more highly expressed during cell division. The level of gene expression may vary according to environmental factors outside the body. For example, a gene expressed at an ordinary level in a plant may be expressed of less or more during the adaptation of the plant to the environment.

Microarray technology is used to determine the level of expression of genes. Changes in the expression levels of thousands of genes at the same time can be detected with the application of microarray techniques. The ability to compare several genes at the same time has enabled this technique to be widely used for different purposes.

Microarray experiments can be arranged for various purposes. For example, the levels of expression of genes belonging to organisms in two different groups, such as patient and control groups are compared. The biomarkers of a disease can be identified by looking at whether or not the genes that are being compared are related to the disease according to their success in distinguishing between patients and healthy people. This technique can also be used to identify altered genes in cancerous as opposed to healthy cells. The presence of genes involved in cancer formation can be identified by investigating the factors that regulate their expression. In a different use of gene expression, levels are taken at different times in the same cell to find genes that vary in appearance according to the interaction of the organism to the environmental environment.

Regardless of the purpose for which the research is conducted, only a small portion of the thousands of genes evaluated maybe related to the situation under investigation. Therefore, selecting the genes associated with the research problem is of great importance in terms of solving that problem. Also, if the microarray data are analysed using machine learning methods, the presence of irrelevant data will both complicate the analysis and decrease the success of the technique. This study aims to develop multi-purpose approaches to improve the feature selection process for machine learning methods, where datasets have a high number of features. Microarray datasets mostly have small sample sizes because of limited diseased and healthy people data. The methods proposed in the study were tested on microarray data, which inherently possesses a large number of features.

Microarray technology and the generation of gene expression data involve many processes. The process of the formation of data is not considered in this thesis. The data used in this study is existing microstructure data ready for analysis. Each row of data used is in a matrix format, which represents a sample of each gene.

The methods proposed in this project to be used on gene expression datasets aims for the correctly classify samples with a smaller number of features. We mostly study with the two-class datasets (e.g., normal and disease) as compared to multi-class datasets.

3.2.2 Imbalanced Data

In real life, imbalanced data exists more frequently than balanced data (Lee and Zhu, 2011), as found in many research domains and also in bioinformatics (Japkowicz and Stephen, 2002). Most observed datasets have two general characteristics; the first is that they have two classes, and the other one is that class distribution is skewed. These problems are called binary and imbalanced classification problems, respectively. Commonly, the minority class is more interesting to research. However, common classification techniques biased towards the majority class on imbalanced datasets. As a consequence, the accuracy performance of the classifier can be very poor. Just because the classification methods are utilised to maximise the performance accuracy across

the entire dataset without considering contribution of minority and majority classes to training (Kim, Chung and Lee, 2017). As minority class has a low impact on outcome, balancing its impact on the outcome should be addressed properly in order to get a better classifier performance. Balancing the dataset classes require analysing the characteristics related to samples and features concerning majority and minority of the cases. Moreover, the imbalance ratio (IR) should also be adequately examined.

Resampling approach (Yang and Mao, 2010) is one of the main strategies suggested in the literature for analysing the imbalanced datasets. Additionally, a numerical arrangement in minority and majority class samples were performed to make the dataset balanced. Feature selection methods can also mitigate the circumstances related to class imbalance problem. The properly selection of subset of variables from a whole set of variables impacts the efficiency of the classifier performance.

3.2.3 Imputation Methods for Data with Missing Values

Numerous studies have been carried out on datasets with missing values concerning to apply imputations on them. In recent years, research studies have crafted improvements on imputation methods.

As in many other types of experimental data, microarray experiments to create gene expression data often include missing values (MVs) (Troyanskaya *et al.*, 2001; Oba *et al.*, 2003; Celton *et al.*, 2012; Chai *et al.*, 2014). Among the reasons why experiments produce such MVs include faults that happen during fabrication or experiments, or microarray image related problems (e.g., image corruption, insufficient resolution).

A basic strategy for approaching the problem of missing values is that simply eliminating samples having them. However, this strategy does not work for microarray data (Souto, Jaskowiak and Costa, 2015). When the samples containing MVs are eliminated from the gene expression dataset, the results of the classification can be biased to one of the classes. This approach can only work for gene expression profiles that contain very few MVs (< 5%). Nonetheless, when a high percentage (>95%) (Haixiang *et al.*, 2017) of the

values of one feature are missing then that variable can considered to be removed.

Generally, statistical metrics were effectively used in algorithms to impute missing values. A widely used statistics approach in this matter is simply imputing the missing values with zeros or feature mean (Alizadeh *et al.*, 2000). One problem with the statistics approaches is that they do not consider correlation between variables (Troyanskaya *et al.*, 2001). However, interestingly, imputation with statistical methods such as getting feature mean or feature median simply reported to be almost equally well in the classifier performance as compared to complicated imputation methods presented in the literature (Souto, Jaskowiak and Costa, 2015).

3.3 Experimental Datasets

3.3.1 Colon Cancer

Colon cancer is a disease when diagnosed in early-stage survival rate is very high. The colon cancer dataset consists of 62 samples and 6500 genes for each sample (Alon, 1999). The dataset consists of colon tissue samples marked as positive (disease) and negative (healthy) groups. One class is the positively marked group, 40 tumour tissues, and the remaining class is the negatively marked group, 22 normal tissues.

3.3.2 DLBCL (Lymphoma)

This dataset consists of 5469 sets of gene expression values for 77 patients (Shipp *et al.*, 2002). These patients are distinguished from each other based on their pre-treatment biopsies. One malignancy group include 58 samples for B-cell lymphoma (DLBCLs), and the remaining malignancy group include 19 samples (control group) for follicular lymphoma (FLs).

3.3.3 DUKE Breast Cancer

This dataset contains 7129 genes for 44 samples. They are classified according to oestrogen receptor ER+ (23) and ER-(21) (West *et al.*, 2001).

3.3.4 CNS (Central Nervous System) Embryonal Tumour

Central Nervous System (CNS) Embryonal Tumour is a disease with a high mortality rate. CNS Embryonal Tumour dataset that we study consists of sixty

patients, and 7129 genes for each sample (Pomeroy *et al.*, 2002). The dataset separated into two main groups, alive and not alive. One class is the survivors, twenty-one patients, whom were alive at the end of the treatment. The remaining class, thirty-nine patients, are those that were died during the treatment.

3.3.5 Ovarian Cancer

Ovarian cancer is a woman disease where early diagnosis increases the survival rate. The ovarian cancer dataset consists of 253 samples, 6000 genes per sample (Petricoin *et al.*, 2002). The dataset separated into two main groups, cancer and not cancer. One class is the normal group, 91 participants, whom were the control samples. The remaining class, 162 participants, are those that were diagnosed with the ovarian cancer disease.

3.3.6 DLBCL (Blood) Cancer

Diffuse Large C B-cell Lymphoma (DLBCL) is a blood cancer disease caused from abnormal B-cell growth with a modest survival rate. The DLBCL dataset consists of 47 samples and 4026 genes (Alizadeh *et al.*, 2000) The dataset separated into two different DLBCL types, "germinal centre B-like" and "activated B-like". One class presents the gene expression characteristics of the germinal centre B-like DLBCL group, 24 patients, and the remaining class presents the gene expression characteristics of activated B-like DLBCL group, 23 patients.

3.3.7 Brain Cancer Data

The brain cancer dataset consists of 50 samples, 6706 genes per sample (Bredel *et al.*, 2005). The dataset classified into three glioma subtypes. First glioma subtype is the pure glioblastomas, thirty-one samples. Second glioma subtype is the oligodendroglial morphology, fourteen samples. The remaining glioma subtype is the grade 1-3 astrocytomas, five samples.

3.3.8 Lung Cancer Data

Lung cancer is a disease with a very low survival rate. The lung cancer dataset consists of sixty-six samples, filtered 3312 (12600 original genes number) genes per sample (Garber *et al.*, 2001). The dataset morphologically classified into four groups. First group is the squamous cell carcinomas, seventeen samples. Second group is the large cell lung cancers, five samples. Third group is the

small cell lung cancers, four samples. The remaining group is the adenocarcinoma, forty samples.

3.3.9 Prostate Cancer Data

Prostate cancer is a disease with high survival rates. The prostate cancer dataset consists of 112 samples and ~26,000 genes (Lapointe *et al.*, 2004). The dataset has three classes. First class contains the control group, forty-one healthy samples. Second class is the disease group, sixty-two prostate cancer samples. The remaining class is another disease group, nine lymph node metastases samples.

3.3.10 Endometrium

Endometrial cancer is a women disease with high survival rates. The endometrium cancer dataset consists of forty-two samples and X genes per sample (Risinger *et al.*, 2003). The dataset classified into four groups. First group is serous papillary, thirteen samples. Second group is the clear cell, three samples. Third group is endometrioid cancers, nineteen samples. The remaining group is the seven age-matched normal endometria.

3.4 Data Validation

The datasets are first separated into training and test parts in order to avoid mixing samples. This separation is justified in the literature (Bolón-canedo *et al.*, 2014). Data split ratios selected from the literature are the widely used 80% training and 20% test (Ding and Wilkins, 2006), and to compare learning levels 60% training and 40% test (Luo *et al.*, 2011). This process is repeated 25 times, so that 25 dataset variations are gained. Then, K-fold cross-validation is utilised for training which could be 5-fold, 10-fold and Leave-one-out (LOO). This high amount of variety and three cross-validations creates multiple altered conditions employed in the proposed multi-criteria decision framework to scrutinise stability given such conditions.

3.5 K-Fold Cross-Validation

K-fold Cross validation is a method that splits the dataset into k different subdivisions where one subdivision used as the testing set whereas the remaining $K-1$ subdivisions used for training the computational model. Blagus (2015) emphasised that it is essential to apply the correct cross-

validation (CV) in feature selection and classification. If the data is not split into training and test proportions to validate feature selection, CV can often be used as a measure of performance (Webb *et al.*, 2010)., training should be adequately separated from the test on the feature selection process. CV should be used on the training data. Then, the test data should only be used for classification purposes.

This research study used the cross validation as to resample the datasets to evaluate the classifier as well as the feature selection performance. The 5-fold, 10-fold and leave-one-out cross-validation methods are applied to test how filtering methods and a wrapper method perform in different training folding circumstances.

3.6 Classification

The k-nearest neighbour (k-NN) is one of the most widely used non-parametric classifiers (Chuang *et al.*, 2011). It is not only conceptually straightforward and easy to implement but also computationally efficient. Noise in data volumes does not seriously affect its performance (Cover and Hart, 1967). It was deemed a suitable classifier for use in this study.

3.6.1 The k-nearest neighbour (k-NN)

k-NN is a distance-based classifier in which Euclidean distance is commonly used as the distance metric, and there is a potential risk of tied results when even numbers are used. If this happens, tied results are solved using a random procedure. Thus, the use of even number was avoided. Odd numbers of neighbourhoods were applied to avoid any tied results from the classifier and the K values are selected ranged from 1 to 11. Applied distance metric is Euclidian metrics.

3.6.2 Support Vector Machine (SVM)

Support Vector Machines is a learning machines-based classifier, emerged from statistical learning theory (Vapnik, 1999). SVM, basically separates samples into two groups aiming to maximize the margin between them (Figure 3-1). Margin based classification for finding the optimal separation hyperplane (OSH) is defined as follows.

Minimise:

$$\frac{1}{2} \| w \| + C \sum_{i}^{n} \xi_i$$

Subject to:

$$(w^t x_i + b) y_i \geq 1 - \xi_i, \xi_i \geq 0, i = 1, \ldots, n$$

minimize: $$\frac{1}{2} \| w \| + C \sum_{i}^{n} \xi_i \quad (1)$$

subject to: $$(w^t x_i + b) y_i \geq 1 - \xi_i, \xi_i \geq 0, i = 1, \ldots, n \quad (2)$$

where: x is the set of features,

n is the length of samples,

y is the class,

w is the weights of the hyperplane,

b is the bias of the hyperplane,

C is for the regularization,

ξ_i is the slack variable which is a value indicating deviation of samples from the hyperplane.

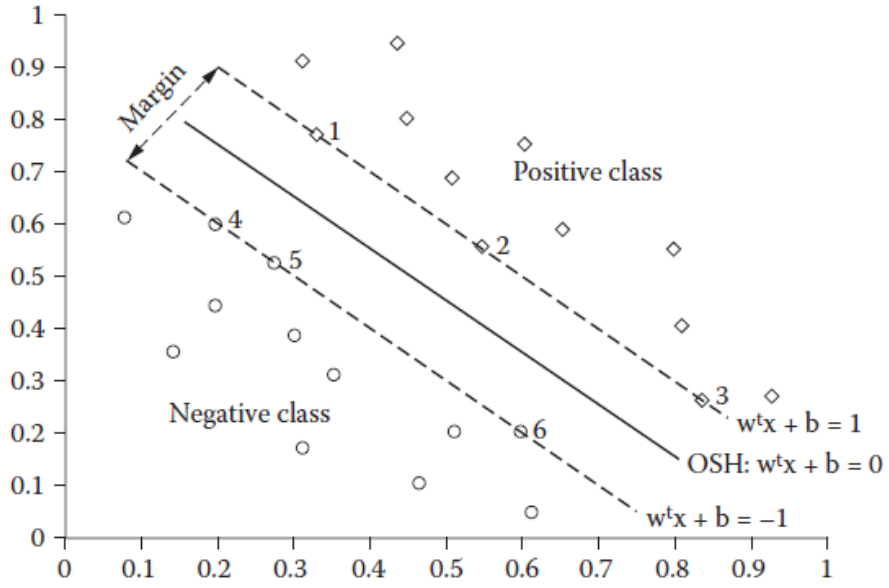

Figure 3-1 Illustrative example of SVM classification (Liang et al.*, 2016).*

3.7 Ranking Methods

t-test, Entropy, Bhattacharya, ROC, Wilcoxon tests will be used as the ranking methods for producing feature subsets. The definitions of these methods are summarised below.

3.7.1 Two sample *T*-test

The t-test is clearly the most popular test, and it is a parametric test. It is applied to determine if the average difference between data for two independent variables is significant(Fox and Dimmic, 2006).

The t-test is expressed by;

$$t = (\mu_1 - \mu_2) / \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \qquad \textit{Equation 3-1}$$

Using a t-statistic can be challenging because genes having a very low variance can skew variance estimates (Tusher, Tibshirani and Chu, 2001). Another disadvantage is its applications on a small number of samples such as the microarray datasets. In such datasets it is reported that t-test may fail to select optimal set of genes (Murie *et al.*, 2009). Therefore, the power of a t-test has been questioned along with the importance of variance modelling (Mary-Huard, Picard and Robin, 2006). A variable in the dataset might have identical means but may differ in variance among classes. In such a case, t-test can miss that variable even though the variance might be significant for the analysis of the dataset (Baldi and Brunak, 2001). These issues have led to the proposal of various alternatives with the aim of better accuracy in variance estimation.

3.7.2 Entropy

In its non-parametric form, relative entropy, also known as Kullback-Liebler distance is defined as follows:

$$e = \frac{1}{2}\left[\left(\frac{\sigma_1^2}{\sigma_2^2} + \frac{\sigma_2^2}{\sigma_1^2} - 2\right) + \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)(\mu_1 - \mu_2)^2\right] \qquad \textit{Equation 3-2}$$

After the statistical calculation is completed for every gene, genes with the highest to lowest entropy values are listed to serve as inputs to PO.

3.7.3 Bhattacharyya

This is similar to Kullback-Leibler distance but considers a different distance metric between two distributions.

The Bhattacharyya coefficient (Guorong, Peiqi and Minhui, 1996) is calculated as:

$$r_k = -ln\left(\int \sqrt{N(W_k^+)N(W_k^-)}dx\right) \qquad \textit{Equation 3-3}$$

Under the Gaussian assumption, this reduces to:

$$r_k = \frac{1}{8}\frac{(n_k^+ - n_k^-)^2}{\hat{\sigma}_k^2} + \frac{1}{2}\ln\left(\frac{\hat{\sigma}_k^2}{\sqrt{\hat{\sigma}_k^+\hat{\sigma}_k^-}}\right)$$

where $\hat{\sigma}_k^2 = \frac{(\hat{\sigma}_k^+)^2 + (\hat{\sigma}_k^-)^2}{2}$

The output ranking list becomes the input for Pareto optimality.

3.7.4 Receiver Operating Characteristic (ROC) Analysis

The receiver operating characteristic (ROC) curve helps to interpret binary classification tasks with the proportion of true positives to false positives, while the difference in the threshold value varies (Metz, 1978; Fawcett, 2006). The true positive (sensitivity) and the false positive (1-specificity) rates are located on the ROC curve as points for different threshold values on the vertical axis and the horizontal axis, respectively (Bradley, 1997).

Table 3.1 Confusion Matrix

		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative	False Positive
	Positive	False Negative	True Positive

At the end of a binary classification task, four possible outcomes are occurred using a confusion matrix: true positive (TP), false positive (FP), true negative (TN), false negative (FN). Their formulas are as follows.

TP = Correctly classified positive samples

TN = Correctly classified negative samples

FP = Incorrectly classified positive samples

FN = Incorrectly classified negative samples

The outcomes are required information to calculate true positive rate (TPR) and false positive rate (FPR). Their formulas are as follows (Fawcett, 2006).

$$TPR = \frac{TP}{TP+FN} \qquad \textit{Equation 3-4}$$

and

$$FPR = \frac{FP}{FP+TN} \qquad \textit{Equation 3-5}$$

Then TPR and FPR are used to plot the ROC curve and calculate the AUC. The ROC curve can be used to rank feature subsets for the Pareto optimal models. The models are tested based on the area under the curve (AUC) aiming to find the gene subsets having the highest AUCs. The figure illustrates how the ROC curve is plotted.

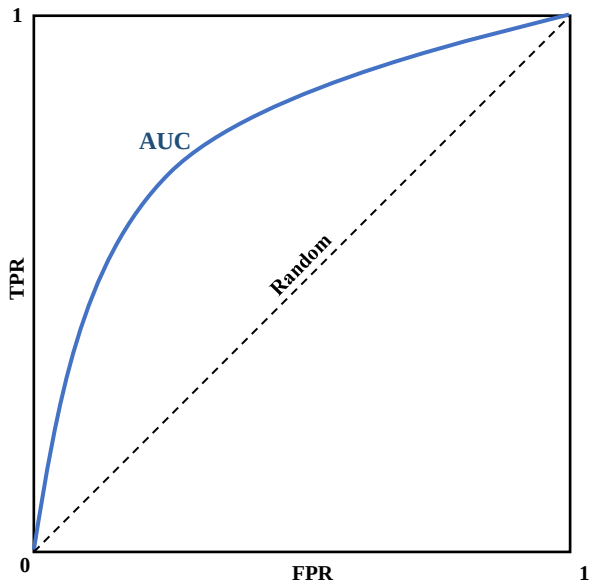

Figure 3-2 An illustrative example of Area Under Curve

3.7.5 Wilcoxon

Mann-Whitney U-Test (or Wilcoxon Rank-Sum Test) differs from the statistical tests being non-parametrically testing equality of two similarly distributed populations. Without fulfilling the parametric test assumptions, the duration of the significance test of the difference between the two means can lead to reaching the conclusion (Nguyen *et al.*, 2015). The Wilcoxon rank-sum test is used as alternative statistical filter test. This test differs from the above-mention methods being a non-parametric and hence no assumption is made about the distribution of the date.

Using two formulas, two U-statistics candidates are calculated. From these, U_1 The number of observations and total sequence number for Sampling 1; If U_2 uses the number of observations and the total number of sequence numbers for Sampling 2. The formulas are:

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2} \qquad \textit{Equation 3-6}$$

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2} \qquad \textit{Equation 3-7}$$

Where n_1 is the sample size for Sampling 1; R_1 Sum of sequence numbers for Sampling 1; n_2 Sample size for sampling 2; R_2 is the sum of the sequence numbers for Sampling 2. For the control, the sum is taken for $U_1$and U_2. This value should be equal to the product of the two sample volume numbers; so

$$U_1 + U_2 = n_1.n_2 \qquad \textit{Equation 3-8}$$

The U-statistic, which is less than the U_1 and U_2 values found, is used in the significance table. If the sample volumes are large, the following standard normal distribution approach is used to find the level of significance:

$$z = (U - m_U)/\sigma_U \qquad \textit{Equation 3-9}$$

Where z is the z-score used in standard normal distribution tables; If m_U and σ_U U if the null hypothesis is true, then the mean and standard deviation for U is They are found by the following formulas:

$$m_U = n_1.n_2/2 \qquad \textit{Equation 3-10}$$

$$\sigma_U = \sqrt{\frac{n_1.n_2(n_1+n_2+1)}{12}} \qquad \textit{Equation 3-11}$$

3.8 Pareto Optimality (PO) Method

Multi-objective optimisation can involve the maximisation or minimisation of a vector function g. The target vector function contains a group of m parameters (decision variable) and a group of n objectives. PO transforms a scalar objective into a vector and can be expressed as follows:

minimise or maximise the function:

$$z = \left(g_1(t), g_2(t), \ldots, g_n(t)\right) \qquad \textit{Equation 3-12}$$

subject to:

$$\begin{matrix} t=(t_1,t_2,\ldots,t_m,)\in T \\ z=(z_1,z_2,\ldots,z_n,)\in Z \end{matrix} \qquad \textit{Equation 3-13}$$

where m and n are the numbers of genes(parameters) and responses(objectives) respectively, and $t = (t_1, \ldots, t_m)$ and $z = g(t)$ are determined as the decision and objective vectors, where T and Z represent the parameter and the objective spaces (Zitzler and Thiele, 1999). The solution set includes all decision vectors, which have a minimum of one optimum objective value. The collective set of non-dominated solutions are known as PO solutions. A non-dominated solution is where the values of target vector functions came to their limits and no more optimization could be possible based on the given parameters.

If the target vector function is to maximise the objective under (3), β dominates γ, or in other words γ is dominated by β:

$$\forall_i \in \{1, 2, \ldots, x\}\, g_i(\beta) \geq g_i(\gamma) \ \wedge\ \exists_k \in \{1, 2, \ldots, x\}\ g_k(\beta) > g_k(\gamma).$$

Equation 3-14

Figure 3-3 shows an illustration of PO solutions where there are two cases. If a solution has maximum values for all objectives, it is the optimum and ideal solution. As seen in Figure 3-3.(a), the optimal solution covers all other solutions for all objectives. However, if a solution has a maximum value for at least one objective and up to the most n-1 objectives, it is one alternative solution. Likewise, a multi-criteria selection case with two objectives that are maximised and five solutions is represented in Figure 3-3(b). Four solutions (A, B, C, and E) are not dominated by any other solution, so the non-dominated solutions are Pareto-optimal solutions. However, solution D is dominated by

solution E in both objectives, and so it is not a Pareto-optimal solution. It can be expressed in two ways, E dominates D, or D is dominated by E. If solution E does not exist, then solution D is a Pareto-optimal solution.

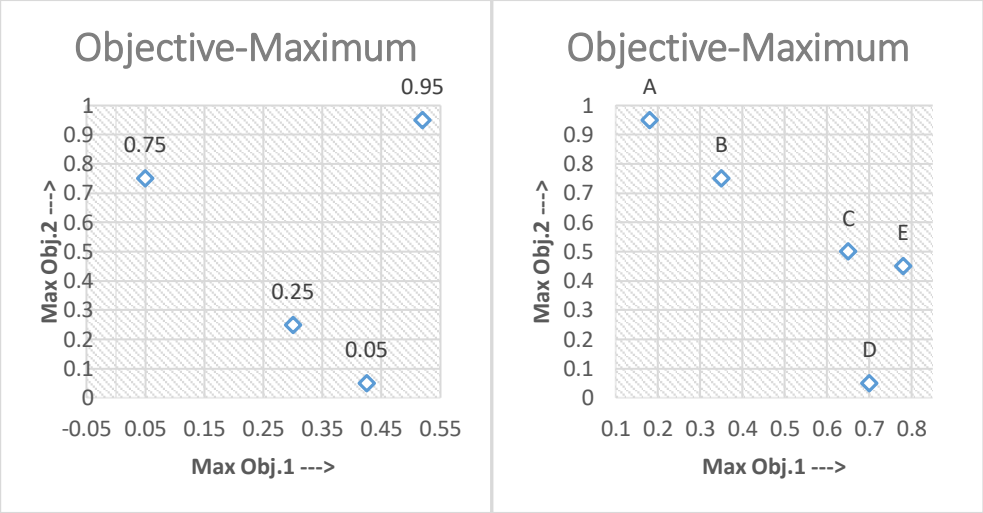

Figure 3-3. Demonstration of a maximisation case with two objectives.

If multiple methods are performed and all of them are combined, this creates a hypersphere of features for PO. If one method is performed and combined with different samples of a same dataset, it creates a hyperplane of features for PO. Multiple methods are creating as possible as major class divided into the equal size with minor class, trade-off/ boundary sensitivity will become healthier.

Four objectives example for PO

To simplify the PO, there is an example with 4 objectives and 6 samples. For the PO example specifically, features in a dataset are arranged in an optimal form, without affected from further parameter changes.

There two aims we want to from an optimal outcome:

1. One outcome maximises total surplus
2. One outcome is preferred by a feature over all other sets of feasible subsets

Pareto optimal response is achieved when there is no room providing a better feature set that is adding or removing a feature will produce negative outcomes from the reached point.

Table 3.2 Simplified 4 Objectives and 6 Samples Dataset Example for PO

Features	P_1	P_2	P_3	P_4		Total Surplus		
Feature A	30	55	30	50	→	165	→	Not $\underline{PO}$
Feature B	25	30	65	60	→	180	→	$\underline{PO}$
Feature C	30	55	55	90	→	230	→	$\underline{PO}$
Feature D	15	20	75	90	→	200	→	$\underline{PO}$
Feature E	10	15	65	55	→	145	→	Not $\underline{PO}$
Feature F	10	35	35	95	→	185	→	$\underline{PO}$

P_n named our samples as Person (Sample) and each row represents of comparison vector. As seen on Table 3.2 presents for each sample between P_1toP_4 for each of available features A, B, C, D, E and F.

Basically, if these 4 individuals' feature expression values compare, they can be identified as healthy or disease sample according to expression values. Our objective is here according to higher activation number find out Pareto Optimal solution.

Pareto optimal feature subset is achieved after a variety of candidate feature subsets are tested and when no further improvement can be possible. Explanations written in two ways because of multi dimension makes harder to understand PO process, first according to mathematical calculations and secondly distinguishing value and meaning according to PO.

Mathematically;

First, Feature A→B (A to B), A→C, A→D, A→E, A→F expression values compares one by one. Feature B could not dominated A but Feature C dominates feature A in all dimensions so A is dominated and eliminated. Also, D,E,F stays and continue to compare feature vectors.

Continue with B,C,D,E,F.

Second, Feature B→C, B→D, B→E, B→F expression values compares one by one. B dominates feature E in all values so it is eliminated. Rest of the features could not dominated B so B stays in PO subset.

Continue with B,C,D,F

Third, C→D, C→F expression values compares one by one. They could not dominate each other. Thus, there is no elimination.

Lastly, D→F compared. They could not dominate each other. Thus, there is no elimination.

Final PO subset is feature B, C, D, F.

In other words, PO feature selection steps explained with numeric and logical information as below.

- Feature B is in PO subset; B→C (B to C), For the P_3 sample goes from 65→55 (B dominated C for Sample P_3).
- C is in PO subset 30 and 55 are max features for sample P_1 and P_2 respectively (C is dominated (distinguishing) feature to identify P_1 and P_2).
- D is in PO subset from D→C, Sample P_3 goes from 75→55 (D is dominated C for sample P_3 it is useful to distinguish the sample).
- From feature A→C expression values for all samples equivalent or higher so A is not in the PO subset
- From feature E→D, all expression values higher so E is not a PO solution. D dominated E in all expression values. Thus, E is not useful to distinguish any sample.
- Until this step feature B, C, D become dominant genes. Until this step, there is no distinguishing feature for P_4 yet.
- From feature F→B or F→C or F→D, Feature F dominates rest of the subset 95 but not for other samples. Thus, F added to PO subset, but no other features removed from final subset.
- Final PO subset is feature B, C, D, F.

3.9 Performance Evaluation

To evaluate the classification performance of the proposed framework, the following formula will be used.

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

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Specifity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

where TP is the correctly classified positive samples, TN is the correctly classified negative samples, FP is the incorrectly classified positive samples, and FN is the incorrectly classified negative samples.

3.10 Conclusion

This chapter presented an overview of the Pareto Optimal methodology that is utilised throughout this research. In this chapter also described the datasets, the ranking methods, and statistical validation and performance evaluation techniques that are employed in this research to evaluate the proposed framework performance. The effectiveness of the proposed framework tested on ten datasets and three data types. The next chapter investigates areas for improvement in current gene selection practices in the bioinformatics.

4 A PARETO OPTIMAL MULTI-OBJECTIVE FRAMEWORK FOR AGGREGATED DISCOVERY OF DISEASE-RELATED GENES

4.1 Introduction

One of the most significant challenges in bioinformatics is to identify the correlation between genes and diseases. Microarray data is particularly valuable as it contains a large amount of information about of genes. However, in general, only a small number of genes have a significant relation with certain diseases. During the last decade, a large amount of research has been performed in this area (Nguyen *et al.*, 2015). The objective of microarray data classification is to create an efficient and practical framework that can differentiate between gene expression in samples so that they can be determined as healthy or diseased. (Chuang *et al.*, 2011).

4.2 The Problem of Gene Selection

Various gene selection studies try to explain the occurrence of cancer in terms of a few genes present in microarray datasets, and their common objective is simply to maximize accuracy (Saeys *et al.*, 2007). However, efficiency should not be the only objective; the complexity of the method should also be taken into account. Furthermore, objectives should also include robustness against data variation in different datasets of new samples (Bolón-Canedo, Sánchez-Maroño and Alonso-Betanzos, 2016) as well as achieving the most comprehensive biologically meaningful subsets (Sarac *et al.*, 2015).

Feature selection methods are the most common techniques used to identify essential genes from a large set of genes (Bolón-canedo *et al.*, 2014). Filter and wrapper methods are the application models generally used for gene selection (Bolón-Canedo, Sánchez-Maroño and Alonso-Betanzos, 2016). Such feature selection methods have both strengths and weaknesses (Saeys *et al.*, 2007), and no technique is considered universally useful in all situations (Ang *et al.*, 2016; Bolón-Canedo, Sánchez-Maroño and Alonso-Betanzos, 2016).

Filter methods are considered to be one of the best options to reduce feature dimensionality and thereby alleviate the need for the computational resources

required (Bolón-Canedo, Sánchez-Maroño and Alonso-Betanzos, 2012). However, they have lower accuracy as compared to wrapper methods because they provide only general subsets (Saeys *et al.*, 2007; Wang *et al.*, 2013; Bolón-canedo *et al.*, 2014). Moreover, utilising these techniques can lead to the significant problem of over-fitting, as the genes identified during training may not necessarily fit well with the test data (Yang *et al.*, 2015).

Additionally, one of the most challenging topics when applying these methods is their robustness. Most recent filter techniques focus on information theory. An alternative approach is to use wrapper techniques, but due to their high computational cost and the risk of over-fitting, these are generally avoided (Bolón-canedo *et al.*, 2014). Nonetheless, wrapper techniques provide the best possible subset for a particular type of model and allow the relationships between features to be evaluated. However, they are considered to introduce bias based on the selection of subsets and are computationally intensive with a high risk of overfitting (Jovic, Brkic and Bogunovic, 2015). Therefore, recent reviews have mentioned that there is a potential to improve accuracy and robustness through the use of ensemble technique which combine different methods (Saeys *et al.*, 2007; Bolón-Canedo, Sánchez-Maroño and Alonso-Betanzos, 2012, 2014).

4.3 Ensemble Feature Selection Methods

Ensemble feature selection methods use either different samples reflecting variations of the dataset, or various methods with the same dataset (functional diversity), or have a mixture of both (Bolón-Canedo, Sánchez-Maroño and Alonso-Betanzos, 2016). Data diversity may cause a shift in the dataset (Bolón-canedo *et al.*, 2014), which appears when the input and output pairs differ between training and test datasets. That means during the split of the data into training and test datasets, critical samples are that grouped in one-fold could characterize one class among the remaining classes. Thus, results from individual folds of the training set may not represent the data accurately enough for successful feature selection. Functional diversity provides many sets of results, and hence the aggregation of such multiple results does not preserve valuable information. As a result, there is a need to utilize both data diversity

and functional diversity so that more comprehensive and representative subsets can be identified.

Ensemble methods are widely proposed for classification (Bolón-Canedo, Sánchez-Maroño and Alonso-Betanzos, 2014), and they tend to provide more robust results than single versions of any one method (Seijo-Pardo *et al.*, 2015). An advantage of the ensemble approach as compared to the feature selection is to eliminate the need for computational models to seek for a feature selection method fitting to their datasets. (Bolón-Canedo, Sánchez-Maroño and Alonso-Betanzos, 2012). As a result, ensemble approaches have become widespread and are now commonly found in bioinformatics applications.

Various aggregation techniques are described in the literature, such as majority voting, stacking and cascading, bagging and boosting (Alpaydin, 2010) to name only a few. The aggregation technique in an ensemble method is an essential part of the disease-related gene selection process. However, the literature appears to suggest that there is a lack of robust multi-criteria aggregation methods for gene selection (Bolón-Canedo and Alonso-Betanzos, 2019). Existing ensemble methods have a limited ability to make decisions during the selection of genes because they mostly rely on voting when aggregating the results of the method used. Voting aggregation may eliminate essential disease-related genes and is limited in the ability to identify biologically valuable genes (Montague and Aslam, 2004; Wu *et al.*, 2009).

A small variation in the training data set can significantly alter gene selection and thus can significantly affect the stability of the system (Nguyen *et al.*, 2015). For example, genes identified using feature selection from 10-fold cross-validation of training data can significantly differ from those when using the 5-fold cross-validation of the same training data. Luo et al. (2011) mentioned that previous studies revealed that the feature selection approaches chosen are more crucial than the classification methods used. This study focused on optimising the results of multiple methods applied to different size of training sets. Partitions are prepared similar to the study of Lue et al. We proposed the partitions as 40% / 60% and, 20% / 80% for test and training sets, respectively.

In this chapter, firstly a novel framework called Model-1 is proposed using PO for gene selection in order to solve the overfitting of the classifier and stability problems in feature selection. Then the proposed framework is used in two different set of partitions as mentioned above providing better results in each step. PO, which was originally used in engineering and economics, provides one of the best trade-off according to the objective function involved (Shoval *et al.*, 2012). The aim of this study is, therefore, to investigate PO as part of an ensemble feature selection framework combined with filter techniques and *K*-fold cross-validation for the selection of genes from gene-expression microarray datasets. Then, the feature subsets extracted are combined as a consensus gene subset. The results of the proposed PO-based framework are shown to yield better stability than that of common single filter methods. Secondly, this work focuses on the development of a multi-objective aggregation framework for gene selection. The framework involves aggregating the different results obtained from a wrapper method in a novel way using Pareto optimisation. PO is utilised for aggregating and evaluating the result of the wrapper method. It selects disease-related genes according to their scores without any user intervention. PO collects information of all genes throughout different data variations and eliminates genes having insufficient discrimination ability while retaining genes efficient discrimination ability. Wrapper methods may provide different results for each variation of training data and may over-fit to it's test data. Meanwhile PO collects individual results and combines them in a common subset with only non-dominant genes. Genes in the common subset cover all training variations and are more robust to alterations. In this way, the overfitting problem is mitigated.

The aim of the present study is to construct a novel cross-validated framework that provides stable and robust generalised aggregated decision-making ability owing to the use of PO. PO combines random inputs and creates a single output. Furthermore, the results of PO are applicable to the aforementioned random feature subsets (Li, Wu and Hu, 2008). By using PO, an efficient and feature selection process selects genes straightforwardly. It will also be utilised to apply the classification process to just an ensemble subset of training datasets of microarray gene expression data (Yang and Mao, 2011). The goal is to provide

a consensus result, which will be independent of any classifier (such as K-nearest Neighbours, or a Support Vector Machine).

4.4 PO-based Multi-Criteria Framework

The proposed framework is designed to include several different feature selection methods combined with Pareto optimality. Two distinct frameworks where PO is used to combine the results of the selected features are developed and compared with traditional feature selection methods, the details of which are discussed below.

Filter and wrapper methods are common feature selection techniques. Filter techniques rank features based on their relationship with a univariate scoring metric with a class label (Hong and Cho, 2009). In this study, several well-known methods are combined with PO; namely, the t-test, Entropy, Bhattacharya, ROC, and Wilcoxon test. For the sake of comparison, they are used one by one in the proposed framework to create rankings of genes related to each dataset. The standalone results of these filter methods without PO are also recorded to compare with those from the proposed framework.

The classes in each dataset are divided into five subsets of samples. This process is repeated 5 times. Thus, the different variants of the test and training parts contain independent samples. Each variant is repeated ten times to discover the performance of the tests in a more sensitive way.

The first study used a ratio of 20% test and 80% training set sizes. As in the study by (Luo *et al.*, 2011), the second study used 40% test and 60% training sets. These two experiments were carried out to observe the effect of different partition sizes for models with and without PO.

The preparation of datasets for the wrapper method included the following steps. 25 different data splits were created from the original datasets, and each training-test split contained independent samples. 80% training and 20% test sizes were utilised for the first study. Moreover, to compare the effect of training–test ratio, a 60% training and 40% test ratio was used as in previous research (Luo *et al.*, 2011). Resampling performed with cross-validation is repeated ten times and then performance of the proposed framework was evaluated. When high-dimensionality becomes an issue in the use of ensemble

methods, filter methods are selected as a preliminary solution to decrease the number of features (Bolón-Canedo, Sánchez-Maroño and Alonso-Betanzos, 2012, 2014; Ogutcen *et al.*, 2016). Features were ranked using ensemble filter methods to gain a relatively small but comprehensive feature pool for SFS (Ogutcen *et al.*, 2016). Their rankings were used by PO to create the common feature pool. Filter methods are useful due to their rapid computation and scalability, but unlike wrapper methods they do not take into account feature dependencies (Saeys *et al.*, 2007).

4.5 Integration of the Wrapper Method with the Proposed Framework

Second stage with the proposed framework introduces a novel aggregation approach for wrapper-based ensemble methods. Sequential forward selection SFS is utilised for feature selection, and Pareto optimality is used to aggregate the selected features. The critical element of aggregation by PO is to consider all of the data together and to collect all of the results into one common solution set. Moreover, in this framework, PO is not dependent on the feature selection method used. Any other feature selection method or indeed different feature selection methods can be used in this framework.

SFS is a common and straightforward wrapper technique, and so it is employed here to demonstrate how PO improves the efficiency of feature selection methods. Unlike voting methods, PO does not rely on a majority decision. It considers all expert decisions as a vector and evaluates all possible solutions in multi-dimensional space. In this decision space, if a solution is not dominated by other features throughout all training set variations, that gene is selected in the common subset. This common subset is called the PO solution.

For comparison, many results come from either different local data variations or different decisions from experts utilised (individual methods) with the same data. In other methods, when an expert has been added to an ensemble model, each feature is selected according to the voting method. If a possible disease-related gene is ignored by the majority of experts, this useful gene cannot be chosen. Hence, it cannot contribute to the classification results. However, one

expert or a minority of experts may have discovered that gene, but the ensemble structure may not use all expert decisions.

On the other hand, when a feature is selected by one of the experts in at least one training variation, and if it is statistically crucial at any level and not dominated by other selected features, PO considers that feature in the common subset. PO does not eliminate possible features as in voting methods. Thus, the PO-based multi-criteria framework keeps all non-dominant genes containing all statistically meaningful information. This makes PO a better and more effective aggregation method, which can produce a more comprehensive subset. To show the advantages of PO in the proposed framework, SFS is employed individually, and the SFS results are presented for comparison. For classification, the K-nearest neighbours (*k*NN) is used which is one of the most commonly used and straightforward classifiers (Chuang *et al.*, 2011).

4.5.1 Cross-Validated Aggregated Gene Selection using Filter Methods

The proposed framework uses several different feature selection methods together as well as PO in Model-1 as illustrated in Figure 4-1. Model-2 as illustrated in Figure 4-2 is an unassembled version of Model-1, which is created to show how the ensemble filter methods perform using a single feature selection method. Meanwhile Model-3 as illustrated in Figure 4-3 focuses only on the performance of the filter tests. The same assessment as in Model-1 is conducted for this model without Pareto Optimality. There is no evidence that it was compared with traditional single feature selection methods or, most importantly, that cross-validation was carried out to assess the generalisation ability of the model independently. After the resampling of the data is completed, %20 of the data were randomly selected, and average classification accuracy values for both training and test datasets were evaluated. Details of all the models are given as follows.

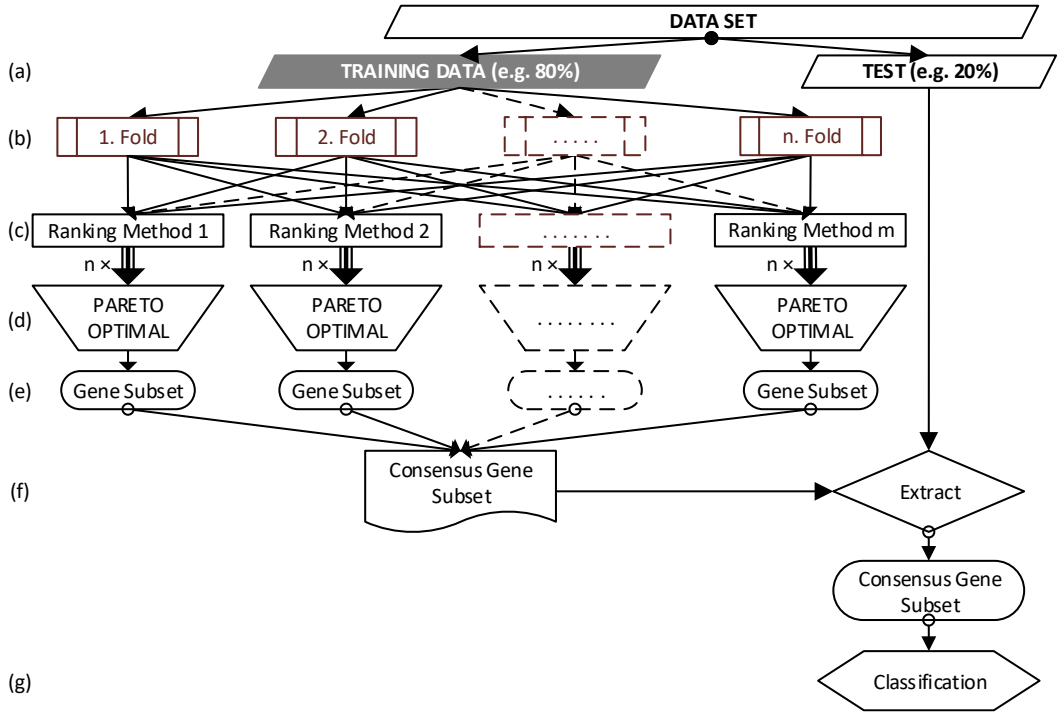

Figure 4-1. Flowchart of the proposed framework using Pareto Optimality on paired-up features of Filter Methods (Model-1)

Model 1:

a) The dataset is divided into training and test cases and different dataset (20% test and 80% training or 40% test and 60% training).

b) *K*-fold (5, 10 and LOO) cross-validation is applied to the training data sets (80% and 60%).

c) The filter methods t-test, entropy, Bhattacharya, ROC, and Wilcoxon are used to rank the features. Each filter method is applied to each fold and ranked features are obtained. For the applied filter method, the ranked features of n-folds paired up.

d) The paired-up features from each filter test are further combined using PO (Figure 4-1). This will result in one subset of features.

e) The final gene set is formed at the end of PO process. The duplicated genes that reside in the final gene set will then be removed. This will yield unique subset of a consensus gene subset.

f) The genes that derived from the training data (consensus gene subset) using the PO analysis is validated using the test data. The test datasets (20% and 40%) are formed in two separate cases and organised using the consensus gene subset to validate our framework (aggregated gene selection using filter methods).

g) The *k*NN is applied as the classification technique to evaluate our framework.

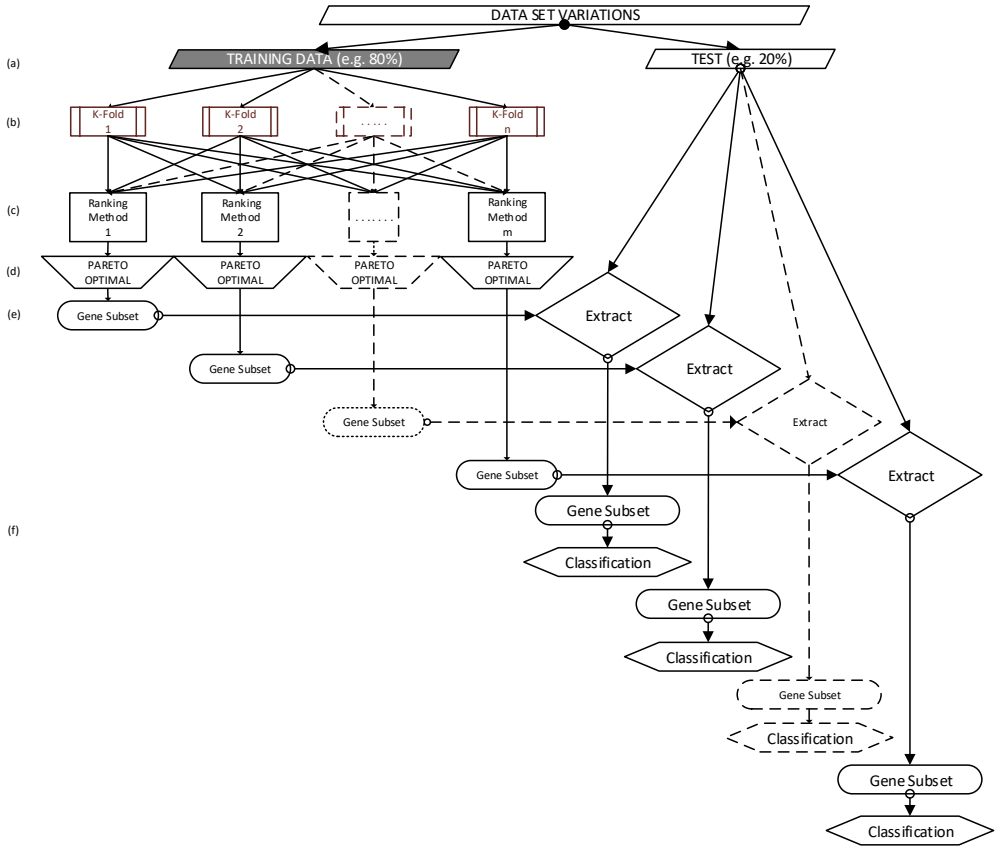

Figure 4-2 Flowchart of the proposed framework using Pareto Optimality on paired-up features of the applied Filter Method (Model-2)

Model 2:

a) The dataset is divided into training and test cases and different dataset (20% test and 80% training or 40% test and 60% training).

b) *K*-fold (5, 10 and LOO) cross-validation is applied to the training data sets (80% and 60%).

c) Different than the Model-1, the filter methods t-test, entropy, Bhattacharya, ROC, and Wilcoxon are solely used to rank the features.

Each filter method is applied to each fold and ranked features are obtained. For the applied filter method, the ranked features of n-folds paired up. The paired-up features of the applied filter test are combined using PO (Figure 4-2). This will result in one subset of features.

d) The final gene set is formed at the end of PO process.

e) The genes that derived from the training data using the PO analysis is validated using the test data. The test datasets (20% and 40%) are formed in two separate cases and organised using the consensus gene subset to validate our framework (aggregated gene selection using filter methods).

f) The *k*NN is applied as the classification technique to evaluate our framework.

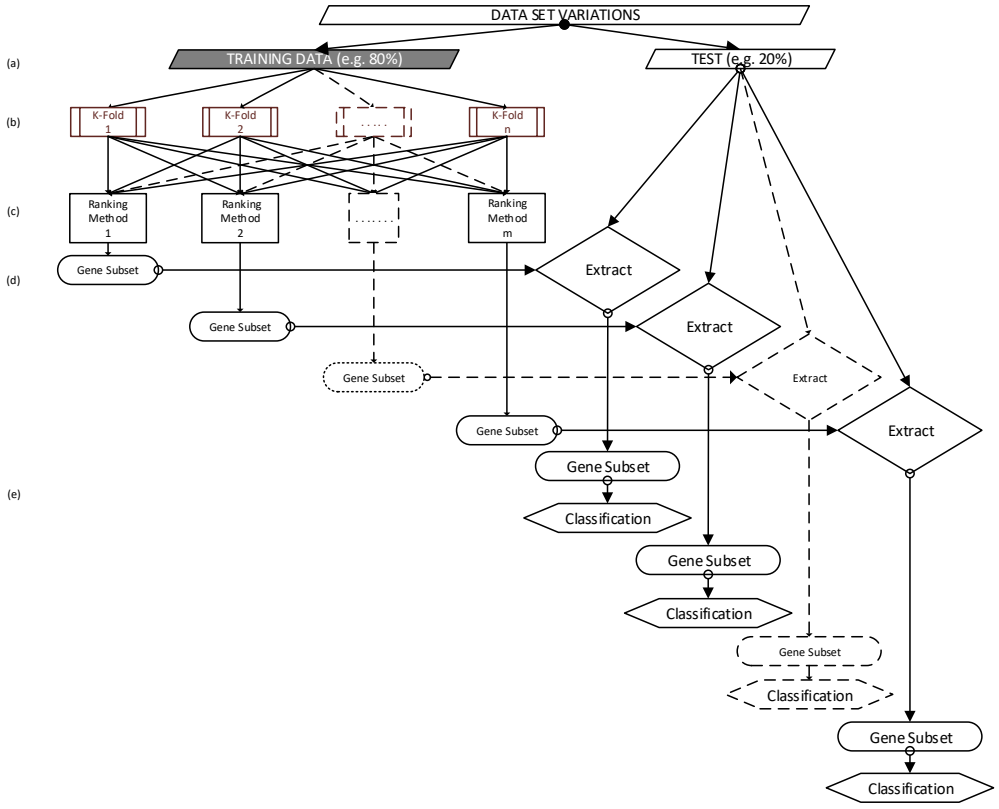

Figure 4-3 Flowchart of the proposed framework without Pareto Optimality on paired-up features of the applied Filter Method (Model-3)

Model 3:

a) The dataset is divided into training and test cases and different dataset (20% test and 80% training or 40% test and 60% training).

b) *K*-fold (5, 10 and LOO) cross-validation is applied to the training data sets (80% and 60%).

c) The filter methods T-test, Entropy, Bhattacharya, ROC, Wilcoxon (Fox and Dimmic, 2006; Nguyen *et al.*, 2015) tests are respectively used alone without PO (Figure 4-3), and subset sizes that yield the highest accuracy from Model-1 are chosen as in the same K-level.

d) For the applied filter method, the ranked features of n-folds paired up.

e) The *k*NN is applied as the classification technique to evaluate our framework.

4.5.2 Cross-Validated Aggregated Gene Selection using Filter-Wrapper Combination

This section describes the proposed novel multi-objective framework for gene selection. The SFS method was utilised to identify the most strongly disease-related genes, as seen on Figure 4-4. Due to the heavy computational burden of wrapper methods, before applying the wrapper technique, the number of features are decreased through the use of multiple filtering methods; namely, the Bhattacharya, Entropy, ROC, t-test, and Wilcoxon tests. This makes feature selection using the wrapper methods faster and more effective. The multi-criteria framework is illustrated in Figure 4-4, and its steps are as follows.

a) The dataset is split into test and training sets (20% test and 80% training or 40% test and 60% training) and numbers of feature dimensions are decreased in the selected feature pool.

b) Cross-Validation*: K* -fold (LOO, 10, 5) CV is applied to the training part of the datasets. *K*-fold preparation*:* Conducted to apply the SFS to the folds. SFS operates with an internal *k*NN classifier to select the relevant genes.

c) SFS selects genes from each fold, and each fold forms a vector of fixed size features for PO. After many experimental trials, we set the size of feature vectors to 10.

d) Common subset: Duplicated genes are removed from the resulting gene set, and a common subset is established using this unique gene content.

e) The PO solution set is ready for classification.

f) For the classification task, the *k*NN classifier (1, 3, 5, 7, 9 NN) respectively is applied to test sets (20% or 40%).

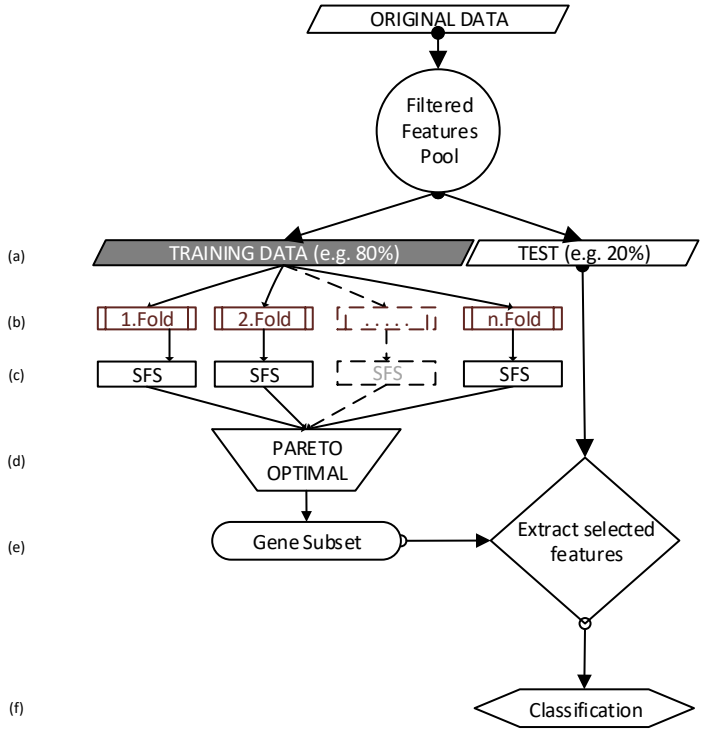

Figure 4-4 Flowchart of the proposed framework using Pareto Optimality on paired-up features of the filter-wrapper combination

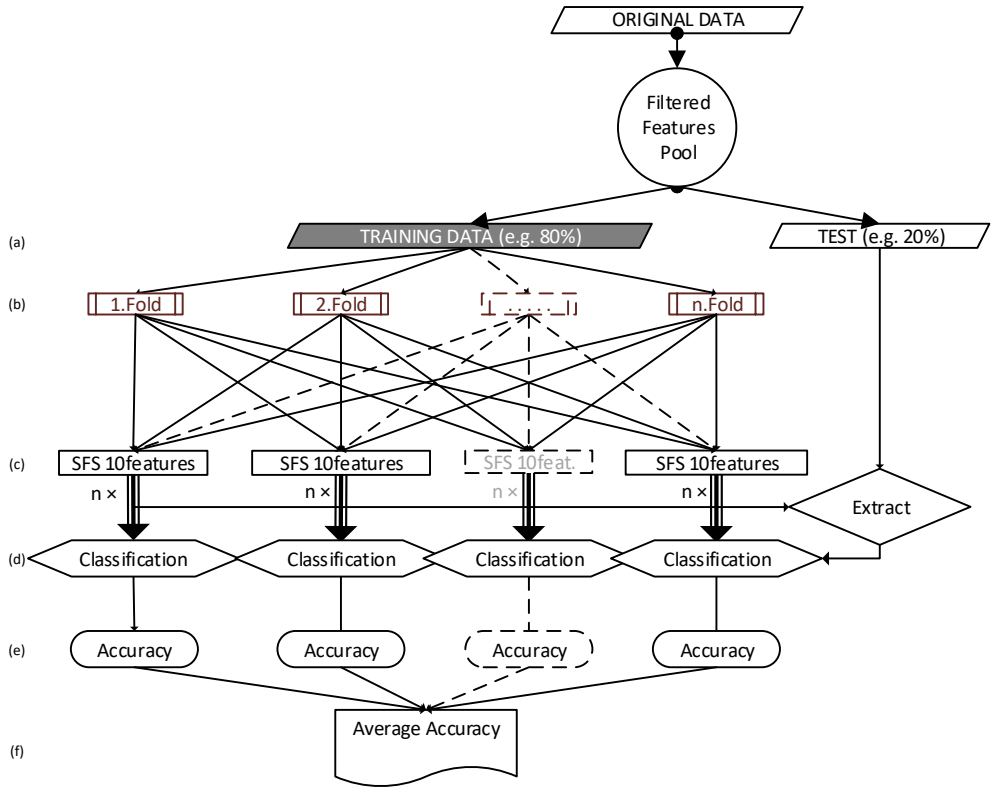

Figure 4-5 Flowchart of the proposed framework without Pareto Optimality on Paired-up features of the applied filter-wrapper combination

a) The dataset is split into test and training sets (80% training and 20% test) and numbers of feature dimensions are decreased in the selected feature pool.

b) Cross-Validation*: K* -fold (LOO, 10, 5) CV is applied to the training part of the datasets. *K*-fold preparation*:* Conducted to apply the SFS to the folds. SFS operates with an internal *k*NN classifier to select the relevant genes.

c) SFS selects genes from each fold. After many experimental trials, we set the size of feature vectors to 10.

d) For the classification task, the *k*NN classifier (1, 3, 5, 7, 9 NN) respectively is applied to test sets 20%. Classification process utilise for data variation because there is no PO method to combine results.

e) Each classification provides an accuracy result.

f) Calculated average accuracy to understand SFS methods performance on this framework.

4.6 Results and Discussion

Table 4.1 and 4.2 summarize the classification results of the three different datasets for the feature selection problem. Each resampling of the data is implemented ten times, and the mean accuracy and standard deviation along with corresponding feature subset sizes were obtained. In addition to the percentages, the different sizes of test datasets are shown in the tables. The methods used were evaluated for multiple *K*s, but due to lack of space and for clarity only the results from *K* that yield the highest performance are presented here.

Table 4.3 presents the results of the comparison of the proposed PO-based framework with those from previous work (Ogutcen *et al.*, 2016). It is observed that the baseline methods require high numbers of features to reach maximum accuracy (Table 4.3) whereas the proposed approach yielded similar or higher accuracy rates using the same number of or fewer genes (Bolón-canedo *et al.*, 2014). Model-1 using all of the tests with PO shows a significant improvement in accuracy with the Colon dataset. This analysis gave an increase of 5% in predictive accuracy with around only 12 genes, which is almost 10 times lower than that of the previous study (Luo *et al.*, 2011). A smaller number of genes is also observed for the other two datasets. The results demonstrate the robustness of the proposed framework, which consistently yields smaller subsets of features.

Consequently, the results appear to suggest that the stability of feature selection in Model-1 is significantly better than that of Models 2 and 3. The stability seems to be similar for all the assessments and data sets. In Model-3, the standard deviation of different *K*s in *k*NN is observed to be more than that of Model-2. Therefore, although they have similar accuracy, this model may not be as reliable as the Model-1.

Table 4.1: Performance of Model-1

Model-1	COLON (K=11) 20%		COLON (K=11) 40%		DLBCL (K=11) 20%		DLBCL (K=11) 40%		DUKE (K=1) 20%		DUKE (K=1) 40%	
	Total acc. %	*Subset Size*	*Total acc. %*	*Subset Size*	*Total acc. %*	*Subset Size*	*Total acc. %*	*Subset Size*	*Total acc. %*	*Subset Size*	*Total acc.%*	*Subset Size*
5 Fold	97.33 ±2.31	12.98 ±4.01	91.42 ±2.85	16.18 ±2.49	99.75 ±0.56	14.40 ±5.01	91.94 ±2.62	21.30 ±5.84	**95.11 ±6.22**	**28.32 ±4.65**	**83.41 ±6.63**	**27.02 ±7.76**
10 Fold	**97.50 ±2.76**	**11.56 ±5.28**	91.58 ±2.59	14.52 ±3.98	100.00 ±0.0	12.82 ±4.64	91.55 ±3.50	17.54 ±6.80	94.67 ±5.41	27.26 ±5.85	80.59 ±8.88	24.70 ±9.95
LOO	96.67 ±4.56	10.20 ±6.02	**91.67 ±2.95**	**13.20 ±4.97**	**100.00 ±0.0**	**11.60 ±4.16**	**92.26 ±3.68**	**15.40 ±7.06**	93.33 ±6.09	24.00 ±7.58	77.65 ±7.67	23.00 ±11.29
Sensitivity	***99.50±1.12***		***96.25±5.59***		***100.00±0.0***		***91.30±3.07***		***91.20±11.19***		***80.67±8.59***	
Specificity	***93.50±8.59***		***82.50±16.77***		***100.00±0.0***		***95.00±11.18***		***100.00±0.0***		***86.50±15.27***	

Table 4.2: Performance of Model-2 (with PO) and Model-3 (without PO)

	COLON (K=11) 20%		COLON (K=11) 40%		DLBCL (K=11) 20%		DLBCL (K=11) 40%		DUKE (K=1) 20%		DUKE (K=1) 40%	
	Total acc. %	*Subset Size*	*Total acc. %*	*Subset Size*	*Total acc. %*	*Subset Size*	*Total acc. %*	*Subset Size*	*Total acc. %*	*Subset Size*	*Total acc. %*	*Subset Size*
PO + t-test	87.92 ±9.92	3.45	81.88 ±7.56	4.75	87.26 ±3.70	1.20	87.26 ±3.70	2.05	80 ±14.69	7.55	73.24 ±9.25	4.00
t-test	85 ±8.58	±2.16	81.04 ±8.17	±2.83	86.88 ±5.90	±0.52	87.26 ±3.88	±0.89	79.44 ±15.43	±4.96	70 ±10.34	±1.78
PO + Entropy	82.92 ±15.17	5.95	81.25 ±6.83	7.50	77.81 ±9.40	4.70	73.55 ±6.91	4.60	76.11 ±14.98	11.20	76.18 ±13.01	9.85
Entropy	81.67 ±15.50	±4.48	82.08 ±6.73	±4.05	76.56 ±9.86	±2.81	73.71 ±5.70	±3.56	78.33 ±11.37	±5.44	74.41 ±14.68	±8.05
PO + Bhattacharyya	83.75 ±9.16	3.00	81.88 ±9.00	3.60	85.94 ±12.96	1.50	86.29 ±5.22	1.90	82.22 ±10.45	10.60	77.65 ±9.07	7.40
Bhattacharyya	82.92 ±8.11	±1.69	83.13 ±9.27	±1.85	84.69 ±11.94	±0.51	86.61 ±5.33	±0.72	82.22 ±9.56	±4.17	76.76 ±10.44	±7.37
PO + ROC	84.58 ±11.24	3.30	82.50 ±7.72	4.90	90 ±8.46	1.55	89.03 ±6.31	3.75	78.89 ±16.08	5.35	73.53 ±6.75	4.65
ROC	84.58 ±10.95	±2.00	81.88 ±7.37	±2.22	90.31 ±7.77	±0.83	88.87 ±6.41	±2.84	78.89 ±12.12	±3.15	74.71 ±9.50	±2.92
PO + Wilcoxon	83.75 ±10.64	3.35	82.29 ±7.63	4.75	85 ±7.96	5.05	79.35 ±10.90	3.95	82.22 ±14.15	5.55	68.82 ±13.24	4.70
Wilcoxon	84.17 ±10.51	±2.08	82.71 ±7.49	±2.31	85.31 ±7.72	±2.06	79.03 ±10.38	±2.46	79.44 ±12.80	±2.39	69.12 ±13.14	±3.26

Table 4.3: Performance comparison of Model-1 using LOO Cross-validation with state-of-the-art techniques using the sample subsets

	DLBCL		Colon		Duke	
	Accuracy %	**Number of Genes**	**Accuracy %**	**Number of Genes**	**Accuracy %**	**Number of Genes**
SNR (Luo *et al.*, 2011)	84.34	10	78.20	10	81.00	20
FFS-ACSA1 (Luo *et al.*, 2011)	77.72	10	80.00	10	71.00	20
FFS-ACSA2 (Luo *et al.*, 2011)	85.09	10	78.36	10	82.83	20
SVM-RFE (Luo *et al.*, 2011)	91.87	10	76.76	10	82.33	20
Model-1	**100.0±0.0**	**11.60 ±4.16**	**97.50 ±2.76**	**11.56 ±5.28**	**83.76 ±5.68**	**24.70 ±9.95**
Sensitivity	100.0±0.0		99.50±1.12		80.67±8.59	
Specificity	100.0±0.0		93.50±8.59		86.50±15.27	

The selection of valid features is important in order to reduce the number of features in the dataset. At the same time, the models are computationally intensive, and so less training data provides a shorter time for feature selection. If training datasets of smaller sample size can provide equivalent or stronger results, this would improve computational efficiency. Therefore in Table 4.3, the performance levels of 20% test and 80% training (Luo *et al.*, 2011) versus 40% test and 60% training (Proposed Method Model-1) sets were compared. In every situation with all the models, the larger training set (80%) provided higher accuracy. Using a 40% test size as in (Luo *et al.*, 2011) for low sample size datasets has a negative effect on overall results. When we look at the *K*-fold difference, in Models 1 and 2, LOOCV-based assessment yielded more stable and higher classification accuracies. In any case, the proposed framework seems durable and robust.

In the second stage of the study, an experimental analysis of the proposed framework was carried out on both SFS+PO (Figure 4-4) and individual SFS (Figure 4-5) models, and the classification results for the selection problem for each dataset are summarised in Table 4.4-4.6.

Folds of the dataset is resampled and repeated for ten times, and the average and standard deviation of the accuracy along with the selected gene sets were obtained. Additionally, the performance of different test data sizes (20% and 40%) are shown in tables 4-6. Classification performance was evaluated using

different values of *K* but, due to limited space and for clarity, only values of *K* that yielded the best results are presented.

As seen in Table 4.4, classification accuracy consistently reached 100% accuracy using SFS-PO, and the corresponding subset size depends on the cross-validation fold size. Furthermore, individual SFS method reached its maximum performance using a few numbers of genes, however its accuracy has remained lower than the accuracy of the proposed PO based ensemble framework. Moreover, SFS was not able to identify disease-related genes as compared to disease related-genes identified with PO based ensemble framework. When fold number and test set sizes were increased, the results included higher numbers of genes due to over-learning from the folds and smaller sample sizes in training. The datasets utilised included different versions such as with different fold sizes and varying K values. In these different conditions, accuracy remained robust and high.

Table 4.4: Classification results of Colon breast cancer with the proposed Filter-Wrapper combined framework

COLON Cancer		**20% Test**		**40% Test**	
		Total acc. %	**Subset Size**	**Total acc. %**	**Subset Size**
5-FOLD	**SFS-PO**	100 ±0.0 (K=7)	15.32±2.04	94.42 ±1.58 (K=7)	10.68±1.52
	SFS	94.57 ±2.32	8	84.5 ±3.49	3
10-FOLD	**SFS-PO**	100 ±0.0 (K=7)	12.96±1.85	94.83 ±0.66 (K=7)	36.58±3.45
	SFS	88.78 ±2.71	2	90.46 ±1.62	13
LOO	**SFS-PO**	100 ±0.0 (K=5)	28.96±5.10	95.42 ±0.97 (K=7)	89.82±7.94
	SFS	91.71 ±2.42	3	91.72 ±1.10	18
Sensitivity		100.00 ±0.0		97.63 ±0.40	
Specificity		100.00 ±0.0		91.00 ±2.87	

Table 4.5: Classification results of DLBCL breast cancer with the proposed Filter-Wrapper combined framework

DLBCL		**20% Test**		**40% Test**	
		Total acc. %	**Subset Size**	**Total acc. %**	**Subset Size**
5-FOLD	**SFS-PO**	99.38 ±1.45 (K=7)	10.94±1.64	92.06 ±4.80 (K=3)	11.40±1.54
	SFS	92.03 ±2.89	3	86.14 ±1.91	3
10-FOLD	**SFS-PO**	**100 ±0** (K=7)	13.44±1.66	95.16 ±4.69 (K=5)	13.16±2.02
	SFS	90.66 ±2.20	2	86.35 ±2.16	2
LOO	**SFS-PO**	100 ±0 (K=9)	18.30±4.48	96.97 ±2.61 (K=5)	45.86±6.55
	SFS	92.29 ±2.47	2	89.88 ±2.15	6
Sensitivity		100.00 ±0.0		96.78±3.06	
Specificity		100.00 ±0.0		97.50±3.86	

For the DLBCL data, a relatively low number of genes is required to reach the highest accuracy using SFS, as seen in

Table 4.5. However, the best result required around 13 genes using the SFS-PO framework. The SFS-PO combination focuses on the overall data rather than just training data, and so it can achieve better accuracy in terms of disease-related genes. PO considers each gene selected from all samples and so each sample has an effect on the features selected. PO evaluates each selected gene, and none are eliminated as in the voting methods. Therefore, PO offers a suitable approach for analysing cancer-related genes in microarray datasets.

The classification performance of Duke Breast cancer data for 20% test and 80% training versus 40% test and 60% training sets are shown in Table 4.6.

There are very limited studies with this dataset for comparison, and it is the most challenging dataset of the three to maximise accuracy. The performance of the PO based framework was experimented with different cross-validation methods. Leave-one out cross validation obtained the highest accuracy for the Duke Breast cancer dataset as compared to 5-Fold and 10-Fold cross validations. However, selected subset of genes for the LOO is obtained higher than 5 and 10-Fold. This result may seem to contradict a common objective in the literature, which is to achieve maximum accuracy with a minimum subset size. However, PO shows its potential by including all essential genes, and this information can support biological studies of cancer.

Table 4.6: Classification results of DUKE breast cancer with the proposed Filter-Wrapper combined framework

DUKE Breast Cancer		**20% Test**		**40% Test**	
		Total acc. %	**Subset Size**	**Total acc. %**	**Subset Size**
5-FOLD	**SFS-PO**	93.11 ±7.79 (K=7)	11.30±1.67	82.35 ±9.72 (K=3)	12.0±1.98
	SFS	82.71 ±5.65	4	81.93 ±2.95	18
10-FOLD	**SFS-PO**	97.33 ±3.85 (K=7)	30.04±4.00	85.76 ±8.35 (K=5)	29.48±5.07
	SFS	87.33 ±2.58	10	81.34 ±2.48	19
LOO	**SFS-PO**	98.44 ±3.88 (K=3)	90.08±10.30	88.47 ±4.65 (K=5)	67.20±9.16
	SFS	91.39 ±1.87	19	82.23 ±1.55	20
Sensitivity		91.20±11.19		80.67±8.59	
Specificity		100.00±0.0		86.50±15.27	

Table 4.7: Performance comparison of ensemble methods applied to Colon Dataset

COLON	**Accuracy**			**Sensitivity**	**Specificity**
	Total acc. %	Std. Dev. %	Subset Size	Acc%±Std.	Acc%±Std.
SFS-PO, 10-Fold, 7NN	100	±0.0	12.96 ±1.85	100.00 ±0.0	100.00 ±0.0
SFS, 10-Fold, 7NN	88.78	±2.71	2	92.70 ±3.10	80.95 ±5.97
Model-1, (Ogutcen *et al.*, 2016)	97.5	±2.76	11.56 ±5.28	99.50±1.12	93.50±8.59
GBC +SVM (Alshamlan, Badr and Alohali, 2015)	98.38	-	10	-	-
MOGA (Hasnat, 2016)	82.3	±7.2	9.03±1.6		
TOPSIS + SVM (Fattah *et al.*, 2013)	88.7	-	10	-	-
Pareto DE (Dash and Misra, 2017)	81	-	-	0.67	0.79

Table 4.7, Table 4.9 show the results of the comparison with state-of-the-art ensemble methods applied to the same datasets. The first observation to note is the lack of information about their performance for comparison. Many of the comparison studies provided results only for accuracy. SFS-PO accuracy is better than the SFS only accuracy. This result reflects how PO integrated framework changes the classification accuracy using same set of methods. Only some of them additionally reported standard deviation without which we cannot get a full understanding of the classification performance.

Table 4.8: Performance comparison of ensemble methods applied to DLBCL Dataset

DLBCL	**Accuracy**			**Sensitivity**	**Specificity**
	Total acc. %	Std. Dev. %	Subset Size	Acc%±Std.	Acc%±Std.
SFS-PO, 10-FOLD, 7NN	100	±0.0	13.44 ±1.60	100.00 ±0.0	100.00±0.0
SFS, LOO, 3NN	90.66	±2.20	2	92.60±2.25	84.85±4.88
Ensemble Filters-PO (Ogutcen *et al.*, 2016)	100	±0.0	11.60 ±4.16	100.00±0.0	100.00±0.0
EPSO (Mohamad *et al.*, 2013)	100	±0.0	4.7±0.82	-	-
FSAM-AHP (Nguyen *et al.*, 2015)	98.70		-	-	-
QMI-SVM (Mortazavi and Moattar, 2016)	96.07		-	-	-

A few of the studies provided accuracy rates for a fixed size of subset. Accuracy with a fixed number of genes shows how efficiently classification can be performed at that point. On the other hand, the proposed aggregated multi-objective framework focuses on discovering feature dependencies in the dataset as a whole.

The comparison of accuracy levels for the DLBCL data shows that EPSO is a competitive method (Table 8). However, EPSO is based on iteration, and so it continues to remove genes until maximum accuracy is reached. If the dataset were artificial, removing features after reaching the maximum accuracy would be the correct action. However, cancer datasets are not created synthetically. They represent the gene expression levels of real humans, and information about one of which may be necessary for the cancer treatment process. Thus, if any gene is non-dominant it should be kept in the subset for further study by molecular biologists. Relationships among genes are vitally important.

Similarly, learning algorithms should not only focus on the minimum possible number of features. All disease-related genes should be considered when investigating treatment. If a gene, which is removed, has a relationship to the disease, this situation postpones the discovery of that gene. Thus, the

methodology applied should aim to discover all disease-related genes and all potentially relevant genes should be considered. PO achieved better results than previous studies and yields different subset sizes for each cancer dataset.

Table 4.9: Comparison of ensemble methods applied to Duke Dataset

DUKE	**Accuracy**			**Sensitivity**	**Specificity**
	Total acc. %	Std. Dev. %	Subset Size	Acc%±Std.	Acc%±Std.
SFS-PO, LOO, 3NN	98.44	±3.88	90.08±10.30	91.20±11.19	100.00±0.0
SFS, LOO, 3NN	91.39	±1.87	19	89.04±3.01	94.33±2.38
Ensemble Filters-PO(Ogutcen *et al.*, 2016)	83.76	±5.68	24.70 ±9.95	80.67±8.59	86.50±15.27
RandEns.ReliefF (Zhou *et al.*, 2014)	91.5	-	40	-	-
LFS (Armanfard, Reilly and Komeili, 2016)	89.17	±7.90	-	-	-
FFS-ACSA2 (Luo *et al.*, 2011)	82.83	-	20	-	-

The Duke Dataset has the smallest sample number of the three datasets considered and this makes it much harder to gain better accuracy than in the studies compared in Table 7. Subset sizes start from 11 genes when 5-fold CV is applied and reach 90 genes when LOOCV is used. In terms of accuracy, differences in fold size increase accuracy levels from 93.11% to 98.44%. This 5% increase in accuracy requires a set with an additional 81 genes. 90 genes are significantly more than the other two cancer dataset subset sizes. Breast cancer may be associated with many genes and LOO may be more suitable when a dataset has a small number of samples. As mentioned above, the proposed framework focuses on all disease-related genes in order to create meaningful subsets for biologists rather than just a minimum number of genes.

Each sample influences the results. Each fold works with one of the training set variations, and each training set variation contains a different group of samples. Dependencies among different training set samples may not be captured when

using voting methods. Unlike previous aggregation methods, feature dependencies between different folds are taken into account here in the final subset. This means that the results from different training set variations (which contain different sample sets) are considered by PO. Combining the results for the different cross-validation folds creates a common subset, and this provides a more successful classification than those in other studies (see Tables 7-9).

Like most machine-learning algorithms, SFS aims to achieve a subset size, which provides maximum performance. Unlike standard machine-learning algorithms, PO is a multi-objective decision method, and so it evaluates all results together. Then, it gives all optimal solutions. A multi-objective technique is more suitable for the analysis of gene-expression datasets because of its ability to provide comprehensive information for biological studies. Cancer is a polygenic disease, and so if a potentially important gene is eliminated in the feature selection stage, the discovery of cancer-related genes will remain incomplete. Thus, the final subset should be optimal in the sense of allowing the discovery of possible disease-associated genes.

4.7 Conclusion

The novel framework proposed in this chapter provides promising results in terms of accuracy and stability over the three different benchmark microarray gene-expression datasets under different conditions. It was also observed that the ensemble of methods boosted the feature selection process, and the PO-based approach yielded comparatively smaller feature subset sizes.

Furthermore, in this study, a PO-based multi-objective predictive model is used to identify cancer-related genes in microarray data. A different set of genes was assessed for each cancer data set through a combination of SFS with PO. The proposed method outperforms SFS, and additionally the multi-objective framework is more effective in identifying all disease-related genes. This makes the final subset more representative of the whole dataset.

In the literature, ensemble methods are mostly used for imbalanced datasets, and research has intensively focused on splitting methods to balance datasets. However, these studies either add artificial samples (oversampling) or remove some samples (undersampling) to achieve balance in the dataset. Next chapter

addresses how the proposed framework could handle the problems of data imbalance in gene expression datasets.

5 A PARETO-OPTIMAL GENE SELECTION FRAMEWORK FOR IMBALANCED DATA

5.1 Introduction

Microarray gene-expression data analysis presents a critical role for decision-making in clinical and medical studies. However, natural datasets are mostly imbalance that creates a bias towards majority class when feature selection and classifying samples. This situation generally causes problems during the learning of the classifier. Among many data-level approaches commonly used to tackle this problem are undersampling, over-sampling and ensemble learning. Undersampling or oversampling may result in a loss of valuable data and unreliable results due to the artificially created or altered samples.

Various datasets present more cancer cases of one class than the control samples of the other class in the various datasets (Shipp *et al.*, 2002). Therefore, applying feature selection on imbalanced datasets requires further investigation since the features selected are likely to be more representative of the majority class (Fernandez, Garcia and Herrera, 2011). An increasing number of studies are being conducted on imbalanced datasets in various domains and also in bioinformatics (Haixiang *et al.*, 2017).

Gene-expression datasets have low number of samples and high numbers of features. Traditional classification methods are not designed for imbalanced situations and so classifier predictions are biased towards majority class (Ling and Sheng, 2010). Various solutions have been proposed in machine learning domain in recent years for overcoming poor prediction performance of classifiers on imbalanced datasets. These solutions are mainly categorised in data-based and algorithm-based approaches.

A novel approach has been proposed to imbalanced microarray datasets for feature selection in this chapter. Since the class ratio of samples is not uniformly distributed, a framework is designed to handle imbalanced datasets. For this purpose, sub datasets that keep a balance between minority and majority classes were generated then combined with the imbalanced dataset. Hence, multi-objective structure is generated with a set of balanced-sub datasets along with the original imbalanced dataset. Thus, the proposed framework prevented the

side effects of methods such as undersampling or oversampling and benefited from boosting the feature selection performance using supportive balanced sub-datasets.

5.1.1 Performance Metrics for Imbalanced Microarray Gene Expression Data

The evaluation metrics (e.g., accuracy) that work well with standard balanced datasets might not be applicable when analysing imbalanced microarray gene-expression data. This is due to the fact that accuracy achieved will be biased towards the majority class. Therefore, performance metrics other than accuracy should be considered in the evaluation of the classification performance of the proposed framework. One approach that could be used for this purpose is to find values of recall and precision for minority and majority classes separately using a confusion matrix Table 5.1 (Anaissi and Kennedy, 2011).

Table 5.1 Confusion Matrix

		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative	False Positive
	Positive	False Negative	True Positive

The outcomes of the confusion matrix can be used to calculate TPR and FPR. Formulas are as follows (Fawcett, 2006);

$$Precision = True\ Positve\ Rate\ (TPR) = \frac{TP}{TP+FN} \qquad \textit{Equation 5-1}$$

And

$$Recall = True\ Negative\ Rate\ (TNR) = \frac{TN}{TN+FP} \qquad \textit{Equation 5-2}$$

The F-measure (or F1 score) could also be used, which is a combination of recall and precision represented in the function below:

$$F1 = 2x\frac{Precision*Recall}{Precision+Recall} \qquad \textit{Equation 5-3}$$

The F-measure is a better measure for imbalanced datasets than accuracy as it is more capable of representing the impact of precision and recall in combination (Yin *et al.*, 2013).

5.1.2 Gene selection in imbalanced microarray datasets

This section explains the challenges for gene selection caused by imbalanced data which is useful to increase the classification results in terms of f-measure and efficiency (Yin *et al.*, 2013; Maldonado, Weber and Famili, 2014). Selected genes are supposed to improve the classification model's ability to differentiate between classes. However, classifiers and feature selection methods are adversely affect the performs poorly when applied on imbalanced datasets.

In many medical domains, dataset classes are not evenly distributed, and cancer patients represent a very small minority class, especially when the disease is very rare. However, most of the time, the important information to identify relevant genes that belongs to the minority class.

There are two main strategies that could be used to approach the problem of imbalanced datasets: data-based strategy and algorithm-based strategy. For data-based strategy resamples the dataset in order to uniformly distribute classes (Chawla *et al.*, 2003). There are two main methods used to resample the dataset for this purpose: undersampling and oversampling. Although some research studies suggest employing both of them (Anaissi and Kennedy, 2011). Undersampling is the removal of samples of the majority classes to match their size with that of the minority class. For example, SHRINC is a under sampling algorithm proposed by (Kubat and Matwin, 1997), to be used in reducing the number of majority class samples. This technique has the disadvantage of removing potentially useful information. Moreover, gene-expression datasets often have small number of samples, and in this case, undersampling is not suitable. Conversely, increasing the number of minority class samples is the aim of oversampling techniques. Minority samples are replicated so that, their number become equivalent to the majority class.

Second strategy that can be used for imbalanced datasets is the algorithm-based approach. One main algorithm proposed for handling class imbalance is the cost sensitive learning (López *et al.*, 2013). This algorithm assigns high costs to

misclassified samples of minority class to achieve the objective. Another well-known strategy is the boosting approach. SMOTEBoost algorithm (Chawla *et al.*, 2003), for instance, iteratively learn patterns from minority class samples during the training process of the classifier.

5.2 Proposed approach for handling data imbalance

A proposed multi-criteria approach using PO for feature selection that handles the microarray gene expression imbalanced datasets is presented in this section. In this method, the majority class is split in subsets equivalent to the size of the minority class. The proposed approach handled the original dataset and its balanced-sub datasets together. Multi-objective structure is used to match up balanced sub-datasets with the original dataset.

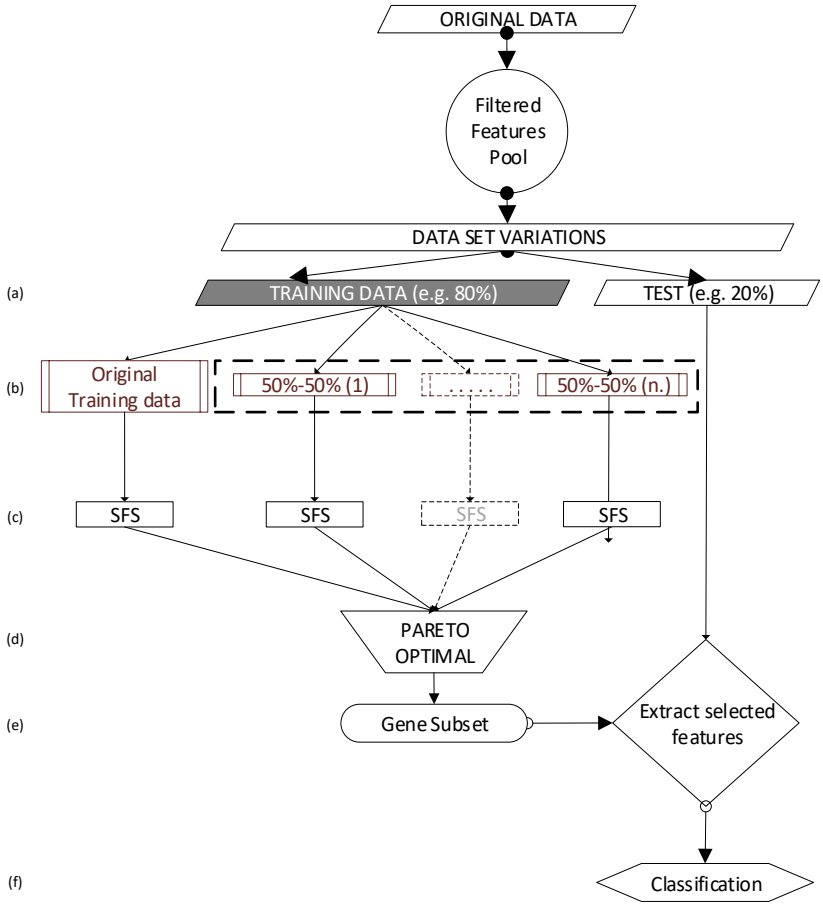

Figure 5-1 Flowchart of the proposed framework using Pareto Optimality on paired-up features of the filter-wrapper combination to be used for imbalanced datasets.

a) The dataset is split into test and training sets (80% training and 20% test) and numbers of feature dimensions are decreased in the selected feature pool.
b) Folds are arranged in the training part to keep the balance of the classes (50%-50%) of the datasets. SFS applied to these balanced folds. SFS operates with an internal *k*NN classifier to select the relevant genes.
c) SFS selects genes from each fold, and each fold forms a vector of fixed size features for PO. After many experimental trials, we set the size of feature vectors to 10.
d) Common subset: Duplicated genes are removed from the resulting gene set, and a common subset is established using this unique gene content.
e) The PO solution set is ready for classification.
f) For the classification task, the *k*NN classifier (1, 3, 5, 7, 9 NN) respectively is applied to test sets 20%.

As shown in Figure 5-1 proposed framework, dataset with its original distribution of classes is used and supported with balance-distributed sub-datasets. Similar to the widely accepted 5-fold and 10-fold approach in the literature, the number of balanced datasets is chosen as ratio between each class data samples. During the training process, without losing any features derived from the original dataset, supportive features coming from the balanced sub-datasets made an impact on the classification performance particularly for the minority class samples.

Skewed class distribution has been determined a training problem in the literature. K-fold approach usually splits the dataset without considering whether it is balanced or imbalanced. In our work, the proposed sub-datasets are a type of replacement for the common K-fold approach specifically for the imbalanced datasets. Research studies on the classification of gene expression data widely assume the classes in the dataset are balanced and suggest ways for improving the classifier performance. However, as far as we know, there are not many studies in the bioinformatics research domain, seeking an alternative way of handling the problem of gene selection in imbalanced microarray gene-expression datasets. Therefore, we proposed a PO-based approach that supports feature discovery to identify minority classes in a better way. According to

aspect ratio between classes, sub-datasets utilised with PO provides more comprehensive feature subsets. The proposed imbalanced dataset framework structure is as follows:

- The dataset is split into test and training sets (80% training and 20% test) and feature dimensions decreased in the selected feature pool.
- Training data first employed with filter methods to gain common genes and created a gene pool for a wrapper method. Duplicated genes are removed from the resulting gene set and a common subset is established using this unique gene content.
- Then narrowed down dataset with the common gene pool employed with a wrapper method. At the same time, random subsets of major class samples matches with minor samples.
- SFS is applied to the original training and balanced folds. SFS operates with an internal *k*NN classifier to select the relevant genes.
- SFS selects genes from each fold, and each fold forms a vector of 10 features for PO
- The PO solution set is ready for classification.
- For the classification task, the *k*NN classifier (1, 3, 5, 7 NN) is applied to the test sets.

5.3 Results and Discussion

In this section, the results of the proposed framework applied on imbalanced datasets were presented. Mainly, balanced datasets are important for minority samples classification, so they help to discover important features of the dataset more effectively.

Table 5.2 Characteristics of Imbalanced Gene Expression Datasets

	Size	Features	Target	#Majority/#minority	Ratio
CNS	60	7129	Failures	39/21	1.86
LYMPH	77	6817	FL	58/19	3.63
OVARIAN	253	15154	Normal	162/91	1.78

Characteristics of the applied imbalanced gene expression datasets are shown on Table 5.2. Three different cancer datasets in different size of samples and

features investigated. We studied the CNS, Lymphoma and Ovarian datasets. The Lymph dataset has the most imbalanced characteristics as compared the remaining datasets having 58 majority classes and 19 minority classes. The steps of the proposed multi-criteria approach using Pareto Optimality for feature selection are shown in Figure 5-1.

Table 5.3 Mean and standard deviations generated for the subset-balanced dataset sizes and k-level of the kNN classifier at the Maximum Accuracy

Dataset	**Mean Subset Size**	**Standard Deviation of Subset Size**	***k* Level**
OVARIAN	7.59	2.45	9
CNS	14.67	2.41	5
LYMPH	9.31	1.91	9

The mean and standard deviations of the subset-balanced dataset sizes that are generated and the k-level of the *k*NN classifier obtained at the best accuracy performance of our proposed framework is presented in Table 5.2.

Table 5.4 Performance comparisons for the imbalanced CNS gene expression dataset

	F-measure for rare class	**F-measure for major class**
MI	0.66	0.87
D-MI	0.70	0.85
Fisher	0.54	0.85
D-Fisher	0.60	0.87
Corr	0.69	**0.89**
D-Corr	0.77	0.88
SFSPO+ 9NN	***0.80***	*0.85*

In the Table 5.3 we reported the F-measure results of various methods presented in this research domain (Yin *et al.*, 2013) including the proposed framework which are applied on imbalanced CNS gene expression dataset. Our proposed framework provided highest F-measure (0.80) for the minority class with a good F-measure (0.85) for the majority class.

Table 5.5 Performance comparisons for the imbalanced Lymph gene expression dataset

	F-measure for rare class	F-measure for major class
MI	0.65	0.86
D-MI	0.70	**0.90**
Fisher	0.54	0.85
D-Fisher	0.59	0.87
Corr	0.17	0.85
D-Corr	0.32	0.87
SFSPO+ 5NN	***0.74***	*0.89*

In the Table 5.4, we reported the F-measure results of seven methods including the proposed framework, which are applied, on imbalanced Lymphoma gene expression dataset. Our proposed framework provided highest F-measure (0.74) for the minority class along with the second-best F-measure (0.89) for the majority class.

Table 5.6 Performance comparisons for the imbalanced Ovarian gene-expression dataset

	F-measure for rare class	F-measure for major class
MI	0.66	0.87
D-MI	0.70	0.85
Fisher	0.54	0.85
D-Fisher	0.60	0.87
Corr	0.59	0.91
D-Corr	0.61	**0.92**
SFSPO+9NN	***0.79***	*0.91*

In the Table 5.5, we reported the F-measure results of seven methods including the proposed framework, which are applied, on imbalanced Ovarian gene-expression dataset. Our proposed framework provided highest F-measure (0.74) for the minority class along with a second-best F-measure (0.91) for the majority class.

We obtained significantly better minority class performances for the CNS, Lymphoma, and Ovarian gene expression datasets. Even, in the majority class, there is no significant difference on the performance measure. Ovarian dataset

has the highest number of samples, which could mitigate the effects of imbalanced distribution of classes. As a consequence the classifier performance achieved a high F-measure performance rate for both majority and minority class samples. In the Lymphoma dataset, interestingly we observed low F-measures for the Corr and D-Corr methods however, their F-measures were better on the CNS dataset.

Additionally, it is always preferable choosing a feature selection approach that proves its superiority among different dataset characteristics. In all datasets on which the proposed framework was applied, the remaining the methodologies were outperformed on the important minority class, achieving the best or second-best performances on the majority class. The reason for not achieving the best majority classification performance might be due to the fact that as the important genes were being added to the subset by the ensemble framework, a bias towards the majority class samples is inevitably occurred.

One more interesting point we observed that performance is highly affected with different *k*-levels of the *k*NN classifier. The expression values of neighbour samples have impacts on the classifier performance. In multi-objective approach, different resampling of the dataset has resulted different feature means. Therefore, structure of the framework improved the minority class identification while minimising the bias effect of the majority class.

As a result, we observed that the sample dependency is highly related with the performance of computational models when studying with the imbalanced gene expression datasets. Thus, the proposed PO based multi-objective approach has shown its ability in minimising such issues for the imbalance datasets. Moreover, since our objective is to better classify the minority class, our approach accomplished its objective by effectively revealing the genes having the most discriminative ability between the healthy and disease samples.

5.4 Conclusion

In this study, we have proposed a framework that selects the relevant and important genes particularly in imbalanced microarray gene-expression datasets. To the best of our knowledge, this is the first time Pareto Optimality is employed for gene selection in imbalanced microarray datasets. Apart from

the classification performance, it was shown that the proposed approach is capable of generating a narrow subset of genes using ensemble of feature selection methods.

6 A PARETO-OPTIMAL GENE SELECTION FRAMEWORK WITH MISSING VALUE IMPUTATION

6.1 Introduction

Many types of real-world experimental data often include missing values (MVs) (Tibshirani et al. 2001; Celton et al. 2010; Oh et al. 2011; Chiu et al. 2013) and microarray datasets are no exception where missing values can be found at different levels of gene expression profiles (Dyrskjot et al., 2003). There are several reasons why these experiments produce such MVs. For instance, the fluorescence intensity of spots may be corrupted by background signal intensity, fabrication faults may take place, or the microarray image might be corrupted, and its resolution may not be sufficient. A significant problem in microarray gene-expression datasets is that they often contain missing values, which can have adverse effects on the decision-making process (Souto et al., 2015, Wang et al., 2018).

Gene expression data is now used in medical treatment, and particularly cancer therapeutics, more than ever before. However, this data often contains missing values, which should be appropriately handled before the analysis. One way to handle missing values is to use pre-processing methods to impute those missing values.

Numerous studies have been conducted on datasets with missing values and how to apply imputation to them. In recent years, varieties of new imputations approaches have been proposed. Formerly, statistical metrics were effectively used in algorithms to impute missing values.

Microarray datasets cannot be treated solely on the numerical values of gene expressions. They should also be considered with their biological backgrounds, such as gene interactions and functional importance for human life. Thus, some imputation methods cannot work properly with microarray data (Yang, Xu and Song, 2015). When samples containing MVs are eliminated from the gene expression dataset, the results of the classification can be biased. Discarding

missing values can only work for gene expression profiles that contain small numbers of MVs (< 5%)(Souto, Jaskowiak and Costa, 2015).

The most basic imputation applications substitute missing values with a fixed number (such as zero), or finding a value from a sample very close to the sample containing MVs or simply replacing missing values with the mean value of the corresponding feature (Alizadeh *et al.*, 2000). However, these simplistic approaches may lead to poor accuracy performance as they do not consider data correlation from the classification perspective (Troyanskaya *et al.*, 2001). More advanced methods utilise feature correlations to mitigate the problem of missing values. These techniques include, for instance, weighted k-nearest neighbours (WKNN) (Troyanskaya *et al.*, 2001), local least squares (LLS) (Kim, Golub and Park, 2005), and Bayesian principal component analysis (BPCA) (Oba *et al.*, 2003). There is no superior imputation method that overcomes the problem of missing values completely.

In recent studies, further improvements in various imputation methods have been achieved (Oba *et al.*, 2003; Wang *et al.*, 2006; Enders, 2010; García-Laencina, Sancho-Gómez and Figueiras-Vidal, 2010; Celton *et al.*, 2012; Chai *et al.*, 2014). The process validation usually involves a comparison of imputed and observed values with various performance indices (Liew, Law and Yan, 2011). In the work of (Wang *et al.*, 2006), three imputation algorithms, WKNN, LLS and BPCA, were used to impute the missing values of five different cancer gene-expression datasets.

A variety of feature selection methods are used with microarray gene expression datasets to increase classifier performance and prevent overfitting (Yu and Liu, 2004; Nguyen *et al.*, 2015; Li, Li and Yin, 2016; Mortazavi and Moattar, 2016; Ogutcen *et al.*, 2016; Li, Li and Liu, 2017; Ogutcen, Belatreche and Seker, 2018). Each method has different advantages and disadvantages, so an ensemble of multiple methods is discovered via Pareto Optimality. Therefore, each method's benefits are combined to attain a better feature subset.

Different imputation methods, such as using means and medians, create different values for the same features. However, there is no approach to combine and present their estimates under one solution. Imputed features are valuable as

they contribute to the classification, but they must be subjected to the feature selection process. The feature selection has been considered a standard process in this field, so imputed datasets were employed, and their effectiveness was evaluated.

Existing studies are only applicable to gene expression profiles that are complete and do not consider those datasets containing missing values (Souto, Jaskowiak and Costa, 2015). In addition, no approach so far considers Pareto optimality in this context to address the modelling of microarray data with missing values.

In this study, the proposed PO-based framework is used to evaluate the effect of imputations methods on the performance of ensemble feature selection methods and different classifiers. Applied framework provided promising results on various gene expression datasets having one or more missing gene values exist on their samples.

6.2 Imputation Methods

One of the best explanations of the precise mathematical background for imputing missing values can be found in Little and Rubin (Little and Rubin, 2019). However, that work does not consider the use of machine learning techniques to impute the missing data. Moreover, the literature related to missing data imputation for microarray datasets are very limited (Troyanskaya et al., 2001).

Distribution of missing values and amount of samples are effectively influences evaluating an imputation method, investigating the connection between performance of imputation methods and applicability on different datasets having missing values (Moorthy, Mohamad and Deris, 2014). While a variety of imputation techniques exist, most are local learning-based techniques which tend to suffer from overfitting (Wang et al., 2018). In this study, this relationship is investigated using the proposed Pareto optimal based multi-objective framework. We have implemented both statistical imputation and machine learning-based imputation methods to estimate missing values in gene expression datasets.

6.2.1 Statistical Imputation for Gene Expression

Mean imputation: Mean imputation for gene expression replaces a single missing value of gene expression with the average of remaining gene expressions in microarray data. However, one drawback of using this method is that the variability (e.g., standard deviations and variance) of gene expression decreases which can reduce the discrimination ability of that gene (Eekhout *et al.*, 2012).

Median imputation: Median imputation for gene expression replaces a single missing value of gene expression with the median of remaining gene expressions in microarray data. This strategy is mathematically similar to estimating missing values with mean imputation. The median value of a gene of interest is simply obtained and replaced with the missing gene expressions found in samples.

6.2.2 Machine Learning-based Imputation for gene expression

Unlike basic statistical substitution methods such as mean and median based imputation methods, machine learning-based imputation methods are more complex and consider how the data is correlated (Souto, Jaskowiak and Costa, 2015). There are variety of machine learning-based imputation methods available, however we have used imputation methods that are based on k-nearest neighbour, local least squares and Bayesian principal component analysis that are preferred to be studied in (Souto, Jaskowiak and Costa, 2015). They are particularly proposed to handle the missing values in gene-expression datasets and the computational complexity of them is provided in Table 6.1. In this table, n indicates the number of samples, p is the number of selected genes, k is the number of neighbours, and c is the number of components.

These machine learning-based imputation approaches have been applied in numerous research studies to estimate missing values of bioinformatics related datasets (Souto, Jaskowiak and Costa, 2015; Choi *et al.*, 2018; Wei *et al.*, 2018; Magzoub *et al.*, 2019).

***k*NN-based imputation:** *k*NN-based imputation is a machine learning-based imputation approach (Troyanskaya *et al.*, 2001). This method finds *k* neighbour gene expressions in samples that are closest to the gene that is of interest

containing missing gene expression. Each neighbour gene is weighted based on their similarity to the gene having the missing value. Then, *k*NNimpute algorithm replaces the missing value with weighted mean of the neighbour gene expressions.

Local least squares-based imputation: Local least squares (LLS) is another machine learning-based imputation method that uses linear regression to estimate the missing gene expression (Kim, Golub and Park, 2005). Like the *k*NN approach, the LLSimpute algorithm also benefits from the nearby gene expressions. LLSimpute is a simple yet powerful approach and can be a better alternative to more used *k*NNimpute (Liew, Law and Yan, 2011).

The Bayesian principal component analysis-based imputation: The Bayesian principal component analysis (BPCA) is a machine learning-based imputation method that benefits from the Bayesian methodology and a number of principal components for inferring the gene expression missing value (Oba *et al.*, 2003). Even though BPCAimpute method is reported to be better at estimating missing values, its structure is relatively complex (Chai *et al.*, 2014). Therefore, it is less used as compared to the well-known kNNimpute method.

Table 6.1 The computational complexity of machine learning-based imputation methods used in the proposed framework

Machine learning-based imputation method	**Computational Complexity**
kNN-based imputation	O(npk)
LLS-based imputation	O(npk)
BPCA-based imputation	O((p + n)c + 2nc)

6.3 Proposed Framework

This section describes the proposed framework for the imputation of missing gene expression data. The Pareto Optimal (PO) based framework is applied to evaluate the effect of several imputation methods on the framework performance. Five different imputation techniques are applied to five different datasets. The imputation procedure followed is as in (Souto, Jaskowiak and Costa, 2015). PO selects different subsets for each dataset. Owing to feature selection, classification results of the proposed framework show improved accuracy throughout the majority of the datasets.

The proposed framework structure is shown in Figure 6-1;

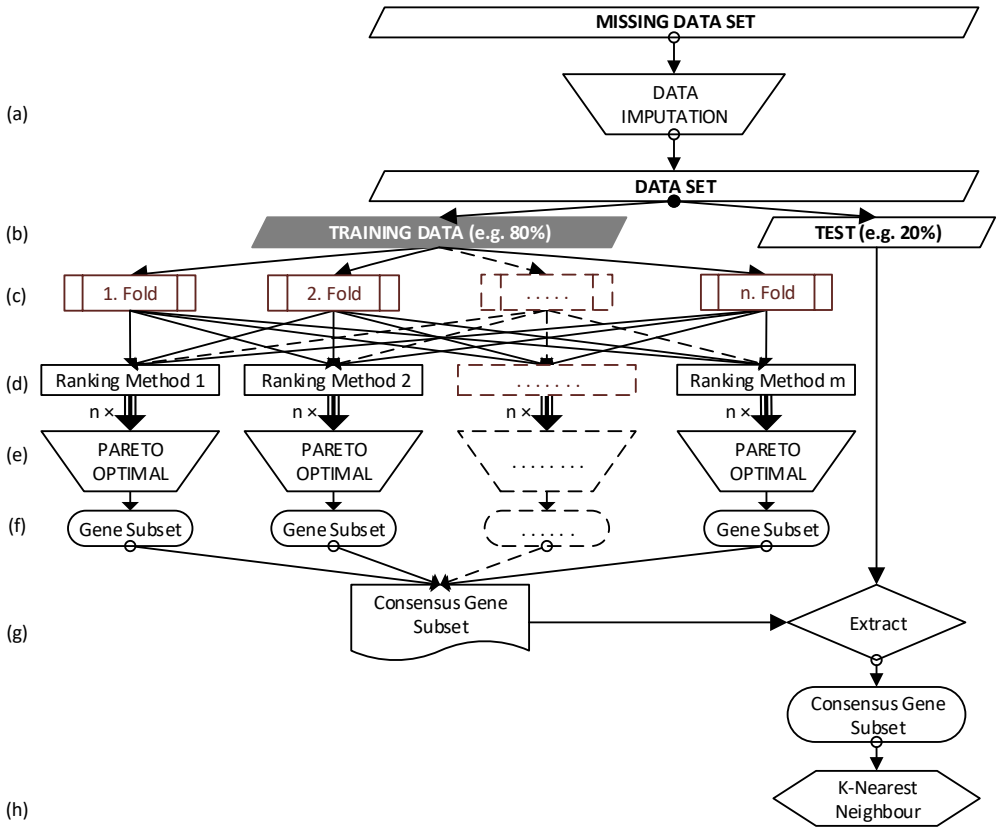

Figure 6-1 Flowchart of the proposed framework using Pareto Optimality on paired-up features of Filter Methods (Model-1) applied for the datasets having missing values

a) Missing values reside in the cancer datasets are imputed using both statistical and machine learning-based imputation methods.

b) The dataset is divided into training and test cases and different dataset (20% test and 80% training or 40% test and 60% training).

c) *K*-fold (5, 10 and LOO) cross-validation is applied to the training data sets (80% and 60%).

d) The filter methods t-test, entropy, Bhattacharya, ROC, and Wilcoxon are used to rank the features. Each filter method is applied to each fold and ranked features are obtained. For the applied filter method, the ranked features of n-folds paired up.

e) The paired-up features from each filter test are further combined using PO (Figure 4-1). This will result in one subset of features.

f) The final gene set is formed at the end of PO process. The duplicated genes that reside in the final gene set will then be removed. This will yield unique subset of a consensus gene subset.

g) The genes that derived from the training data (consensus gene subset) using the PO analysis is validated using the test data. The test datasets (20% and 40%) are formed in two separate cases and organised using the consensus gene subset to validate our framework (aggregated gene selection using filter methods).

h) The *k*NN is applied as the classification technique to evaluate our framework.

In the comparison conducted by Souto et al. (2015), 5 different methods for data imputation were used: K-nearest neighbours (*k*NN), mean, median, Bayesian principal component analysis (BPCA), and local least squares (LSS). The applicability of Pareto Optimal is explored for datasets, which have missing data in terms of how it will improve classification performance. Different imputation methods create different values for genes. However, previous studies did not combine imputed features and neither did they apply feature selection. In the present study, Pareto Optimality is applied in a new way, in order to benefit from different set of feature selection methods.

Previous studies (Souto, Jaskowiak and Costa, 2015) compared the distance calculations such as Pearson correlation, Euclidean distance, variance minimisation, so the most sufficiently accurate calculation is Euclidian distance for missing values. Therefore, the commonly used Euclidean distance metric was used to determine the neighbours. Nevertheless, the findings suggest that Euclidean distance is sensitive to outliers that can exist in microarray data (Troyanskaya *et al.*, 2001).

In this study, the imputation of varying percentage of missing values (between 3% and 8%) of the data was investigated. One can consider the ability of the method to preserve the significant genes in the dataset, or its discriminatory / predictive power for classification or clustering purposes. The performance

levels of data imputation methods with a multi-criteria decision*making framework is compared to those found in previous studies. A Different level of missing data is used for comparison purposes.

6.4 Results and Discussion

In this section, the proposed framework is applied after the missing values have been imputed. Both statistical and machine learning-based imputation methods have been used and their impact on the performance were reported.

The characteristics of five cancer datasets with varying percentage of missing values (between 3% and 8%) used in the proposed experiments is presented in Table 6.2. A description of the five cancer datasets, with 10% missing values, used in the proposed experiments is presented in Table 6.1. It shows the class distributions, class numbers and how many samples in classes. The proposed framework has used various ranking methods including *t*-test, Entropy, Bhattacharya, ROC, and Wilcoxon test and combined with PO. Also, the ratio of missing values that impacts the number of genes (between 40 and 75) in each dataset are shown in Table 6.2. Default values were applied with classifier settings of the linear kernel for the SVMs, *k*NN with *k*=1 and Euclidean distance based on comparative study.

Table 6.2 Cancer datasets with missing values

Dataset	Tissue	No. Classes	Size of classes	No. Samples	Original Data		
					No. genes	%MV	%Genes with MV
DLBCL	Blood	2	21,21	42	4022	3.25	49.3
Brain	Brain	3	31,14,5	50	41472	7.57	43.06
Lung	Lung	4	17,40,4,5	66	24192	3.87	67.81
Prostate	Prostate	4	11,39,19,41	110	42640	4.93	67.16
Endometrium	Endometrium	4	13,3,19,7	42	24192	7.97	74.33

The proposed PO-based framework is used to assess the contribution of each imputation method to finding a good subset of selected genes. The performance of the data imputation methods with a multi-criteria decision framework with respect to (Souto's, 2015) results is given in Table 6.3.

Table 6.3 Accuracy performance comparison of various imputation methods on gene expression datasets using PO-based feature selection with kNN classifier.

		BPCA	BPCA-PO	*k*NN	*k*NN-PO	LLS	LLS-PO	MEAN	MEAN-PO	MEDIAN	MEDIAN-PO
DLBCL-Blood	*Accuracy*	69.05%	92.80%	66.67%	**93.20%**	69.05%	91.20%	66.67%	92%	66.67%	92.40%
	Subset Size	960	32	945	29	962	32	932	**23**	932	25
Brain	*Accuracy*	80%	82.54%	80.00%	**84.55%**	80.00%	82.73%	80.00%	83.82%	80.00%	**84.55%**
	Subset Size	3819	**65**	3833	66	3825	**65**	3850	**65**	3852	66
Lung	*Accuracy*	83.33%	87.86%	83.33%	88.00%	81.82%	86.57%	83.33%	**88.29%**	83.33%	88.14%
	Subset Size	2563	33	2540	32	2578	36	2584	**30**	2603	32
Prostate	*Accuracy*	66.67%	**77.39%**	66.67%	75.12%	65.22%	75.56%	66.67%	77.22%	63.77%	77.04%
	Subset Size	3846	34	3811	35	3833	**32**	3838	37	3930	39
Endometrium	*Accuracy*	80.95%	82.87%	76.19%	**85.18%**	76.19%	85.17%	76.19%	84.72%	76.19%	84.26%
	Subset Size	942	**51**	2074	69	2078	65	2073	63	2073	63

The obtained results show that the proposed framework provided higher accuracy for the microarray imputed gene expression datasets. Real-world datasets often contain missing values and simple removal of samples involving missing values can cause overfitting of the trained model to the dataset or bias toward the classes of samples that do not contain any missing value. However, the obtained results clearly show how important finding unique subsets representing genes with imputed missing values on gene expression datasets reflecting a realistic real-world solution.

Table 6.4 Accuracy performance comparison of various imputation methods on gene expression datasets using PO-based feature selection with SVM classifier.

		BPCA	BPCA-PO	*k*NN	*k*NN-PO	LLS	LLS-PO	MEAN	MEAN-PO	MEDIAN	MEDIAN-PO
DLBCL-Blood	*Accuracy*	91.48%	**93.60%**	91.48%	**93.60%**	88.10%	92.00%	91.48%	93.20%	91.48%	92.40%
	Subset Size	960	32	945	28	962	32	932	**23**	932	25
Brain	*Accuracy*	84.00%	95.44%	84.00%	**95.45%**	84.00%	**95.45%**	84.00%	**95.45%**	84.00%	**95.45%**
	Subset Size	3819	**65**	3819	66	3819	**65**	3819	**65**	3819	66
Lung	*Accuracy*	81.82%	**94.99%**	81.82%	94.29%	80.30%	93.57%	83.33%	92.86%	83.33%	93.57%
	Subset Size	2563	33	2563	32	2563	36	2563	**30**	2563	32
Prostate	*Accuracy*	82.61%	95.65%	84.06%	95.00%	79.71%	95.65%	82.61%	**96.52%**	81.16%	**96.52%**
	Subset Size	3846	34	3846	**32**	3846	**32**	3846	37	3846	39
Endometrium	*Accuracy*	79.57%	89.81%	80.95%	90.74%	80.95%	**91.20%**	78.57%	90.74%	80.95%	90.28%
	Subset Size	942	**51**	942	63	942	64	942	62	942	63

Each gene expression dataset has its natural content and causes a different level of accuracy and subset sizes. The proposed method obtained a high accuracy classification performance on DLBCL-Blood dataset for each of the imputation method that is employed. This shows that dominant features of that dataset were successfully identified. *k*NN-PO resulted in the best classification performance (93.2%) for this dataset. However, we observed that our work accomplishes very close outcomes for each of the imputation method we studied. Likewise, a drastic increase in accuracy is observed for the Prostate dataset, the *k*NN-PO yielded the best classification performance reaching 98.26%. It should be noted that, the main similarity here is that both of these datasets include binary class data.

The applied framework is provided better performance with two of the multiclass datasets. Increases in performance are observed between 2% to 5% on accuracy. However, for the Endometrium dataset accuracy performance is slightly lower than the previous study.

It is observed that the different imputation methods could not easily be separated in terms of accuracy. There is no clear distinction for the best imputation method to generalise for imputation technique. When we change the classifier, the highest accuracies changed, occasionally switching to a different imputation method. In addition, regardless of which imputation method was used, feature

subsets of similar size were chosen. These shows that the recovery of missing data with different imputation methods do not provide a significant benefit among themselves. But we still consider that the difference in classifier performance highly dependent on variance of the gene expressions. The main benefit of our PO-based study is its ability to narrow down the genes from a thousand of genes. Moreover, these genes might more likely related to samples having the disease As seen in Table 6.3, accuracy values for the DLBCL(Blood) and Prostate datasets are increased dramatically. This can provide a baseline for further studies.

Missing values in datasets is challenging and it is hard to gain maximum accuracy because recovering missing values is limited to the number of other samples and general structure of the dataset.

The results reported in this chapter provided improvements in 4 out of 5 the dataset considered. In two datasets, namely DLBCL(Blood) and Prostate, a significant improvement in results were achieved. Of the remaining two datasets, the results from the proposed framework for Brain and Lung showed 3% to 5% higher accuracy respectively. Only for the Endometrium dataset was accuracy slightly lower than in the previous study.

The Endometrium dataset was a challenging dataset with a huge number of genes however only a small number of samples utilizing such a gene set. It has 42 samples and 24,192 genes. In this case, multi-criteria feature selection could not improve the results.

6.5 Conclusion

The impact of common imputation methods on feature selection for different classification methods with PO-based framework has been analysed and their performance accuracy is compared against existing related works (Souto, Jaskowiak and Costa, 2015). There are not many studies that evaluate the contributions of imputation methods on ensemble methods. The experimental evaluation showed that the proposed framework based on ensemble feature selection methods performed reasonably well on datasets with missing values using different data imputation methods to recover missing gene information. Moreover, the outcome of this study provides a perspective on how imputation

methods affect different disease-related gene expression datasets and the performance of different classifiers combined with ensemble feature selection methods.

7 CONCLUSION AND FUTURE WORK

7.1 Thesis Summary

This thesis proposed, for the first time, a Pareto optimal based gene selection framework for biomarker discovery in microarray datasets. The proposed framework aims to identify a subset of disease relevant genes in microarray datasets. The framework integrates different filter methods combined with cross-validation and various classification approaches. A PO method was used to combine the results of multiple learning methods to represent the data from different perspectives. While the proposed framework has been thoroughly evaluated on various well-known microarray datasets, its applicability can be further extended to other datasets, which contain a low number of high-dimensional samples. The experimental results have shown that the proposed framework can achieve comparable or higher predictive accuracy with relatively fewer features.

Another problem in representative gene selection is the dependency on the employed feature selection method where different filtering methods can achieve different subsets of genes. The use of PO method again proved to provide a viable solution to this problem and was able to select relevant subsets of genes.

Class imbalance is one of the challenging problems in data classification and bioinformatics datasets are no exception. The ability of the proposed framework to handle imbalanced microarray datasets was explored. It was shown that the proposed PO based framework can leverage the power of combining various data sampling methods in order to select optimal subsets of genes that are less biased towards the majority class.

Missing data is another crucial problem that was considered by the proposed framework given that microarray gene-expression datasets often contain missing values, which can result in a biased gene classification. The ability of the proposed framework to benefit from common data imputation methods was investigated in this thesis. The effect of common imputation methods on the performance of the proposed Pareto optimal based gene selection framework

was been analysed. The obtained results showed that the proposed framework can cope reasonably well with datasets presenting missing values.

7.2 Summary of Core Contributions

7.2.1 Pareto Optimal Framework for Feature Selection on Biological Data Sets

The primary contribution of this thesis is a new Pareto Optimal framework to select a subset of genes from microarray gene-expression datasets. The dependencies of the selection process were observed on a sample dataset. The results showed that even in almost identical subsets formed by excluding one sample of the entire dataset and including the other, attribute selection methods could produce different selections. The experiments showed that the results obtained with a single method or a single training set are unreliable because different data sets and different methods select different attributes. In order to find solutions to this problem, multi-criteria approaches have been proposed to improve the feature selection methods. The proposed methods have been applied to different biological data sets.

An aggregated cross-validation PO framework combined different filter methods, t-test, entropy, Wilcoxon, ROC, and Bhattacharya. Moreover, different data variations combined with SFS that is a wrapper method. This framework is applied to three different data conditions. As a result, it has been shown that the PO-based hybrid approach, which combines local experts in multi-criteria decision-making, can be used to select the genes most relevant to various cancer diseases. In addition, PO minimized dependence on data set diversity using multiple criteria to select genes. Furthermore, when selecting the most successful gene subset, the number of subsets was determined by the method when the PO did not need parameters such as the number of elements to be selected or the selection threshold. Applying PO with a combination of different methods increases the success of attribute selection methods.

The results in three studies with multiple datasets demonstrate the benefit of ensemble framework with on multi-objective method, namely PO. Progressive improvement on the subsets of each dataset provided higher accuracy than previous studies and narrowed down subsets for further biological studies.

7.2.2 Gene Selection on Imbalanced Datasets

The proposed multi-criterion mixing method was applied to the imbalanced data set. Again, the dependence on data set variation was minimized by using multiple learning sets with PO method. In addition, it is aimed to reduce the method dependence of the feature selection process by using a combination set of 5 different methods. Moreover, natural distribution, balanced distribution and support method has been reached with more inclusive gene clusters. Better recognition of the minority class by the genes we selected confirms our results. The results show that additionally selected genes are useful in identifying the minority class.

Then imbalanced classes supported by balanced subsets and gained higher discriminative genes and bring some noise for majority class identification.

7.2.3 Gene Selection on Datasets Having Missing Values

Imputation methods compared for feature selection on the PO-based framework. Effectiveness of each method has been measured. k-NN based imputation presented a slightly better contribution to features selection for classification.

7.3 Future Research Directions

The successful integration of local methods with multi-criteria decision-making methods suggests that local attribute selection methods can also be combined to further optimise both the size and relevance of the selected subsets of the genes. This possibility can be considered as a future extension of the proposed framework.

In microarray data, the proposed hybrid method can also be used to reduce the relatively large number of genes selected by conventional microarray methods. The applicability of the proposed framework can be further extended to other microarray analysis applications such as predictive modelling of gene expression, gene sequencing and building genomic diagnostic tools.

For imbalanced classes, PO based framework can adjust learning difference as inverse class ratio, that means making the minority class as majority class for sub-dataset when dataset has enough sample size for the minority class.

Following the promising results of this study, future work could investigate the impact of different PO thresholds and consider the different levels of dominated features to boost accuracy for different datasets. Also, the efficiency of various combinations of filter methods and classifier models can be further explored using the proposed framework.

Finally, the prioritisation of genes is another critical issue. The analytical hierarchy process could be integrated into to the proposed framework, which will allow more detailed experiments to be conducted with the subsets.

BIBLIOGRAPHY

Abdi, M. J., Hosseini, S. M. and Rezghi, M. (2012) 'A Novel Weighted Support Vector Machine Based on Particle Swarm Optimization for Gene Selection and Tumor Classification', 2012. doi: 10.1155/2012/320698.

Ahmed, S., Zhang, M. and Peng, L. (2015) *Genetic Programming for Biomarker Detection in Mass Spectrometry Data*. Victoria University of Wellington. doi: 10.1007/978-3-642-35101-3_23.

Alizadeh, A. A. *et al.* (2000) 'Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling', *Nature*, 403(6769), pp. 503–511. doi: 10.1038/35000501.

Alon (1999) 'Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligo', in *Proceedings of National Academy of Sciences of the United States of American*, pp. 6745–6750. doi: 10.1136/adc.26.126.106.

Alpaydin, E. (2010) *Introduction to Machine Learning Second Edition*, *The Adaptive Computation and Machine Learning*.

Alshamlan, H. M., Badr, G. H. and Alohali, Y. A. (2015) 'Genetic Bee Colony (GBC) algorithm : A new gene selection method for microarray cancer classification', *Computational Biology and Chemistry*. Elsevier Ltd, 56, pp. 49–60. doi: 10.1016/j.compbiolchem.2015.03.001.

Anaissi, A. and Kennedy, P. J. (2011) 'Feature Selection of Imbalanced Gene Expression Microarray Data', pp. 73–78. doi: 10.1109/SNPD.2011.12.

Ang, J. C. *et al.* (2016) 'Supervised , Unsupervised , and Semi-Supervised Feature Selection : A Review on Gene Selection', 13(5), pp. 971–989.

Armanfard, N., Reilly, J. P. and Komeili, M. (2016) 'Local feature selection for data classification', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(6), pp. 1217–1227. doi: 10.1109/TPAMI.2015.2478471.

Baldi, P. and Brunak, S. (2001) *Bioinformatics: the machine learning approach*.

Bellman, R. E. (1961) 'Adaptive control processes: A guided tour', *Princeton*

University Press, 28, pp. 1–19. doi: 10.1002/nav.3800080314.

Bolon-Canedo, Veronica *et al.* (2015) 'An insight on complexity measures and classification in microarray data', *Proceedings of the International Joint Conference on Neural Networks*, 2015-Septe. doi: 10.1109/IJCNN.2015.7280302.

Bolon-Canedo, V. *et al.* (2015) 'Recent advances and emerging challenges of feature selection in the context of big data', *Knowledge-Based Systems*, 86, pp. 33–45. doi: 10.1016/j.knosys.2015.05.014.

Bolon-Canedo, V. *et al.* (2017) 'Exploring the consequences of distributed feature selection in DNA microarray data', *Proceedings of the International Joint Conference on Neural Networks*, 2017-May, pp. 1665–1672. doi: 10.1109/IJCNN.2017.7966051.

Bolón-canedo, V. *et al.* (2014) 'A review of microarray datasets and applied feature selection methods', 282, pp. 111–135. doi: 10.1016/j.ins.2014.05.042.

Bolón-Canedo, V. *et al.* (2013) 'A review of feature selection methods on synthetic data', *Knowledge and Information Systems*, 34(3), pp. 483–519. doi: 10.1007/s10115-012-0487-8.

Bolón-Canedo, V. and Alonso-Betanzos, A. (2019) 'Ensembles for feature selection: A review and future trends', *Information Fusion*, 52(November 2018), pp. 1–12. doi: 10.1016/j.inffus.2018.11.008.

Bolón-Canedo, V., Sánchez-Maroño, N. and Alonso-Betanzos, A. (2012) 'An ensemble of filters and classifiers for microarray data classification', *Pattern Recognition*, 45(1), pp. 531–539. doi: 10.1016/j.patcog.2011.06.006.

Bolón-Canedo, V., Sánchez-Maroño, N. and Alonso-Betanzos, A. (2014) 'Data classification using an ensemble of filters', *Neurocomputing*, 135, pp. 13–20. doi: 10.1016/j.neucom.2013.03.067.

Bolón-Canedo, V., Sánchez-Maroño, N. and Alonso-Betanzos, A. (2015) *Feature Selection for High-Dimensional Data*. Switzerland: Springer,Cham. doi: https://doi.org/10.1007/978-3-319-21858-8.

Bolón-Canedo, V., Sánchez-Maroño, N. and Alonso-Betanzos, A. (2016)

‘Feature selection for high-dimensional data’, *Progress in Artificial Intelligence*. Springer Berlin Heidelberg, 5(2), pp. 65–75. doi: 10.1007/s13748-015-0080-y.

Bonilla-Huerta, E. *et al.* (2016) ‘Hybrid Framework Using Multiple-Filters and an Embedded Approach for an Efficient Selection and Classification of Microarray Data’, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 13(1), pp. 12–26. doi: 10.1109/TCBB.2015.2474384.

Bradley, A. P. . (1997) ‘The Use of the Area Under The ROC Curve in the Evaluation Of Machine Learning Algorithms’, *Biological Chemistry*, 378(7), pp. 697–699. doi: 10.1515/bchm.1997.378.7.697.

Bredel, M. *et al.* (2005) ‘Functional network analysis reveals extended gliomagenesis pathway maps and three novel MYC-interacting genes in human gliomas’, *Cancer Research*, 65(19), pp. 8679–8689. doi: 10.1158/0008-5472.CAN-05-1204.

Celton, M. *et al.* (2012) ‘Comparative analysis of missing value imputation To cite this version : Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments’.

Cesario, A. and Marcus (2011) *Cancer Systems Biology, Bioinformatics and Medicine*. Springer.

Chai, L. E. *et al.* (2014) ‘Investigating the effects of imputation methods for modelling gene networks using a dynamic Bayesian network from gene expression data’, *Malaysian Journal of Medical Sciences*, 21(2), pp. 20–27.

Chawla, N. V. *et al.* (2003) ‘SMOTEBoost: Improving prediction of the minority class in boosting’, in *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, pp. 107–119.

Chen, J. *et al.* (2007) ‘Gene selection with multiple ordering criteria’, *BMC bioinformatics*, 8, p. 74. doi: 10.1186/1471-2105-8-74.

Choi, H. S. *et al.* (2018) ‘Deep learning based low-cost high-accuracy diagnostic framework for dementia using comprehensive neuropsychological assessment profiles’, *BMC Geriatrics*, 18(1). doi: 10.1186/s12877-018-0915-z.

Chuang, L. Y. *et al.* (2011) 'A hybrid feature selection method for DNA microarray data', *Computers in Biology and Medicine*. Elsevier, 41(4), pp. 228–237. doi: 10.1016/j.compbiomed.2011.02.004.

Cover, T. and Hart, P. (1967) 'Nearest neighbor pattern classification', *Information Theory, IEEE Transactions on*, 13(1), pp. 21–27. doi: 10.1109/TIT.1967.1053964.

Dash, R. and Misra, B. (2017) 'Gene selection and classification of microarray data: A Pareto de approach', *Intelligent Decision Technologies*, 11(1), pp. 93–107. doi: 10.3233/IDT-160280.

Deb, K. and Raji Reddy, A. (2003) 'Reliable classification of two-class cancer data using evolutionary algorithms', *BioSystems*, 72(1–2), pp. 111–129. doi: 10.1016/S0303-2647(03)00138-2.

Ding, Y. and Wilkins, D. (2006) 'Improving the performance of SVM-RFE to select genes in microarray data.', *BMC bioinformatics*, 7 Suppl 2, p. S12. doi: 10.1186/1471-2105-7-S2-S12.

Dinu, L. P. and Popescu, M. (2008) 'A multi-criteria decision method based on rank distance', *Fundamenta Informaticae*. IOS Press, 86(1, 2), pp. 79–91.

Eekhout, I. *et al.* (2012) 'Brief Report: Missing Data: A Systematic Review of How They Are Reported and Handled', *Epidemiology*. JSTOR, pp. 729–732.

Eisenhaber, F. (2008) 'Introduction to Bioinformatics. By Arthur M. Lesk', *Biotechnology Journal*, 3(11), pp. 1452–1453. doi: 10.1002/biot.200800277.

Enders, C. K. (2010) *Applied Missing Data Analysis*, *Journal of Chemical Information and Modeling*. New York: The Guilford Press. doi: 10.1017/CBO9781107415324.004.

Fattah, I. M. A. *et al.* (2013) 'A TOPSIS based Method for Gene Selection for Cancer Classification', *International Journal of Computer Applications*, 67(17), pp. 39–44. doi: 10.5120/11490-7195.

Fawcett, T. (2006) 'An introduction to ROC analysis', 27, pp. 861–874. doi: 10.1016/j.patrec.2005.10.010.

Fernandez, A., Garcia, S. and Herrera, F. (2011) 'Addressing the Classification with Imbalanced Data : Open Problems and New Challenges on Class Distribution In many applications , there exists a significant difference between the class prior Addressing the Classification with Imbalanced', *Hybrid Artificial Intelligent Systems*, pp. 1–10.

Fleury, G. *et al.* (2002) 'Pareto analysis for gene filtering in microarray experiments', *European Signal Processing Conference*, 2002-March.

Fox, R. J. and Dimmic, M. W. (2006) 'A two-sample Bayesian t-test for microarray data', *BMC Bioinformatics*, 7, pp. 1–11. doi: 10.1186/1471-2105-7-126.

Garber, M. E. *et al.* (2001) 'Diversity of gene expression in adenocarcinoma of the lung', *Proceedings of the National Academy of Sciences*, 98(24), pp. 13784–13789. doi: 10.1073/pnas.241500798.

García-Laencina, P. J., Sancho-Gómez, J. L. and Figueiras-Vidal, A. R. (2010) 'Pattern classification with missing data: A review', *Neural Computing and Applications*, 19(2), pp. 263–282. doi: 10.1007/s00521-009-0295-6.

genome.gov (2020) *No Title, Online*. Available at: https://www.genome.gov/genetics-glossary/acgt (Accessed: 14 July 2020).

Gerlein, E. A. *et al.* (2016) 'Evaluating machine learning classification for financial trading: An empirical approach', *Expert Systems with Applications*, 54, pp. 193–207. doi: 10.1016/j.eswa.2016.01.018.

Guorong, X., Peiqi, C. and Minhui, W. (1996) 'Bhattacharyya distance feature selection', in *Proceedings - International Conference on Pattern Recognition*, pp. 195–199. doi: 10.1109/ICPR.1996.546751.

Guyon, I. *et al.* (2006) 'Feature Extraction Foundations', pp. 1–8.

Haixiang, G. *et al.* (2016) 'BPSO-Adaboost-KNN ensemble learning algorithm for multi-class imbalanced data classification', *Engineering Applications of Artificial Intelligence*. Elsevier, 49, pp. 176–193. doi: 10.1016/j.engappai.2015.09.011.

Haixiang, G. *et al.* (2017) 'Learning from class-imbalanced data: Review of

methods and applications', *Expert Systems with Applications*. Elsevier Ltd, 73, pp. 220–239. doi: 10.1016/j.eswa.2016.12.035.

Handl, J., Kell, D. B. and Knowles, J. (2007) 'Multiobjective optimization in bioinformatics and computational biology', *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4(2), pp. 279–291. doi: 10.1109/TCBB.2007.070203.

Hasnat, A. (2016) 'Feature Selection in Cancer Microarray Data using Multi-Objective Genetic Algorithm combined with Correlation Coefficient'.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) 'Chapter 15 - Random Forests', *The Elements of Statistical Learning*, Springer. doi: 10.1007/b94608.

Haury, A. (2013) 'Feature selection from gene expression data : molecular signatures for breast cancer prognosis and gene regulation network inference l ' École nationale supérieure des mines de Paris Sélection de variables à partir de données d ' expression Feature select'.

Haury, A. C., Gestraud, P. and Vert, J. P. (2011) 'The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures', *PLoS ONE*, 6(12), pp. 1–12. doi: 10.1371/journal.pone.0028210.

Hawkins, D. M. (2004) 'The Problem of Overfitting', *Journal of Chemical Information and Computer Sciences*, pp. 1–12. doi: 10.1021/ci0342472.

Hira, Z. M. and Gillies, D. F. (2015) 'A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data', *Advances in Bioinformatics*, 2015(1), pp. 1–13. doi: 10.1155/2015/198363.

Ho, T. K. (1998) 'The random subspace method for constructing decision forests', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8), pp. 832–844. doi: 10.1109/34.709601.

Hong, J. H. and Cho, S. B. (2009) 'Gene boosting for cancer classification based on gene expression profiles', *Pattern Recognition*, 42(9), pp. 1761–1767. doi: 10.1016/j.patcog.2009.01.006.

Japkowicz, N. and Stephen, S. (2002) 'The class imbalance problem: A systematic study', *Intelligent Data Analysis*, 6(5), pp. 429–449. doi:

10.3233/ida-2002-6504.

Jie, G. and Gao, J. (2010) ‘Model and algorithm of vehicle routing problem with time windows in stochastic traffic network’, *2010 International Conference on Logistics Systems and Intelligent Management (ICLSIM)*, 2, pp. 848–851. doi: 10.1109/iclsim.2010.5461065.

Jirapech-Umpai, T. and Aitken, S. (2005) ‘Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes’, *BMC Bioinformatics*, 6, pp. 1–11. doi: 10.1186/1471-2105-6-148.

Jovic, A., Brkic, K. and Bogunovic, N. (2015) ‘A review of feature selection methods with applications’, *Ieee*, pp. 1200–1205. doi: 10.1109/MIPRO.2015.7160458.

Kacem, I., Hammadi, S. and Borne, P. (2002) ‘Pareto-optimality approach for flexible job-shop scheduling problems: Hybridization of evolutionary algorithms and fuzzy logic’, *Mathematics and Computers in Simulation*, 60(3–5), pp. 245–276. doi: 10.1016/S0378-4754(02)00019-8.

KaltenbachHans-Michael (2013) *The Naïve Bayes Model for Unsupervised Word Sense Disambiguation- Aspects Concerning Feature Selection*. Springer. doi: 10.1007/978-81-322-0763-4.

Khoshgoftaar, T. M. *et al.* (2013) ‘A Survey of Stability Analysis of Feature Subset Selection Techniques’, pp. 424–431.

Kim, H., Golub, G. H. and Park, H. (2005) ‘Missing value estimation for DNA microarray gene expression data: Local least squares imputation’, *Bioinformatics*, 21(2), pp. 187–198. doi: 10.1093/bioinformatics/bth499.

Kim, T., Chung, B. Do and Lee, J. S. (2017) ‘Incorporating receiver operating characteristics into naive Bayes for unbalanced data classification’, *Computing*, 99(3), pp. 203–218. doi: 10.1007/s00607-016-0483-z.

Kira, K. and Rendell, L. A. (1992) ‘A Practical Approach to Feature Selection’, *Machine Learning Proceedings 1992*, pp. 249–256. doi: 10.1016/b978-1-55860-247-2.50037-1.

Kubat, M. and Matwin, S. (1997) ‘Addressing the curse of imbalanced training

sets: one-sided selection', in *Icml*. Nashville, USA, pp. 179–186.

Lapointe, J. *et al.* (2004) 'Gene expression profiling identifies clinically relevant subtypes of prostate cancer', *Proceedings of the National Academy of Sciences*, 101(3), pp. 811–816. doi: 10.1073/pnas.0304146101.

Lee, J. S. and Zhu, D. (2011) 'When costs are unequal and unknown: A subtree grafting approach for unbalanced data classification', *Decision Sciences*, 42(4), pp. 803–829. doi: 10.1111/j.1540-5915.2011.00332.x.

Lee, P. H., Jung, J. Y. and Shatkay, H. (2010) 'Functionally informative tag SNP selection using a pareto-optimal approach', in *Advances in Experimental Medicine and Biology*, pp. 173–180. doi: 10.1007/978-1-4419-5913-3_20.

Li, G. Z., Meng, H. H. and Ni, J. (2008) 'Embedded gene selection for imbalanced microarray data analysis', *3rd International Multi-Symposiums on Computer and Computational Sciences, IMSCCS'08*, pp. 17–24. doi: 10.1109/IMSCCS.2008.33.

Li, S., Wu, X. and Hu, X. (2008) 'Gene selection using genetic algorithm and support vectors machines', *Soft Computing*, 12(7), pp. 693–698. doi: 10.1007/s00500-007-0251-2.

Li, X., Li, M. and Yin, M. (2016) 'Multiobjective ranking binary artificial bee colony for gene selection problems using microarray datasets', *IEEE/CAA Journal of Automatica Sinica*, pp. 1–16. doi: 10.1109/jas.2016.7510034.

Li, Y., Li, T. and Liu, H. (2017) 'Recent advances in feature selection and its applications', *Knowledge and Information Systems*. Springer London, 53(3), pp. 551–577. doi: 10.1007/s10115-017-1059-8.

Liang, Y. *et al.* (2016) *Support Vector Machines and Their Application in Chemistry and Biotechnology, Support Vector Machines and Their Application in Chemistry and Biotechnology*. doi: 10.1201/b10911.

Liew, A. W. C., Law, N. F. and Yan, H. (2011) 'Missing value imputation for gene expression data: Computational techniques to recover missing data from available information', *Briefings in Bioinformatics*, 12(5), pp. 498–513. doi: 10.1093/bib/bbq080.

Ling, C. X. and Sheng, V. S. (2010) ‘Cost-sensitive learning’, *Encyclopedia of machine learning*. Springer, pp. 231–235.

Little, R. J. A. and Rubin, D. B. (2019) *Statistical analysis with missing data*. Wiley.

López, V. *et al.* (2013) ‘An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics’, *Information Sciences*. Elsevier Inc., 250, pp. 113–141. doi: 10.1016/j.ins.2013.07.007.

Luo, L. *et al.* (2011) ‘Methods of forward feature selection based on the aggregation of classifiers generated by single attribute’, *Computers in Biology and Medicine*. Elsevier, 41(7), pp. 435–441. doi: 10.1016/j.compbiomed.2011.04.005.

Luscombe, N. M., Greenbaum, D. and Gerstein, M. (2001) ‘Review What is bioinformatics ? An Introduction and overview’, *Year Book of Medical Informatics*, pp. 83–100.

Magzoub, M. M. *et al.* (2019) ‘The impact of DNA methylation on the cancer proteome’, *PLOS Computational Biology*, 15(7), p. e1007245. doi: 10.1371/journal.pcbi.1007245.

Mahajan, S., Abhishek and Singh, S. (2016) ‘Review On Feature Selection Approaches Using Gene Expression Data’, *Imperial Journal of Interdisciplinary Research*, 2(3), pp. 356–364.

Makałowski, W., Jakalski, M. and Makałowska, I. (2014) ‘Bioinformatics’, *In: eLS*. Chichester: John Wiley & Sons, Ltd, pp. 131–156. doi: 10.1002/9780470015902.a0005247.pub2.

Maldonado, S., Weber, R. and Famili, F. (2014) ‘Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines’, *Information Sciences*, 286, pp. 228–246. doi: 10.1016/j.ins.2014.07.015.

Malley, J. D., Dasgupta, A. and Moore, J. H. (2013) ‘The limits of p-values for biological data mining’, *BioData Mining*. doi: 10.1186/1756-0381-6-10.

Mandoiu, I. and Zelikovsky, A. (2008) *Bioinformatics Algorithms : Techniques*

and Applications. Hoboken, NJ, USA: John Wiley & Sons, Incorporated, 2008.

Mary-Huard, T., Picard, F. and Robin, S. (2006) 'Introduction to statistical methods for microarray data analysis', *Mathematical and Computational Methods in Biology. Paris: Hermann*.

Metz, C. E. (1978) 'Basic principles of ROC analysis', *Seminars in Nuclear Medicine*, 8(4), pp. 283–298. doi: 10.1016/S0001-2998(78)80014-2.

Michiels, S., Koscielny, S. and Hill, C. (2005) 'Prediction of cancer outcome with microarrays: a multiple random validation strategy', *Health (San Francisco)*, 365, pp. 488–492. doi: 10.1016/S0140-6736(05)66541-5.

Mohamad, M. S. *et al.* (2013) 'An Enhancement of Binary Particle Swarm Optimization for Gene Selection in Classifying Cancer Classes', *Algorithms For Molecular Biology. Biomed Central. Indexed by ISI and SCOPUS.*, 8(15), pp. 1–11.

Montague, M. and Aslam, J. A. (2004) 'Condorcet fusion for improved retrieval', pp. 538–548. doi: 10.1145/584792.584881.

Mooney, M. A. and Wilmot, B. (2015) 'Gene set analysis: A step-by-step guide', *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics*, 168(7), pp. 517–527. doi: 10.1002/ajmg.b.32328.

Moorthy, K., Mohamad, M. and Deris, S. (2014) 'A Review on Missing Value Imputation Algorithms for Microarray Gene Expression Data', *Current Bioinformatics*, 9(1), pp. 18–22. doi: 10.2174/15748936089991401091209 57.

Moosa, J. M. *et al.* (2016) 'Gene selection for cancer classification with the help of bees', 9(Suppl 2). doi: 10.1186/s12920-016-0204-7.

Morán-Fernández, L., Bolón-Canedo, V. and Alonso-Betanzos, A. (2017) 'Can classification performance be predicted by complexity measures? A study using microarray data', *Knowledge and Information Systems*. Springer London, 51(3), pp. 1067–1090. doi: 10.1007/s10115-016-1003-3.

Moreau, Y. and Tranchevent, L. C. (2012) 'Computational tools for prioritizing candidate genes: Boosting disease gene discovery', *Nature Reviews Genetics*. Nature Publishing Group, 13(8), pp. 523–536. doi: 10.1038/nrg3253.

Mortazavi, A. and Moattar, M. H. (2016) 'Robust feature selection from microarray data based on cooperative game theory and qualitative mutual information', *Advances in Bioinformatics*, 2016. doi: 10.1155/2016/1058305.

Mount, D. W. (2004) 'Bioinformatics Sequence and Genome Analysis', *A Primer on String Theory*, pp. 1–10. doi: 10.1017/9781316672631.002.

Muresan, S. *et al.* (2015) 'Pre-processing flow for enhancing learning from medical data', *Proceedings - 2015 IEEE 11th International Conference on Intelligent Computer Communication and Processing, ICCP 2015*, pp. 27–34. doi: 10.1109/ICCP.2015.7312601.

Murie, C. *et al.* (2009) 'Comparison of small n statistical tests of differential expression applied to microarrays', *BMC Bioinformatics*, 10. doi: 10.1186/1471-2105-10-45.

Nguyen, T. *et al.* (2015) 'Hierarchical gene selection and genetic fuzzy system for cancer microarray data classification', *PLoS ONE*, 10(3), pp. 1–23. doi: 10.1371/journal.pone.0120364.

Niijima, S. and Kuhara, S. (2006) 'Recursive gene selection based on maximum margin criterion: a comparison with SVM-RFE.', *BMC bioinformatics*, 7, p. 543. doi: 10.1186/1471-2105-7-543.

Nilsson, R. (2007) *Statistical Feature Selection with Applications in Life Science, Department of Physics, Chemistry and Biology*.

Niu, B. *et al.* (2017) 'Analysis and Modeling for Big Data in Cancer Research', *BioMed Research International*, 2017. doi: 10.1155/2017/1972097.

Oba, S. *et al.* (2003) 'A Bayesian missing value estimation method for gene expression profile data', *Bioinformatics*, 19(16), pp. 2088–2096. doi: 10.1093/bioinformatics/btg287.

Ochs, M. F., Casagrande, J. T. and Davuluri, R. V. (2010) *Biomedical informatics for cancer research, Biomedical Informatics for Cancer Research*. doi: 10.1007/978-1-4419-5714-6.

Ogutcen, O. F. *et al.* (2016) 'An aggregated cross-validation framework for computational discovery of disease-associative genes', in *IFMBE Proceedings*.

doi: 10.1007/978-3-319-32703-7_94.

Ogutcen, O. F., Belatreche, A. and Seker, H. (2018) ‘A multi-objective pareto-optimal wrapper based framework for cancer-related gene selection’, *Advances in Intelligent Systems and Computing*, 869, pp. 353–364. doi: 10.1007/978-3-030-01057-7_28.

Pardalos, P. and Du, D.-Z. (2008) *PARETO OPTIMALITY , GAME THEORY AND EQUILIBRIA Springer Optimization and Its Applications*. doi: 10.1007/978-0-387-77247-9.

Peng, Y., Li, W. and Liu, Y. (2017) ‘A Hybrid Approach for Biomarker Discovery from Microarray Gene Expression Data for Cancer Classification’, *Cancer Informatics*, 2, p. 117693510600200. doi: 10.1177/117693510600200024.

Petricoin, E. F. *et al.* (2002) ‘Use of proteomic patterns in serum to identify ovarian cancer’, *Lancet*, 359(9306), pp. 572–577. doi: 10.1016/S0140-6736(02)07746-2.

Pohjalainen, J., Räsänen, O. and Kadioglu, S. (2015) ‘Feature selection methods and their combinations in high-dimensional classification of speaker likability , intelligibility and personality traits ’, *Computer Speech & Language*. Elsevier Ltd, 29(1), pp. 145–171. doi: 10.1016/j.csl.2013.11.004.

Pomeroy, S. L. *et al.* (2002) ‘Prediction of central nervous system embryonal tumour outcome based on gene expression’, *Nature*, 415(6870), pp. 436–442. doi: 10.1038/415436a.

Prakash, S. *et al.* (2008) ‘Pareto optimal solutions of a cost-time trade-off bulk transportation problem’, *European Journal of Operational Research*, 188(1), pp. 85–100. doi: 10.1016/j.ejor.2007.03.040.

Pudil, P., Novovicova, J. and Kittler, J. (1994) ‘Floating Search Methods in Feature-Selection’, *Pattern Recognition Letters*, 15(11), pp. 1119–1125. doi: 10.1016/0167-8655(94)90127-9.

Reunanen, J. (2003) ‘Overfitting in making comparisons between variable selection methods’, *Journal of Machine Learning Research*, 3(Mar), pp. 1371–

1382.

Risinger, J. I. *et al.* (2003) 'Microarray analysis reveals distinct gene expression profiles among different histologic types of endometrial cancer', *Cancer Research*, 63(1), pp. 6–11.

Sabzevari, S. and Abdullah, S. (2011) 'Gene selection in microarray data from multi-objective perspective', *Conference on Data Mining and Optimization*, (June), pp. 199–207. doi: 10.1109/DMO.2011.5976528.

Saeys, Y. *et al.* (2007) 'A review of feature selection techniques in bioinformatics', *Bioinformatics*, 23(19), pp. 2507–2517. doi: 10.1093/bioinformatics/btm344.

Samiappan, S., Prasad, S. and Bruce, L. M. (2013) 'Non-uniform random feature selection and kernel density scoring with SVM based ensemble classification for hyperspectral image analysis', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6(2), pp. 792–800. doi: 10.1109/JSTARS.2013.2237757.

Sarac, F. *et al.* (2015) 'Comparison of unsupervised feature selection methods for high-dimensional regression problems in prediction of peptide binding affinity', *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2015-Novem, pp. 8173–8176. doi: 10.1109/EMBC.2015.7320291.

Schnattinger, T. *et al.* (2013) 'RNA-Pareto: Interactive analysis of Pareto-optimal RNA sequence-structure alignments', *Bioinformatics*, 29(23), pp. 3102–3104. doi: 10.1093/bioinformatics/btt536.

Seijo-Pardo, B. *et al.* (2015) 'Ensemble feature selection for rankings of features', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 29–42. doi: 10.1007/978-3-319-19222-2_3.

Shipp, M. a *et al.* (2002) 'Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning.', *Nature medicine*, 8(1), pp. 68–74. doi: 10.1038/nm0102-68.

Shoval, O. *et al.* (2012) ‘Evolutionary trade-offs, pareto optimality, and the geometry of phenotype space’, *Science*, 336(6085), pp. 1157–1160. doi: 10.1126/science.1217405.

Silva, H. De and Perera, A. S. (2017) ‘Evolutionary k-Nearest Neighbor Imputation Algorithm for Gene Expression Data’, 10(1), pp. 1–8.

Silvério Lopes, H. and Magalh, L. (2011) *COMPUTATIONAL BIOLOGY AND APPLIED*. Croatia: InTECH.

Soto, A. J. *et al.* (2009) ‘Multi-objective feature selection in QSAR using a machine learning approach’, *QSAR and Combinatorial Science*, 28(11–12), pp. 1509–1523. doi: 10.1002/qsar.200960053.

Souto, M. C. P. D., Jaskowiak, P. A. and Costa, I. G. (2015) ‘Impact of missing data imputation methods on gene expression clustering and classification’, *BMC Bioinformatics*, 16(1), pp. 1–9. doi: 10.1186/s12859-015-0494-3.

Steenhoff, D. *et al.* (2012) ‘A Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis’, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(4), pp. 1106–1119.

Subramanian, J. and Simon, R. (2013) ‘Overfitting in prediction models - Is it a problem only in high dimensions?’, *Contemporary Clinical Trials*, 36(2), pp. 636–641. doi: 10.1016/j.cct.2013.06.011.

Sun, F. *et al.* (2015) ‘Sentiment classification in the financial domain using? SVM and multi-objective optimisation’, in *Proceedings - 2015 IEEE Symposium Series on Computational Intelligence, SSCI 2015*, pp. 910–916. doi: 10.1109/SSCI.2015.134.

Taub, F. E., Deleo, J. M. and Thompson, E. B. (1983) ‘Sequential Comparative Hybridizations Analyzed by Computerized Image Processing Can Identify and Quantitate Regulated RNAs’, *DNA*, 2(4), pp. 309–327. doi: 10.1089/dna.1983.2.309.

Tavakkoli-Moghaddam, R., Azarkish, M. and Sadeghnejad-Barkousaraie, A. (2011) ‘A new hybrid multi-objective Pareto archive PSO algorithm for a bi-objective job shop scheduling problem’, *Expert Systems with Applications*.

Elsevier Ltd, 38(9), pp. 10812–10821. doi: 10.1016/j.eswa.2011.02.050.

Ting, C. K., Lin, W. T. and Huang, Y. T. (2010) 'Multi-objective tag SNPs selection using evolutionary algorithms', *Bioinformatics*, 26(11), pp. 1446–1452. doi: 10.1093/bioinformatics/btq158.

Troyanskaya, O. *et al.* (2001) 'Missing value estimation methods for DNA microarrays', *Bioinformatics*, 17(6), pp. 520–525. doi: 10.1093/bioinformatics/17.6.520.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001) 'Significance analysis of microarrays applied to the ionizing radiation response', *Proceedings of the National Academy of Sciences*, 98(9), pp. 5116–5121. doi: 10.1073/pnas.091062498.

Uncu, O. *et al.* (2007) 'A novel feature selection approach: Combining feature wrappers and filters', *Information Sciences*, 177(2), pp. 449–466. doi: 10.1016/j.ins.2006.03.022.

Uslan, V. (2015) 'Support Vector Machine-based Fuzzy Systems for Quantitative Prediction of Peptide Binding Affinity'.

Vapnik, V. N. (1999) 'An overview of statistical learning theory', *IEEE Transactions on Neural Networks*. doi: 10.1109/72.788640.

Wang, A. *et al.* (2018) 'Microarray Missing Value Imputation: A Regularized Local Learning Method', *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. IEEE, PP, p. 1. doi: 10.1109/TCBB.2018.2810205.

Wang, D. *et al.* (2006) 'Effects of replacing the unreliable cDNA microarray measurements on the disease classification based on gene expression profiles and functional modules', *Bioinformatics*, 22(23), pp. 2883–2889. doi: 10.1093/bioinformatics/btl339.

Wang, J. *et al.* (2013) 'Maximum weight and minimum redundancy: A novel framework for feature subset selection', *Pattern Recognition*. Elsevier, 46(6), pp. 1616–1627. doi: 10.1016/j.patcog.2012.11.025.

Wang, Y., Miller, D. J. and Clarke, R. (2008) 'Approaches to working in high-dimensional data spaces: Gene expression microarrays', *British Journal of*

Cancer, pp. 1023–1028. doi: 10.1038/sj.bjc.6604207.

Webb, G. I. *et al.* (2010) 'Leave-One-Out Cross-Validation', *Encyclopedia of Machine Learning*, pp. 600–601. doi: 10.1007/978-0-387-30164-8_469.

Wei, R. *et al.* (2018) 'Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data', *Scientific Reports*, 8(1). doi: 10.1038/s41598-017-19120-0.

West, M. *et al.* (2001) 'Predicting the clinical status of human breast cancer by using gene expression profiles', *Proceedings of the National Academy of Sciences*, 98(20), pp. 11462–11467. doi: 10.1073/pnas.201162998.

Whitney, A. W. (1971) 'A direct method of nonparametric measurement selection', *IEEE Transactions on Computers*. IEEE, 100(9), pp. 1100–1103.

Williams, J. M. *et al.* (2010) 'OpenHelix: Bioinformatics education outside of a different box', *Briefings in Bioinformatics*, 11(6), pp. 598–609. doi: 10.1093/bib/bbq026.

Winkler, S., Affenzeller, M. and Wagner, S. (2007) 'Advanced genetic programming based machine learning', *Journal of Mathematical Modelling and Algorithms*, 6(3), pp. 455–480. doi: 10.1007/s10852-007-9065-6.

Wu, O. *et al.* (2009) 'Rank aggregation based text feature selection', *Proceedings - 2009 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2009*, 1, pp. 165–172. doi: 10.1109/WI-IAT.2009.32.

Wu, X. *et al.* (2008) *Top 10 algorithms in data mining*. doi: 10.1007/s10115-007-0114-2.

Yang, F. *et al.* (2015) 'Emphasizing minority class in LDA for feature subset selection on high-dimensional small-sized problems', *IEEE Transactions on Knowledge and Data Engineering*, 27(1), pp. 88–101. doi: 10.1109/TKDE.2014.2320732.

Yang, F. and Mao, K. Z. (2010) 'Improving robustness of gene ranking by resampling and permutation based score correction and normalization', *Proceedings - 2010 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2010*, pp. 444–449. doi: 10.1109/BIBM.2010.5706607.

Yang, F. and Mao, K. Z. (2011) 'Robust feature selection for microarray data based on multicriterion fusion', *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(4), pp. 1080–1092. doi: 10.1109/TCBB.2010.103.

Yang, Y., Xu, Z. and Song, D. (2015) 'Missing value imputation for microRNA expression data by using a GO-based similarity measure', *BMC Bioinformatics*, 17(1). doi: 10.1186/s12859-015-0853-0.

Yin, L. *et al.* (2013) 'Feature selection for high-dimensional imbalanced data', *Neurocomputing*. Elsevier, 105, pp. 3–11. doi: 10.1016/j.neucom.2012.04.039.

Yu, L. and Liu, H. (2004) 'Efficient Feature Selection via Analysis of Relevance and Redundancy', *J. Mach. Learn. Res.*, 5, pp. 1205–1224. Available at: http://dl.acm.org/citation.cfm?id=1005332.1044700.

Zhao, Z. A. and Liu, H. (2011) *Spectral Feature Selection for Data Mining (Open Access)*. Chapman and Hall/CRC.

Zhou, Q. *et al.* (2014) 'Stable feature selection with ensembles of multirelieff', *2014 10th International Conference on Natural Computation, ICNC 2014*, (61202144), pp. 742–747. doi: 10.1109/ICNC.2014.6975929.

Zitzler, E. and Thiele, L. (1999) 'Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach', *IEEE Transactions on Evolutionary Computation*, 3(4), pp. 257–271. doi: 10.1109/4235.797969.