

**FEATURE SELECTION BY MUTUAL INFORMATION: ROBUST
RANKING ON HIGH- DIMENSION LOW- SAMPLE-SIZE DATA**

By

CHIN FUNG YUEN

A thesis submitted to the Department of Mathematical and Actuarial Sciences,
Lee Kong Chian Faculty of Engineering and Science,
Universiti Tunku Abdul Rahman,
in partial fulfilment of the requirements for the degree of
Doctor of Philosophy in Science
Sept 2020

ABSTRACT

FEATURE SELECTION BY MUTUAL INFORMATION: ROBUST RANKING ON HIGH- DIMENSION LOW- SAMPLE-SIZE DATA

CHIN FUNG YUEN

Feature selection is a process of selecting a group of relevant features by removing unnecessary features for use in constructing the predictive model. The current benchmark for the data set is obtained by including all the features, such as redundancy and noise. Therefore, for this research, an optimal baseline for the data set will be proposed using the feature ranking method. To achieve this optimal baseline, a total number of features will be obtained at the same time to serve as the guideline on the number of features needed in a feature selection method. In addition, the high dimensional data which increases the difficulty on the features selection due to the curse of dimensionality. To overcome this problem, a robust feature selection algorithm, named ranked mutual information with support vector machine (rMI-SVM) can be applied on the data with missing value regardless of the linearity of the data set, as it does not require additional parameter or preset on the number of features needed. The features selected by rMI-SVM can avoid overfitting as the chosen candidate feature will provide new information to the predictive model. The receiver operating characteristic curve has been plotted to show the sensitivity of the model built by rMI-SVM compared to the regression method under the same number of

features. Also, the Z- score graph was plotted to confirm that the features chosen by rMI-SVM were not selected by chance. The experimental results show that the proposed method can select a compact subset of features that can perform better than the benchmark of the data set and the optimal baseline proposed in this study. The biological meaning of the selected features confirmed that the selected features are related to the relevant disease.


ACKNOWLEDGEMENT

Firstly, I would like to express my sincere gratitude to my supervisor Dr Goh Yong Kheng and my co-supervisor Dr Tan Choon Peng for the continuous support of my PhD study and related research, for their patience, motivation, and immense knowledge. Their guidance helped me in all the time of research and writing of this thesis. I could not have imagined having better supervisors and mentors for my PhD study. Besides my supervisors, I would like to thank you for the sponsorship given by Universiti Tunku Abdul Rahman to complete my research. I would also like to thank the rest of my work completion seminar committee: Dr Pan Wei Yeing and Dr Liew How Hui for their insightful comments and encouragement, but also for the hard question which incited me to widen my research from various perspectives. Also, I thank my colleagues in the Department of Physical and Mathematical Science. Last but not least, I would like to thank my family: my mother and to my sons for supporting me spiritually throughout writing this thesis and my life in general.

APPROVAL SHEET

This thesis entitled **“FEATURE SELECTION BY MUTUAL INFORMATION: ROBUST RANKING ON HIGH- DIMENSION LOW-SAMPLE-SIZE DATA”** was prepared by CHIN FUNG YUEN and submitted as partial fulfilment of the requirements for the degree of Doctor of Philosophy in Science at Universiti Tunku Abdul Rahman.

Approved by:



(Dr. GOH YONG KHENG)

Supervisor

Department of Mathematical and Actuarial Sciences
Lee Kong Chian Faculty of Engineering and Science
Universiti Tunku Abdul Rahman

Date: 10/09/2020



(Dr. TAN CHOON PENG)

Co-supervisor

Department of Mathematical and Actuarial Sciences
Lee Kong Chian Faculty of Engineering and Science
Universiti Tunku Abdul Rahman

10/09/2020
Date:

**LEE KONG CHIAN FACULTY OF ENGINEERING AND SCIENCE
UNIVERSITI TUNKU ABDUL RAHMAN**


Date: 10 September 2020

SUBMISSION OF THESIS

It is hereby certified that **CHIN FUNG YUEN (1002128)** has completed this thesis entitled “FEATURE SELECTION BY MUTUAL INFORMATION: ROBUST RANKING ON HIGH- DIMENSION LOW- SAMPLE-SIZE DATA” under supervision of Dr. Goh Yong Kheng from the Department of Mathematical and Actuarial Sciences, Lee Kong Chian Faculty of Engineering and Science, and Dr. Tan Choon Peng from the Department of Mathematical and Actuarial Sciences, Lee Kong Chian Faculty of Engineering and Science.

I understand that the University will upload softcopy of my thesis in pdf format into UTAR Institutional Repository, which may be made accessible to UTAR community and public.

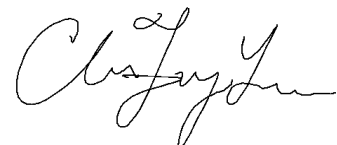
Yours truly,

A handwritten signature in black ink, appearing to read 'Chin Fung Yuen', written in a cursive style.

(CHIN FUNG YUEN)

DECLARATION

I CHIN FUNG YUEN hereby declare that the thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at UTAR or other institutions.

A handwritten signature in black ink, appearing to read 'Chin Fung Yuen', written in a cursive style.

(CHIN FUNG YUEN)

Date 10 September 2020

TABLE OF CONTENTS

	Page
ABSTRACT	ii
ACKNOWLEDGEMENTS	iv
APPROVAL SHEET	v
SUBMISSION SHEET	vi
DECLARATION	vii
LIST OF TABLES	x
LIST OF FIGURES	xv
LIST OF ABBREVIATIONS	xvii
CHAPTER	
1.0 INTRODUCTION	1
1.1 Brief of Bioinformatics	1
1.2 Development of Computational Bioinformatics	3
1.3 Research in Cancer Biomarker	5
1.4 Motivation	8
1.5 Problem Statement	12
1.6 Objectives and Significance of Study	13
2.0 REVIEW ON FEATURE SELECTION	16
2.1 Machine Learning	16
2.2 Dimensional Reduction	17
2.2.1 Feature Extraction	17
2.2.2 Feature Selection	19
2.3 Evolution of Feature Selection on Information Theory	21
2.4 Common Problems and Limitation in the Past Research	30
3.0 RANKED MUTUAL INFORMATION	32
3.1 Information Theory	32
3.1.1 Entropy	32
3.1.2 Conditional Entropy and Joint Entropy	33
3.1.3 Mutual Information	35
3.2 The Optimal Baseline based on Ranked Features	38
3.3 Algorithm on Ranked Features	40
4.0 DIMENSION REDUCTION WITH SUPPORT VECTOR MACHINE	47
4.1 Ranked Mutual Information with Support Vector Machine (rMI-SVM)	47
4.1.1 Dimension Reduction	47
4.1.2 Relevancy and Redundancy	48
4.1.3 Increment of the Information Content	51
4.1.4 Minimal Features	53

4.2	Algorithm on the rMI-SVM	54
5.0	EVALUATION OF rMI-SVM	60
5.1	Classifier	60
5.2	Receiver Operating Characteristic Curve	61
5.3	Robustness	65
5.4	Comparison with other Relevant Work	66
5.5	Influence on the Classifier to the Predictive Model	70
5.6	Experimental Setup	71
6.0	RESULTS	72
6.1	Research Data Set	72
6.2	Experiment Procedure	76
6.3	Evaluation on Four Different SVM Classifiers for Binary Data Set	78
6.4	Optimal baseline of the Binary Data	82
6.5	rMI-SVM for Binary Data Set	87
6.6	Evaluation on Four Different Classifiers for Multiclass Data Set	122
6.7	Optimal Baseline of the Multiclass Data	125
6.8	rMI-SVM for Multiclass Data Set	129
7.0	DISCUSSION	150
7.1	Evaluation of the features selected by the rMI-SVM	150
7.2	Z-score Analysis	151
	7.2.1 Z-score for Binary Data Set	151
	7.2.2 Z-score Analysis for Multiclass Data Set	163
7.3	Cross-validation	171
	7.3.1 Cross-validation for Binary Data Set	171
	7.3.2 Cross-validation for Multiclass Data Set	177
7.4	Comparison of Performance	181
	7.4.1 Comparison of Performance for Binary Data Set	181
	7.4.2 Comparison of Performance for Multiclass Data Set	189
7.5	Biological Meaning of the Selected Features	194
7.6	New Findings	196
7.7	Contribution of Feature Selection in Malaysia's Medical Research	198
8.0	CONCLUSION	201
8.1	Summary	201
8.2	Limitation	203
8.3	Future Study	204
	REFERENCES/ BIBLIOGRAPHY	205

LIST OF TABLES

Table		Page
1.1	Percentage of incidence by gender and ethnic	8
1.2	Percentage of the top 10 most common cancer among male and female	9
3.1	The expression of feature respect to label class	43
3.2	Joint probability mass function of a feature and the label class	44
6.1	Summary of the downloaded data set	73
6.2	Part of the ribosomal protein cluster of colon cancer data set	74
6.3	Average accuracy of cross-validation and predictive model on four different kernel functions of classifier of colon cancer data set	79
6.4	Average accuracy of cross-validation and predictive model on four different kernel functions of classifier of a leukaemia data set	79
6.5	Average accuracy of cross-validation and predictive model on four different kernel function of classifier of a prostate cancer data set	80
6.6	Average accuracy of cross-validation and predictive model on four different kernel functions of classifier of lung cancer data set	80
6.7	Average accuracy of cross-validation and predictive model on four different kernel functions of classifier of Parkinson data set	81
6.8	Average accuracy of cross-validation and predictive model on four different kernel functions of classifier of breast cancer data set	82
6.9	The baseline using full features and the optimal baseline with the number of features	86
6.10	Number of features selected by the rMI-SVM algorithm for six binary data set	88
6.11	Ten-fold cross-validation and average accuracy for four different classifiers using full features, seven features using the rMI-SVM, Regression method and mRMR method of the colon data set	92

6.12	Output of the confusion matrix and the ROC curve of colon cancer data using seven features	94
6.13	Ten-fold cross-validation and average accuracy for four different classifiers using full features, five features using the rMI-SVM, Regression method and mRMR method of the leukaemia data set	97
6.14	Output of the confusion matrix and the ROC curve of leukaemia data using five features	99
6.15	Ten-fold cross-validation and average accuracy for three different classifiers using full features, three features using the rMI-SVM, Regression method and mRMR method of prostate cancer data set	103
6.16	Output of the confusion matrix and the ROC curve of prostate cancer data using three features	105
6.17	Ten-fold cross-validation and average accuracy for three different classifiers using full features, four features using the rMI-SVM, Regression method and mRMR method of lung cancer data set	108
6.18	Output of the confusion matrix and the ROC curve of lung cancer data using four features	110
6.19	Ten-fold cross-validation and average accuracy for four different classifiers using full features, seven features using the rMI-SVM, Regression method and mRMR method of Parkinson data set	113
6.20	Output of the confusion matrix and the ROC curve of Parkinson data using seven features	115
6.21	Ten-fold cross-validation and average accuracy for four different classifiers using full features, eleven features using the rMI-SVM, Regression method and mRMR method of breast cancer data set	119
6.22	Output of the confusion matrix and the ROC curve of breast cancer data using eleven features	121
6.23	Average accuracy of 2-fold cross-validation and predictive model on four different kernel functions of classifier of skin cancer data set	122
6.24	Average accuracy of 2-fold cross-validation and predictive model on four different kernel functions of classifier of the lymphoma data set	123
6.25	Average accuracy of 3-fold cross-validation and predictive model on four different kernel functions of classifier of lung cancer data set	123

6.26	Average accuracy of 10-fold cross-validation and predictive model on four different kernel functions of classifier of the handwriting data set	124
6.27	The baseline using full features and the optimal baseline with the number of features for the multiclass data set	128
6.28	Number of features selected by the rMI-SVM algorithm for four multiclass data set	130
6.29	Two-fold cross-validation and average accuracy for SVM classifier using full features, rMI-SVM, and regression method for skin cancer data set	131
6.30	Output of the ROC curve of skin cancer data using four features	135
6.31	Two-fold cross-validation and average accuracy for SVM classifier using full features, rMI-SVM, and regression method for lymphoma data set	136
6.32	Output of the ROC curve of lymphoma data using 22 features	139
6.33	Three-fold cross-validation and average accuracy for SVM classifier using full features, rMI-SVM, and regression method for lung cancer data set	140
6.34	Output of the ROC curve of lung cancer data using four features	144
6.35	Ten-fold cross-validation and average accuracy for SVM classifier using full features, rMI-SVM, and regression method for handwriting data set	145
6.36	Output of the ROC curve of handwriting data using 50 features	148
7.1	The top 20 ranked features with Z-score selected by regression method, mRMR method and rMI-SVM algorithm for colon cancer data set	152
7.2	The top 30 ranked features with Z-score selected by regression method, mRMR method and rMI-SVM algorithm for leukaemia data set	155
7.3	The top 10 ranked features with Z-score selected by regression method, mRMR method and rMI-SVM algorithm for prostate data set	157
7.4	The top 20 ranked features with Z-score selected by regression method, mRMR method and rMI-SVM algorithm for lung cancer data set	159

7.5	The top 10 ranked features with Z-score selected by regression method, mRMR method and rMI-SVM algorithm for Parkinson data set	161
7.6	The top 10 ranked features with Z-score selected by regression method, mRMR method and rMI-SVM algorithm for breast cancer data set	163
7.7	The top 20 ranked features with Z-score selected by regression method and rMI-SVM algorithm for skin cancer data set	165
7.8	The top 20 ranked features with Z-score selected by regression method and rMI-SVM algorithm for lymphoma data set	167
7.9	The top 15 ranked features with Z-score selected by regression method and rMI-SVM algorithm for lung cancer data set	169
7.10	The top 20 ranked features with Z-score selected by regression method and rMI-SVM algorithm for handwriting data set	171
7.11	Average accuracy with the number of features obtained from several feature selection methods for colon cancer data set	183
7.12	Average accuracy with the number of features obtained from several feature selection methods for leukaemia data set	184
7.13	Average accuracy with the number of features obtained from several feature selection methods for prostate data set	185
7.14	Average accuracy with the number of features obtained from several feature selection methods for lung cancer data set	186
7.15	Average accuracy with the number of features obtained from several feature selection methods for Parkinson data set	187
7.16	Average accuracy with the number of features obtained from several feature selection methods for breast cancer data set	188
7.17	Average accuracy with the number of features obtained from several feature selection methods for skin cancer data set	189
7.18	Average accuracy with the number of features obtained from several feature selection methods for lymphoma data set	191
7.19	Average accuracy with the number of features obtained from several feature selection methods for lung data set	191
7.20	Average accuracy with the number of features obtained from several feature selection methods for handwriting data set	193

LIST OF FIGURES

Figure		Page
3.1	Relation between mutual information and entropy	37
3.2	Flowchart of finding the optimal baseline and the unique cutoff number of features	41
3.3	Joint distribution for a feature in the 3-dimensional histogram	43
3.4	Average accuracy versus the ranked features based on the mutual information score	45
4.1	Flowchart of finding a smaller selected feature	55
4.2	Average accuracy versus the ranked features based on the mutual information score	57
4.3	Average accuracy for a smaller selected feature	59
5.1	Confusion matrix	62
6.1	Flowchart of the experimental procedure for obtaining an optimal baseline and number of features	77
6.2	Flowchart of the experimental procedure for dimension reduction on feature selection	78
6.3	Average accuracy of the ranked features for colon cancer data set	83
6.4	Average accuracy of the ranked features for the leukaemia data set	83
6.5	Average accuracy of the ranked features for prostate cancer data set	84
6.6	Average accuracy of the ranked features for lung cancer data set	84
6.7	Average accuracy of the ranked features for Parkinson data set	85
6.8	Average accuracy of the ranked features for breast cancer data set	85
6.9	Average accuracy of the ranked features for skin cancer data set	125
6.10	Average accuracy of the ranked features for the lymphoma data set	126
6.11	Average accuracy of the ranked features for lung cancer data set	126

6.12	Average accuracy of the ranked features for the handwriting data set	127
6.13	Confusion matrix for skin cancer data with 2-fold cross validation	134
6.14	Confusion matrix for lymphoma data set with 2-fold cross validation	138
6.15	Confusion matrix for lung cancer data set with 3-fold cross validation	143
6.16	Confusion matrix for handwriting data set with 10-fold cross validation	147
7.1	Z-score versus the features for colon cancer data set	151
7.2	Z-score versus the features for the leukaemia data set	154
7.3	Z-score versus the features for the prostate data set	156
7.4	Z-score versus the features for lung cancer data set	158
7.5	Z-score versus the features for Parkinson data set	160
7.6	Z-score versus the features for breast cancer data set	162
7.7	Z-score versus the features for skin cancer data set	164
7.8	Z-score versus the features for the lymphoma data set	166
7.9	Z-score versus the features for lung cancer data set	168
7.10	Z-score versus the features for the handwriting data set	170
7.11	Average accuracy of the predictive model and cross-validation versus the selected features by the rMI-SVM for the binary data set	177
7.12	Average accuracy of the predictive model and cross-validation versus the selected features by the rMI-SVM for the multiclass data set	180

LIST OF ABBREVIATIONS

Abbreviations

Acc	Average accuracy
AD	Adenocarcinoma
AML	Acute myeloid leukaemia
ANOVA	Analysis of variance
BLBC	Basal-like breast cancer
BMC	BioMed Central
C	Label class
CASP	Critical Assessment of Protein Structure Prediction
cDNA	Complementary DNA
CGL	Chronic granulocytic leukaemia
NB-CV	Cross validation of the naïve Bayes classifier
D	Data set
DCSF	Dynamic change of selected feature
DNA	deoxyribonucleic acid
$E(f_i)$	Expected number of times feature i
EST	Expressed sequence tag
$f_{\bar{c}}$	Average number of selected features
f_i	Frequency of the feature that selected
FPR	False positive rate
GAMIFS	Genetic algorithm mutual information feature selection
GEO	Gene Expression Omnibus
$H(X)$	Entropy of random variable X

$H(X,Y)$	Joint entropy of random variable X and Y
$H(X Y)$	Conditional entropy of random variable X and Y
$I(X;Y)$	Mutual information between random variable X and Y
ICI	Independent classification information
IEEE	Institute of Electrical and Electronics Engineers
JMI	Joint Mutual Information
JMIM	Joint Mutual Information Maximization
K	Number of replicates
k	Number of selected features
KNN	k -nearest neighbours classifier
LASSO	Least absolute shrinkage and selection operator
M	Number of instance
MBEGA	Markov blanket-embedded genetic algorithm
MI	Mutual information
MIFS	Mutual Information based Feature Selection
MIM	Mutual Information Maximization
MMR	Mismatch repair
MPM	Malignant pleural mesothelioma
MRI	Max-Relevance and Max-Independence
mRMR	Minimal-redundancy-maximal-relevance
mRNA	Messenger RNA
MSA	Multiple sequence alignment
N	Number of feature
n	Sample size
$N(0,1)$	Gaussian distribution

NB	Naïve- Bayes classifier
NCBI	National center for Biotechnology Information
NJMIM	Normalized Joint Mutual Information
NMIFS	Normalized Mutual Information Features Selection
NSGA-II	Non-dominated Sorting Genetic Algorithm II
$p(x)$	Probability mass function of the random variable X
$p(x, y)$	Joint distribution of random variables X and Y
PAM	Peptidylglycine alpha-amidating monooxygenase
PD	Parkinson disease
pmf	Probability mass functions
RAM	Random access memory
RBF	Radial basis function
RFE	Recursive Feature Elimination
rMI-SVM	Ranked Mutual Information with Support Vector Machine
RNA	Ribonucleic acid
ROC	Receiver operating characteristic
S	Subset of selected features
SVM	Support vector machine
TC	Tree classification classifier
TPR	True positive rate
X	Random variable
x_j	Candidate feature
Y	Random variable

CHAPTER 1

INTRODUCTION

1.1 Brief of Bioinformatics

The global population is now living in an era of information and technology, judging from the impact of computers and technology that has been around us. Undeniably, our lives are inseparable from information and networks. People are not only receiving information through the internet but also using online banking, online shopping, and online stock trading popularly called online transactions. Bioinformatics is a popular term and is an emerging discipline in modern science. Bioinformatics is a potent technology to search, process and to apply big data in biological science. As a methodology, bioinformatics is a comprehensive, data-driven, genome-wide and systematic approach that generates new hypotheses, discovers new patterns, and finds new functional elements. It helps to overcome the shortcomings in traditional experimental biology method. The potent combination of computational and experimental methods should be the most significant way in learning biological data (Gauthier et al., 2018).

Bioinformatics has a history of about 70 years when people are keen on developing molecular biology and computer science where many of the fundamental concepts and techniques were established in molecular biology

between 1950 and 1960. In 1944, Avery et al. extracted pure deoxyribonucleic acid (DNA) from a relay of preset of several features (Avery et al., 1944). In 1952, Hershey and Chase asserted that the DNA contained genetic information (Hershey and Chase, 1952). Subsequently, a year later, Watson and Crick (1953) pioneered the discovery of the double-helix structure of DNA (Watson and Crick, 1953). At the same time, the first protein sequence, insulin was solved and published by Sanger and Thompson (1953) (Sanger and Thompson, 1953). These findings further prompted the research of molecular biology to the golden age. Pauling and Zuckerkandl (1963) introduced evolutionary biology by analysing the protein sequences of haemoglobin (Pauling and Zuckerkandl, 1963).

Margaret Dayhoff (1962) was the first bioinformatician to use the computational methods in electrochemistry. She was the first person to apply the computation method in protein sequences evolution (Dayhoff and Ledley, 1962). Until today, the one-letter amino acid code proposed by Davhoff (1965) is still being used. Fitch (1970) defined the concept of ontology and stressed that human haemoglobin is more closely related to chimpanzee haemoglobin than rat haemoglobin. Maxam and Gilbert (1977) proposed the first DNA sequencing method in 1976 which was followed by a proposal by Sanger et al. (1977) on the sequencing of Sanger DNA (Maxam and Gilbert, 1977; Sanger et al., 1977; Fitch, 1970). This induced the developing of the DNA sequencing when prior to that the researchers studied more on protein sequencing. Furthermore, Dayhoff, Schwartz and Orcutt (1978) proposed the evolution of a protein model by using probability and constructing a famous peptidylglycine

alpha-amidating monooxygenase (PAM) amino acid substitution matrices to solve the problem in protein alignment.

Consequently, Staden (1979) developed the first software for analysing the Sanger DNA sequencing (Staden, 1979). In addition, Murata et al. (1985) developed the first multiple sequence alignment (MSA) by generalising the first dynamic programming algorithm developed by Needleman and Wunsch (1970) which is a more practical algorithm of MSA published by Feng and Doolittle (1987) (Feng and Doolittle, 1987; Murata et al., 1985). Later, a popular MSA software CLUSTAL was developed in 1988, which was used and maintained until today (Sievers and Higgins, 2014). In 1995, Craig Venter sequenced the first free-living organism genome, which was 1.83Mbp long (Fleischmann et al., 1995). In 2005, a total of 454 sequencing technologies were made available (Ouzounis and Valencia, 2003).

1.2 Development of Computational Bioinformatics

When life science was hastily developed in the last 70 years, computer technologies were also progressing and developing significantly. Nowadays, laptop, cell phones, tablets, smartwatch, and internet are deemed pervasive. A significant breakthrough in computer science was the information theory proposed by Shannon (1948). The development of bioinformatics as an interdisciplinary field of life science and computer science is an inevitable historical trend. The problem of the classic bioinformatics first started in 1960.

Sequence analysis has always been a core issue in bioinformatics. Information was concealed in the sequences of the function, evolution and regulation of genes and protein. Dayhoff and Ledley (1962) proposed a COMPROTEIN algorithm to solve the protein sequences by breaking down the proteins into small pieces, then rejoining them together to form a complete protein sequence. Subsequently, Dayhoff (1965) published the “Atlas of Protein Sequence and Structure”, which became the first database in bioinformatics (Dayhoff, 1965).

Needleman and Wunsch (1970) developed the first Dynamic Programme algorithm to find the optimal alignment (Needleman and Wunsch, 1970). Additionally, a practical technique for secondary structure prediction method was developed by Chou and Fasman (1974) (Chou and Fasman, 1974). Consequently, Michael and Arieh (1975) developed the complex chemical systems of protein folding and was awarded the Nobel Prize in Chemistry in 2013 (Michael and Arieh, 1975). Similarly, in 1994 John Moult and his team started the Critical Assessment of Protein Structure Prediction (CASP) to provide an efficient prediction method (John, 2005). In 1990, the expressed sequence tag (EST) sequencing and microarray technology was started and became widely used. Numerous algorithms and techniques were created for gene expression data analysis. Gradually, genome sequencing, genome alignment, and gene prediction became popular in the bioinformatics research area. Currently, the research area of bioinformatics Dayhoffs sequence alignment and assembly, gene expression and identification, protein prediction which is a tool of evolutionary development (Dayhoffs et al., 1978).

Biology techniques have always generated large amount of data with noise. Biological data are complex and highly dimensional. Similar to data mining, bioinformatics also used mathematical tools to extract useful biological information from a large amount of data. The human organism is a complex system, and contains cells with a nucleus comprising the chromosomes.

1.3 Research in Cancer Biomarker

DNA microarray, also known as DNA array, are commonly referred to as gene chips. The core principle behind a microarray is the hybridisation between two DNA strands, the property itself matches a nucleic acid with each other by creating the hydrogen bonds between paired nucleotide base pairs. It is an instrument for research in genomics and genetics. Researchers can collect a large number of gene expression at the same time, and it is a fast, precise and a cheaper bioanalytical testing skill. It can provide comprehensive information that are related to the genetic sequence. The emergence of microarray experiments have produced several bioinformatics challenges such as experimental design, pre-processing of the data, significance analysis, cluster analysis and predictive analysis, relevant analysis and experimental validation. The primary purpose of predictive analysis or classification method is to use gene expression data to build a classification model to predict the existence of diseases. This includes how to select the essential features from a large number of genes, and then build the predictive model. The aim of this analysis is to recognise genes that may be affected by the disease as a biomarker for

early diagnosis and to successfully establish the diagnostic models (Le et al., 2015).

A biomarker is a quantifiable indicator of some physiological state or disease condition. In genetics, a biomarker is a DNA sequence that causes the disease or is related to a predisposition to the disease. Microarray data analysis has become a popular research area. One of the applications of DNA microarray is gene expression profiling. A cancer biomarker refers to a substance which was formed directly by a tumour cell or induced by a non-tumour cell. The detection of tumour markers can gauge the existence of a tumour or pathogenesis and prognosis of tumours. Cancer biomarker can help in diagnosing the presence of a tumour and the type of tumour. At the same time, it can help in the prognosis of the tumour whether it is benign or malignant and which type of treatment is suitable. Also, it helps to predict whether the tumour will attack again or not.

Researchers at the University of Texas MD Anderson Cancer Centre have confirmed that a protein called CSN6 relates to low survival rates in patients with colorectal cancer. This discovery may have significant inference on alternative treatment for colorectal cancer (Aranda et al., 2015). A researcher found that a molecule called IL13RA2 gives a higher expression in the final stage of Basal-like breast cancer (BLBC); therefore, IL13RA2 can be used to predict the progression-free survival of BLBC. At the same time, a high level of IL13RA2 also indicated that BLBC can spread quickly to the

lungs (Papageorgis et al., 2015). Furthermore, researchers from the Johns Hopkins Kimmel Cancer Centre found that mutations in the mismatch repair (MMR) gene may help to predict the reaction of the inhibitor drug PD-1 in patient accurately. This study transmits a sign to tell how chemotherapy works in cancer and based on the genetic characteristics of cancer it can give guidance on chemotherapy. The use of this predictive biomarker can help researchers to prescribe correctly the drug to the responded patient to avoid expensive treatment and increase the treatment time to heal patients (Ranjan et al., 2015).

Researchers from the University of Sheffield in the United Kingdom have identified 788 biomarkers from the blood test, and these biomarkers were used to develop early cancer screening test (Uttley et al., 2016; University of Sheffield, 2016). Researchers have developed a new technique for detecting pancreatic cancer biomarkers in the serum of cancer patients (Sogawa et al., 2016). Biomarkers have great significance in the research and development of life science, as well as medical diagnosis, clinical treatment and new drug development. It helps researchers to be more effective in diagnosis or treatment, especially in the prevention and control of complex and chronic diseases such as cancer, cardiovascular disease, diabetes, etc. For example, a new combination of four proteins (APOE, ITIH3, APOA1 and APOL1) form an accurate biomarker diagnosis of pancreatic cancer as early diagnosis and early treatment may improve prognosis and increase the chances of survival (Liu et al., 2017).

1.4 Motivation

In 2011, the population of Peninsular Malaysia was 29.1 million, which comprised of 50.7%, (male) and 49.3%, (female). According to the distribution of ethnicity, Malay was 54.7%, Chinese was 24.3%, Indian was 7.3%, Bumiputera was 12.8%, and others was 0.9% (Department of Statistic Malaysia). Between 2007 and 2011, there were 103507 new cancer cases diagnosed in Malaysia made up 45.2% of male and 54.8% female. Table 1.1 shows the breakdown of percentage based on gender and race. On average, every year, about 20000 new cancer cases were diagnosed out of which 55 people were diagnosed every day with cancer. From the population view, one out of 1455 people was diagnosed with cancer. In Malaysia, the top three most common type of cancer was breast cancer (32.1%), colorectal cancer (27.1%) and lung cancer (21.4%). Table 1.2 shows the breakdown of the percentage of the top 10 most common cancer among male and female.

Table 1.1: Percentage of incidence by gender and ethnic

Race	Male	Female
Malay	18.1%	23.4%
Chinese	19.7%	21.4%
Indian	2.6%	4.6%
Others	4.8%	5.4%

Table 1.2: Percentage of the top 10 most common cancer among male and female

Male		Female	
Colorectal	16.4%	Breast	32.1%
Lung	15.8%	Colorectal	10.7%
Nasopharynx	8.1%	Cervix Uteri	7.7%
Lymphoma	6.8%	Ovary	6.1%
Prostate	6.7%	Lung	5.6%
Liver	6.5%	Lymphoma	3.9%
Leukaemia	5.4%	Corpus Uteri	3.8%
Stomach	4.3%	Leukaemia	3.5%
Bladder	3.2%	Thyroid	3.0%
Other Skin	3.0%	Stomach	2.6%

In the absence of other causes of death, it is very likely that one out of ten males will develop cancer before the age of 75 and one out of nine females, though females are at slightly higher risk than males. Among the race, Chinese has a higher risk than others and among the genders, the female has a higher risk than the male. Studies found that the diagnosis of cancer usually occurs after the age of 30 for male and female. However, the incidence of the male is higher than female after the age of 60. Breast cancer is the most common cancer in Malaysia. The incidence percentage is the highest among the Chinese, followed by Indians and Malays. Most of the cases were diagnosed between 45 years old to 69 years old which decreases after the age of 70. Overall, Chinese female has a higher risk than other races. Breast cancer is divided into four stages, and the first and second stages belong to early-stage while the third and fourth stages belong to the late stage. Among 11938 cases, 20% diagnosed was at the early stage, 37% diagnosed was at the second stage,

23% diagnosed was at the third stage and 20% diagnosed was at the fourth stage, and 43% of the incidence diagnosed at the late stage.

Colorectal cancer is the most common cancer among the male, and second among the female. Studies found that the diagnosis of colorectal cancer showed an increase after 60 years old in male and female. Overall, Chinese has a higher risk than other races and most of the colorectal cancer diagnosed was at the late stage (66% in male and 65% in female). Lung cancer was the second common cancer among male and fifth among female. Studies found that the diagnosis of lung cancer started to increase after 60 years old in male. Again the Chinese have a higher risk than other races and an extremely high percentage of the incidences diagnosed were at the late stage (89% in male and 91% in female).

From the study, most cancer cases were diagnosed at the late stage. Although cancer cells are not naturally found in the human body, nevertheless, there are cancer cells in each of us, but not everyone will develop malignant tumours. Many small tumours not easily detected during physical examination. If they missed the first physical examination, the cancer cells would start growing. It may take just a few weeks or as slow as years or decades to grow. Secondly, not all physical examinations can detect cancer because the standard physical tests usually check the blood, urine and physical function and do not deliberately check the cancer biomarker. Many early cancers do not affect the blood, urine or physical function; therefore, it becomes difficult to detect cancer in the first stage. Many cancers have no symptoms or signs in the early

stages, such as liver cancer and stomach cancer, which are often considered as typical stomach diseases. Bleeding caused by ovarian cancer may be treated as irregular menstruation. Some people are unaware of the high risk of cancer since some cancers are genetically inherited. (National Cancer Registry Department. Summary of Malaysian National Cancer registry report 2007-2011)

In 2016, cancer was the fourth biggest killer in Malaysia. The highest mortality rate of cancer is lung cancer (14.4%), followed by breast cancer (9.21%) and colorectal cancer (7.3%). The main reason for the increasing mortality rate of cancer patient is because of neglecting to seek treatment at the beginning of the disease. In Malaysia, 60% of the cancer cases were discovered at the final stage, and the main factor is people lack awareness about cancer screening. Even some patients tend to seek alternative treatment rather than treatment at a hospital. In the end, most of the patients only return to the hospital for treatment when the cancer cells have spread to a severe condition thus reducing the chances of full recovery. Early diagnosis and early treatment may improve prognosis. Biomarkers play an essential role in the early detection of cancer. However, finding a biomarker from a microarray data is not an easy task as microarray datasets are commonly high dimensional with a low number of instance. Therefore, traditional statistical analysis is not suitable to deal with the curse of dimensionality.

1.5 Problem Statement

High dimensional datasets commonly come with noise, irrelevant features and redundant features, which will decrease the power of classifying between the tumours and non-tumours. Microarray data are always in high dimensionality with a small sample size which is the main challenge in microarray data analysis. A variety of statistical method and machine learning tools are used in the classification task. The microarray data analysis can be classified into supervised learning and unsupervised learning (Vinh and Bailey, 2013). The simplest method for detecting differential gene expression analysis is the t -test. When the p -value exceeds the selected criteria according to the confidence level, the features are then considered different. However, the t -test is often limited by the sample size. Small sample size leads to an unreliable estimation in the predictive model and the same in the analysis of variance (ANOVA).

Since “noise” always exists in the microarray data and the assumption of the normal distribution, a statistical method such as t -test, the regression model may not be suitable or is risky to use. The non-parametric test does not require the data to satisfy the assumption of different distribution. However, the shortcoming of the non-parametric method is that the analysis has a hypothesis test. For example, changing the variation in the sample can affect the analysis result. Therefore, to extract the information from high dimensional data in the classification task has become the most challenging problem in bioinformatics. Current research shows that there are numerous

feature selection methods in reducing the dimension of high dimensional data. However, most of the feature selection methods had produced a different subset of selected features when the training set chooses a different sample except for one unsettling question. What is the optimal baseline of a build training model, since the optimal baseline usually depends on the amount of information contained in the data set.

One of the most controversial issues in feature selection is how many genes are needed to interpret the predictive model? As the smallest number of genes is preferable when going through the clinical test or validation test in the lab (Elyasigomari et al., 2017). Currently, research are usually not involved as to how many features are needed in building a predictive model. It only shows that the more relevant features are added to the built predictive model, the better is the accuracy of the prediction. Therefore, when more features are added into the predictive model, from the past research, not every newly added feature will provide new information to the predictive model. From the previous study, the accuracy of the predictive model was frustrated when more features were added to the predictive model, causing this to be very unexpected. If the newly added feature cannot improve on the predictive model, then the meaning of choosing this new feature is lost.

1.6 Objectives and Significance of the Study

The objectives of this research are 1) to develop a robust algorithm to handle the high dimension with low instance data set in the classification task,

2) to use the ranked features to obtain an optimal baseline with minimal features, instead of using all the features, 3) to obtain an algorithm which strictly increases the information content per feature added and 4) do not rely on the preset number of features needed in building the predictive model. The idea from information theory is adopted to capture the most relevant subset of features by ranking the features according to the mutual information score. On the contrary, the support vector machine (SVM) classifier plays a vital role to reduce the dimensionality of the data set. The ranked features also give an optimal baseline for the predictive model which is better than using all the features. The number of ranked features, k needed to obtain this optimal baseline plays an essential role in feature selection where any of the feature selection method should not use more than k features to achieve this optimal baseline. The k features serve as the unique cutoff number of features to obtain the optimal baseline. This is a significant breakthrough because the past research has no idea how many features are needed to achieve the baseline as the current baseline is obtained using all the features.

In addition, the proposed algorithm will obtain a smaller number of features to build the predictive model, and the proposed algorithm is able to show that the selected features give a better performance compared to the model when using all features (benchmark). The number of features selected by the proposed algorithm also provides the same or better performance compared to the new proposed baseline using the ranked features. At the same time, the proposed algorithm will guarantee that the newly added feature will provide new information to the predictive model and obtain a strictly

increasing accuracy graph versus the number of features which alleviates the problem of overfitting. The features selected by the proposed algorithm will guarantee that the predictive model will only get better and better when more features are added to the predictive model.

The robustness of the model will measure by the Z-score, showing that the information theory is able to retain the maximum information in the predictive model during the feature selection process. Furthermore, the receiver operating characteristic curve (ROC curve) plots to show the sensitivity of the predictive model. The predictive model is compared to the regression model and Minimal-redundancy-maximal-relevance criteria by using different classifiers such as SVM, k - nearest neighbour, naïve Bayes and decision tree.

Chapter 2

REVIEW ON FEATURE SELECTION

2.1 Machine Learning

Machine Learning (ML) is part of the development of artificial intelligence. It allows the machine to learn independently and enhance the algorithm. As more data is input, the algorithm will continue to modify and make the prediction more accurate. For example, when you click on YouTube, the website will predict the type of video based on history and display it in “Recommended Videos”. Additionally, Machine Learning is classified into supervised and unsupervised learning. In supervised learning, a data set with the label class where the relationship between the independent variables and the dependent variables are known. Unsupervised learning is to let the algorithm find the patterns from a huge number of data and classify the data, customarily is called clustering. The application of machine learning has appeared in our daily life for a long time. For example, handwriting recognition on mobile phones, automatic filtering of junk mail in the e-mail, automatic driving of crewless vehicles.

One of the developments of machine learning – “AlphaGo” which trains its algorithm through a large number of professional chess games, making AlphaGo the professional chess within a period of two years. Imagine if two AlphaGo competes with each other and further strengthens their

algorithms during the chess game, it will be developed to the extent that the human brain cannot reach it. Machine learning is a process of processing data and building the training model. In building the training model, there is a set of process in the training strategy by taking into account the training model in terms of correlation, dependency, and relevancy (Hu et al., 2018). A good predictive model is related to the feature selection method and the algorithm. Cai et al. (2015) proposed an ensemble-based feature selection method which combines receiving the operating curve, information theory, classifying followed by machine learning to categorise three different types of lung cancer. Also, it will show that the proposed method, combined with incremental feature selection gives a better result with only 16 selected features (Cai et al., 2015).

2.2 Dimensional Reduction

2.2.1 Feature Extraction

In the field of machine learning and statistics, dimension reduction refers to the process of reducing the number of attributes to obtain a set of essential variables. Dimensionality reduction is classified into two groups: feature selection and feature extraction. Feature selection selects the most informative features from the original features to reduce the dimension of the data set and to improve the performance of the learning algorithm. Feature extraction is the process of transforming a high dimensional raw data into low dimensional so that machine learning can learn it; but, the transformed data irreversible. Feature selection and feature extraction have some similarities.

Though they have the same objective that is trying to reduce the number of attributes in the data set, but the way they are being used is different. The feature extraction is mainly through the relationship between the attributes, such as combining different attributes to obtain a new attribute. Thus, changing the original feature dimension whereas the feature selection is to select a subset from the original feature data set without changing the original feature space where the subset well represents the data set (Ang et al., 2016).

Principal component analysis, singular value decomposition and deep learning belong to the feature extraction method. The principal component analysis is a statistical method where a set of variables that may be related to each other transforms into a set of linearly uncorrelated variables by orthogonal transformation. The transformed set of variables is called the principal component. Singular value decomposition is a vital matrix decomposition in linear algebra. It is the generalisation of feature decomposition on a matrix. The concept of deep learning comes from the study of artificial neural networks. The deep learning structure contains multiple hidden layers. It creates a high-level theoretical representation of attributes by combining low-level features to find the distribution of the data. However, feature extraction does not fully explain the model in detail as the original data set. For example, in principal component analysis, the principal component with a small contribution rate may often contain some vital information.

Also, feature extraction such as deep learning with the complex structure, the ability of the neural network to fit the model is becoming better, but this often leads to over-fitting which creates a severe problem in machine learning. It means that the performance of the training data is outstanding, but the prediction ability is feeble. The purpose of building a predictive model is to learn the structure and nature of the original data set to solve some real-life problem. At the same time, the selected features should be able to explain the problem better. Therefore, the feature selection should be building a faster and lower- predictive cost model, improving the accuracy of the prediction, and having a better understanding and explanation of the model. (Pedrycz and Chen, 2020)

2.2.2 Feature Selection

Feature selection is classified into three groups, filter method, wrapper method and embedded method (Kohavi and John, 1997). The main idea of the filter method is to rank the features according to the importance of the features using measurements such as Pearson correlation coefficient, rank correlation coefficient, Chi-square test and mutual information (Li F, 2017; Herman et al., 2013; Guyon, 2004). Pearson correlation coefficient is used to measure the linear correlation coefficient between two variables. The higher the absolute value of the correlation coefficient, the stronger is the correlation. The rank correlation coefficient is a statistical analysis indicator that shows the degree of relevance of the variables. The conventionally used rank correlation analysis methods are Spearman rank and Kendall rank. Chi-square test is a

widely used hypothesis test method. It is applied in the statistical inference for data classification. Mutual information is used to evaluate the amount of information contributed by the existence of one event to the existence of another event (Nguyen et al., 2014).

The wrapper method searches the selection of a subset as an optimisation issue by generalising many different combinations of subset and evaluating one by one to obtain the best subset (Aksakalli and Malekipirbazari, 2016). The typical optimisation method is a recursive feature elimination. The recursive feature elimination method uses a machine learning model to train the data set, so that features that correspond to the weight coefficient will be eliminated. This process continues until it has obtained an optimised subset of features; usually, the number of features have to decide before running the algorithm. Guyon et al., (2002) proposed a gene selection on microarray data using Support Vector Machine based on Recursive Feature Elimination (RFE) which requires pre-processing before training the classifier. The embedded method studies the best attributes to improve the accuracy of the predictive model when setting the model. The embedded method integrates the feature selection process with the model training process, and both processes are complete in an optimisation process. The embedded method will select the essential attributes to train the predictive model.

The example of embedded methods is the least absolute shrinkage and selection operator (LASSO) (Kukreja et al., 2006). It is easy to overfit by

using only the ordinary least square method because the noise is over-focused and the slightest difference in the training data may cause a significant difference in the model. The LASSO uses a cost function with a regular term can overcome the overfitting problem. There are always limited features under LASSO, but the parameter corresponding to these features is zero, therefore during the feature selection, those zero coefficient will be removed from the data set. However, LASSO depends on the parameter of lambda, where the enormous lambda value, the more sparse features and consequently, fewer features will be selected (Kamkar et al., 2015).

2.3 Evolution of Feature Selection on Information Theory

Information theory is a widely used theory in feature selection. The feature selection is an essential part of building a predictive model. How many essential features are useful for the classification should be selected from a thousand to a few ten thousand features is a challenging question in microarray data analysis. For the past 30 years, information theory was mostly applied using the filter method since it can evaluate both linear and nonlinear dependencies among the features. Mutual information (MI), conditional mutual information and joint mutual information are information measures used to measure the relevance and redundancy between the features and the label class (Novovičová et al., 2007; Cheng et al., 2011; Peng and Fan, 2017). For over 20 years, feature selection had used mutual information.

According to Lewis, (1992), Mutual Information Maximization (MIM) was used in feature selection on text categorisation, and the features ranked according to the expected mutual information and selection of the different size of features had been investigated. In addition, Lewis found that the optimal feature set size are between 10 to 15 features as the number of selected features is relatively low during that time (Lewis, 1992). Subsequently, Battiti (1994) proposed a new algorithm using mutual information which was based on the greedy selection named Mutual Information based Feature Selection (MIFS). MIFS shows that mutual information is capable of measuring the dependence between variables regardless of the relation between the variables, either linear or nonlinear. Also, MIFS considers the interaction between features and label class and among the features. The number of features selected were based on the optimisation of the greedy selection. The mutual information between the feature and the class by Battiti was defined as $I(X:C) = 1 + \alpha \log\left(\frac{2\alpha-1}{2\alpha}\right) - \frac{1}{2}\log(2\alpha - 1)$ where X represents the feature, and C is the label class, α is a parameter that was calculated using the fisher factor. Battiti showed that when the α is between $\frac{1}{2}$ and $\frac{1}{\sqrt{2}}$, the more informative features will be selected. Battiti is the pioneer who started using mutual information in feature selection (Battiti, 1994).

Joint Mutual Information (JMI) Yang and Moody, (1999) highlighted that Joint Mutual Information (JMI) considered the joint mutual information between the features and the selected features with the class. The joint mutual

information was defined as $I(X_i, X_j; Y) = I(X_i; Y) + I(X_j; Y|X_i)$ whereby this algorithm has to maximise the $I(X_i; Y)$ which represented the mutual information between feature and label class. Similarly, $I(X_j; Y|X_i)$ was maximised which represented the conditional mutual information between selected feature with the label class and the already selected features. The conditional mutual information overcomes the limitation on feature selection when there exists a group of features with the same mutual information. Therefore JMI will select feature X_2 from $\{X_2, X_3\}$ with the same mutual information when $I(X_2; Y|X_1) > I(X_3; Y|X_1)$. In general, JMI uses conditional mutual information to filter redundant features (Yang and Moody, 1999).

MIFS-U is an evolution of the MIFS. Kwak and Choi (2002) proposed two feature selection algorithms using mutual information method and Taguchi method. These two methods can complement each other to overcome any drawback. MIFS only consider the relationship between the features and the class in the greedy selection algorithm. Furthermore, Kwak and Choi (2002) considered the relationship between the selected features with the new candidate with the label class. Therefore, Kwak and Choi tried to choose a new candidate to maximise the $I(C; f_i) - \beta \sum_{f_s} \frac{I(C; f_s)}{H(f_s)} I(f_i; f_s)$, which is the new information that inputs into the selected features. Evidently, $I(C; f_i)$ was the mutual information between the label class C and the selected feature and β is a parameter. f_s represented the already selected features while $H(f_s)$ represented the entropy of the already selected features, $I(C; f_s)$ and $I(f_i; f_s)$ represented the mutual information between the label class and already

selected features and candidate feature with the previously selected features respectively. However, the algorithm used by Kwak and Choi sets the parameter $\beta = 1$ in $I(C; f_i) - \beta \sum_{f_s} \frac{I(C; f_s)}{H(f_s)} I(f_i; f_s)$. MIFS-U overcomes the limitation in MISF and shows that MISF will not perform well in nonlinear cases (Kwak and Choi, 2002).

Peng et al. (2005) proposed a feature selection method based on the criteria of maximising the relevance between feature and class and, at the same time, minimising the redundancy between the feature called Minimal-Redundancy-Maximal-Relevance” (mRMR). In addition, Peng and his team members use mRMR together with the wrapper method to obtain a more compact subset to illustrate that their approach performs better compared to the popular maximum dependency method. They selected the compact features using a two-stage algorithm where the first step is to choose a subset of features using the mRMR criteria. The mRMR incremental algorithm optimises the $\max_{x_j \in X - S_{m-1}} \left[I(x_j; c) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_j; x_i) \right]$. X represents the full features while the S_{m-1} represents the features set with $m - 1$ features (Peng et al., 2005).

The objective of mRMR is to select the m features that maximise the maximal relevancy- minimum redundancy. $I(x_j; c)$ represents the mutual information between the candidate feature and the label class while $I(x_j; x_i)$ represents the mutual information between the two features. During the second

stage, the wrapper method was used to select the compact features based on the minimum classification error from the cross-validation classifier. Subsequently, Peng et al. (2005) proposed two search algorithms during the second stage which comprised the backward selection and forward selection. It showed that the mRMR primary approach was maximising the dependency and showed a practical approach to high dimensional data. Although their selected features might not be independent and were highly correlated to each other but the selected features were maximising the relevancy and minimising the redundancy to yield better performance. They stated that though all the selected features are independent but are usually less optimised. However, they also indicated that mRMR might choose a high relevancy feature with high redundancy at the same time because of the selected feature based on the difference of relevance and redundancy.

Estevez et al. (2009) defined Normalised Mutual Information Feature Selection (NMIFS) as an enhanced edition of MIFS, MIFS-U and mRMR. NMIFS does not depend on any parameter like MIFS, MIFS-U and mRMR, and in practice there is no clear guidance on how to select the value of the parameter. The NMIFS proposed to use average normalised mutual information to measure the redundancy between the features (Estevez et al., 2009). Later, the genetic algorithm combines with NMIFS to become a genetic algorithm mutual information feature selection (GAMIFS).

GAMIFS measures the dependency among the features, and its team members define the normalised mutual information as $NI(f_i; f_s) = \frac{I(f_i; f_s)}{\min\{H(f_i), H(f_s)\}}$ and the average normalised mutual information as $\frac{1}{|S|} \sum_{f_s \in S} NI(f_i; f_s)$. $I(f_i; f_s)$ represents the mutual information between the features, whereas $H(f_i)$ represents the entropy of the feature and S is the sum of the features. NIMIFS selected a feature by maximising the $G = I(C; f_i) - \frac{1}{|S|} \sum_{f_s \in S} NI(f_i; f_s)$ while $I(C; f_i)$ represents the mutual information between the label class and the feature. The difference between NMIFS and JMI is that NMIFS only consider the mutual information between features and the features with the label class upon removing the redundant features.

Hoque et al. (2014) stressed that MIFS-ND is also another evolution of the MIFS. MIFS-ND selected the features based on an optimisation algorithm known as Non-dominated Sorting Genetic Algorithm II (NSGA-II). MIFS-ND considered the relation between features and features and class with mutual information. The mutual information between feature usage is to find out the redundant features while the mutual information between features and class usage is for finding the relevant features. NSGA-II will select the feature with the higher mutual information between feature and class when there is a tie in the condition where the selected features are of the same value (Hoque et al., 2014). The difference between the MIFS and MIFS-ND is that MIFS algorithm depends on the α while MIFS-ND does not rely on any parameter. The difference between mRMR and MISF-ND is mRMR which uses a two-

stage features selection that involved filter method and wrapper method, while MISF-ND only uses the filter method in feature selection.

Bennasar, Hicks, and Setchi (2015) proposed the Joint Mutual Information Maximisation (JMIM) and Normalised Joint Mutual Information (NJMIM). Both the feature selection methods are based on the criteria maximising the relevance and minimising the redundancy, and these methods have demonstrated the ability to solving the overestimates problem. JMIM selected the features based on the $\max_{f_i \in F-S} \left(\min_{f_s \in S} I(f_i, f_s; C) \right)$ where F is the initial data set, S is the already selected subset of features. The joint mutual information $I(f_i, f_s; C)$ defined as $I(f_i, f_s; C) = \left[-\sum_{c \in C} p(c) \log(p(c)) \right] - \left[\sum_{c \in C} \sum_{f_i \in F-S} \sum_{f_s \in S} \log \left(\frac{p(f_i f_s, c / f_s)}{p(f_i / f_s) p(c / f_s)} \right) \right]$ where C is the label class, f_i is the candidate feature and f_s is the already selected subset of features. The NJMIM used the normalised mutual information instead of mutual information to select the feature based on $\max_{f_i \in F-S} \left(\min_{f_s \in S} \left(\frac{I(f_i, f_s; C)}{H(f_i, f_s, C)} \right) \right)$. The JMIM showed a better performance in discrete data compared to NJMIM (Bennasar et al., 2015).

Wang et al. (2017) pointed out that the Max-Relevance and Max-Independence (MRI) has relevancy between the features and classification of independent features. Furthermore, Wang et al. (2017) proposed a new information term “independent classification information” (ICI) disclosing that the selected candidate feature will provide substantial new information

and less redundancy to the predictive model. The selection of the candidate feature is based on the higher MRI score where the score of MRI is defined as $J_{MRI}(x_k) = I(y; x_k) + \sum_{x_j \in S} ICI(y; x_j, x_k)$, $I(y; x_k)$ which represents the mutual information between the label class y and candidate feature, while S represents the already selected subset of features. $\sum_{x_j \in S} ICI(y; x_j, x_k)$ became the loose upper bound of the mutual information between label class and the already selected subset of features with the candidate feature defined as $\sum_{x_j \in S} ICI(y; x_j, x_k) = \sum_{x_j \in S} [I(y; x_j|x_k) + I(y; x_k|x_j)]$ where $I(y; x_j|x_k) = I(y; x_k) - I(y; x_j; x_k)$. $I(y; x_k)$ represents the mutual information between the label class and candidate feature while $I(y; x_j; x_k)$ represents the multi-information between the label class, that have already selected the subset of features and the candidate feature (Wang et al., 2017).

Gao, Hu and Zhang (2018) proposed a dynamic change between the features and the label called dynamic change of selected feature (DCSF). The next candidate feature will be chosen when it is dependent on the already selected subset of features. Therefore, DCSF will choose a candidate feature based on the maximum value of $J(x_k) = I(x_k; y) - \frac{3}{|S|} \sum_{x_j \in S} I(x_j; x_k) + \frac{2}{|S|} \sum_{x_j \in S} I(x_j; x_k|y)$ where $I(x_k; y)$ represents the mutual information between the candidate feature and the label class, S represents the already selected subset of features and $I(x_j; x_k|y)$ represents the conditional mutual information of the already selected subset of features with the candidate feature given the label class. Gao and his team members showed that the

previous methods such as MIM, MIFS, mRMR, JMI and MRI did not take into consideration the dynamic change in building the predictive model. However, the current DCSF will only consider the feature selection in linear but does not include nonlinear yet (Gao et al., 2018).

Elyasigomari et al. (2017) applied a two-stage gene selection which combines the mRMR criteria in searching the essential genes in stage one, and use the SVM with the new algorithm cuckoo optimisation algorithm for stage two. The proposed method shows that the selected features are biologically relevant to cancer diseases. This research also highlights that only a few features are relevant in the microarray data analysis (Elyasigomari et al., 2017). Li, Xie and Liu (2018) proposed an efficient feature selection method to overcome the drawback of the SVMRFE proposed by Guyon in the year 2002. The proposed method improves the SVMRFE with the variable step size to minimise the computation time. Besides this, the proposed method also use the resampling method to overcome the common problem of small sample size and class imbalance in the microarray data set (Li et al., 2018).

The graph theory, Fisher scores, modified Ant Colony Optimization with local search algorithm were proposed by Bir-Jmel et al. (2019). This proposed method focuses on reducing the high dimensionality of the microarray data. For more details on features selection on microarray dataset and mutual information methods, review articles are recommended. (Bolón-Canedo et al., 2013; Vergara and Estév, 2014)

2.4 Common Problems and Limitation in the Past Research

Based on past researches, feature selection is a very challenging task in bioinformatics because of the high dimensionality of the data set combined with a low number of instances. Feature selection always raises an issue on how to select the most relevant features with minimum redundancy from hundreds to tens of thousands of features. The relevant features will improve the performance of the classification model, while the redundancy features will decrease the performance of the classification model. In general, the problem that occurs during the feature selection can be classified into three domains: 1) how to select the most relevant features with minimum redundancy, 2) how to maximise the new information that is added into the predictive model, 3) how many minimum features are needed in a predictive model, 4) which is the better baseline other than using all the features and 5) what is the relationship among the selected features.

Several past researchers had tried to solve the first problem by maximising the mutual information between the feature and class such as MIM, MIFS, MIFS-ND, mRMR and GAMIFS. However, while selecting the high relevance features, the redundant features might be chosen at the same time (Pascoal et al., 2017). The methods mentioned earlier did not consider the maximisation of the increment of the information content, which was the relationship between the candidate feature with the already selected subset of features and the label class. MIFS-U, JMI, MRI, DCSF used conditional mutual information and joint mutual information to solve the second problem.

They considered the relationship between the candidate feature with the already selected subset of features and the label class to maximise the increment of the information contained in the model.

The existing methods failed to carry out an in-depth study on the minimum features needed in a predictive model. Until today, the number of features needed to build a predictive model is still a hot topic in most of the learning machine task. Besides this, the baseline of the data set often involved all the features and the researchers are always comparing their findings with previous results. Usually, the baseline of the microarray data depends on the information contained in the data set, and this varies among the different data set. The information content can only be increased when more real data or sample are added. The baseline is obtained by using all the features which are often far worse because it also contains irrelevant features and noise, so there is a need to find a way to get a better baseline.

Most of the feature selection methods are concentrated in selecting the most informative features to build a predictive model. Not much study has been done on the biological meaning of the selected features in the microarray data analysis. It is an essential step to ensure the chosen features are relevant to the diseases. However, more time is needed for the features to undergo an experiment in the wet lab to confirm that the selected features are in fact related to the disease.

CHAPTER 3

RANKED MUTUAL INFORMATION

3.1 Information Theory

3.1.1 Entropy

Along with the development of computer science and technology, researchers are aware of the value of the information in biological science. There was a vast information hype in microarray data, and there is an urgent need to extract and find useful information. Information theory is applied mathematics that adopts the probability theory and mathematical statistics to study the information, data compression, cryptography, and data transmission. Information theory considers the transmission of information as a statistical occurrence and offers a method to estimate the communication channel volume. Information transmission and information compression are two major research areas of information theory. The beginning of the information theory research was a paper entitled “A Mathematical Theory of Communication” published by Shannon (1948). Because of this new finding, Shannon is known as the “father of information theory”. Shannon defined the information entropy of X as $H(X) = -\sum_x p(x) \log p(x)$ where $p(x)$ is a probability mass function of the random variable X (Shannon, 19448).

This definition is used to estimate the volume of bandwidth required to pass a binary encoded original information. The base of the logarithm was two

as it is the smallest unit in information which is “bit”. The base of the logarithm can change to other value. For example, when the base of the logarithm was natural logarithm, and the unit was “nat”. In this research, the base of the logarithm has two uses in the calculation. Information theory applied the probability to describe the uncertainty and entropy was used to measure the uncertainty. The probability of occurrence of a random variable is small, and has more uncertainty, as the amount of information is vast and vice versa.

From a statistical point of view, the small probability of an event brings much information. Therefore, the lower the probability of an event, the higher the amount of information. That is, the amount of information is inversely proportional to the frequency of occurrence of the events, (Cover and Thomas, 2006). The range of the entropy was from zero to $\log(n)$, where n is the number of outcomes. The minimum value of entropy was zero when only one probability was one and the others were 0's while the maximum value of the entropy occurs when all the probabilities equal to $\frac{1}{n}$.

3.1.2 Conditional Entropy and Joint Entropy

Conditional entropy, $H(X|Y)$ indicates the uncertainty of the random variable Y on condition that random variable X is known. It means that the more information is known, the less uncertain the random event will be. The mathematical expectation of the $H(Y|X)$ is defined as follows:

$$H(Y|X) = \sum_x p(x)H(Y|X = x)$$

$$= - \sum_x p(x) \sum_y p(y|x) \log p(y|x)$$

$$= - \sum_x \sum_y p(x, y) \log p(y|x)$$

$$= - \sum_{xy} p(x, y) \log p(y|x)$$

where $p(x, y)$ is the joint distribution of random variables X and Y .

Conditional entropy $H(Y|X)$ is equivalent to the difference in the joint entropy $H(X, Y)$ and the individual entropy $H(X)$.

$$H(X, Y) = - \sum_{xy} p(x, y) \log p(x, y)$$

$$= - \sum_{xy} p(x, y) \log [p(y|x)p(x)]$$

$$= - \sum_{xy} p(x, y) \log p(y|x) - \sum_{xy} p(x, y) \log p(x)$$

$$= H(Y|X) - \sum_{xy} p(x, y) \log p(x)$$

$$= H(Y|X) - \sum_x \sum_y p(x, y) \log p(x)$$

$$= H(Y|X) - \sum_x \log p(x) \sum_y p(x, y)$$

$$\begin{aligned}
&= H(Y|X) - \sum_x [\log p(x)] p(x) \\
&= H(Y|X) - \sum_x p(x) [\log p(x)] \\
&= H(Y|X) + H(X)
\end{aligned}$$

The joint entropy was $H(X, Y) = H(Y|X) + H(X)$ and the conditional entropy was $H(Y|X) = H(X, Y) - H(X)$. The amount of information added by the two events was higher than a single event. When $H(X)$ was known, the amount of information left by $H(X, Y)$ will be the conditional entropy. The conditional entropy $H(Y|X) \neq H(X|Y)$ but $H(X) - H(X|Y) = H(Y) - H(Y|X)$. For random variables X, Y and Z , $H(X, Y|Z) = H(X|Z) + H(Y|X, Z)$.

3.1.3 Mutual Information

Mutual information measures the amount of information contributed by the presence of one event to the occurrence of another event. The mutual information of two discrete random variables X and Y is defined as $I(X; Y) = \sum_y \sum_x p(x, y) \log \left[\frac{p(x, y)}{p(x)p(y)} \right]$. Intuitively, mutual information measured the share information between random variables X and Y . For example, if the random variables X and Y were independent of each other, then the known random variable X does not contribute any information to the random variable Y and vice versa; therefore the mutual information was zero. On the other extreme case, if the random variable X was the deterministic function of

random variable Y and random variable Y was also the deterministic function of random variable X , then all the information passed was shared by both random variables. Besides, mutual information was non-negative, $I(X; Y) \geq 0$ and it was symmetrical, $I(X; Y) = I(Y; X)$.

The mutual information is not based on two specific messages, but from the overall viewpoint of the random variables X and Y , therefore the mutual information does not have a negative value. When the extract information is from one event, the worst case is zero. The uncertainty of an event will not increase just because the other event was known. The mutual information was symmetrical since the amount of information on random variable X obtained from the random variable Y was the same as the amount of information on random variable Y received from the random variable X . The amount of information obtained from another event was at most as much entropy as another event, and it will not exceed the amount of information contained in itself, $I(X; Y) \leq H(X)$ or $I(Y; X) \leq H(Y)$. When the random variables X and Y corresponds one to one it implies that $I(X; Y) = H(X)$ and $H(X/Y) = 0$. When the random variables X and Y were stand-alone of each other, $H(X/Y) = H(X)$ implies that $I(Y; X) = 0$. No information was obtained from one event to another event which was the same as the situation of channel interruption. The entropy of the random variables X and Y , the joint entropy of the random variables X and Y and mutual information between random variables X and Y is illustrated in Figure 3.1 below.

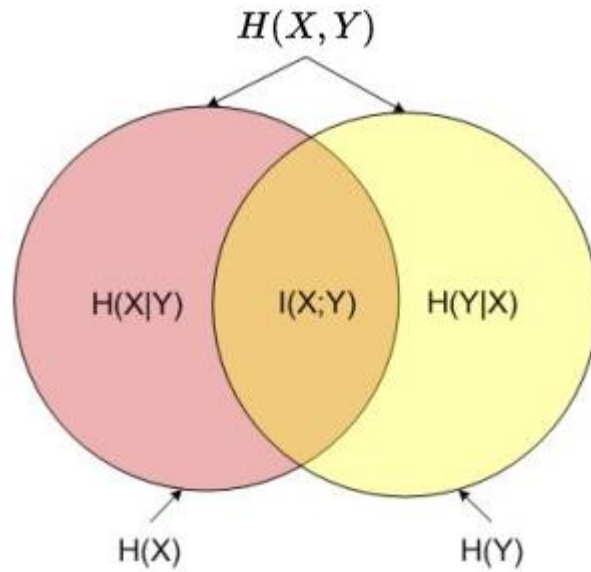


Figure 3.1: Relation between mutual information and entropy

Mutual information measures the amount of information contained between the features and the label class. If the feature belongs to the label class, then it has an enormous amount of mutual information. The mutual information does not require any assumptions about the nature of the relationship between the features and the label class, and it is well suited for the feature selection in bioinformatics. The mutual information calculation is similar to information gain, and the average of mutual information is also information gain.

The mutual information can also be written as:

$$\begin{aligned}
 I(X; Y) &= \sum_{x,y} p(x, y) \log \frac{p(x,y)}{p(x)p(y)} & (1) \\
 &= \sum_{x,y} p(x, y) \log \frac{p(x|y)}{p(x)}
 \end{aligned}$$

$$\begin{aligned}
&= \sum_{x,y} p(x,y) \log p(x) + \sum_{x,y} p(x,y) \log p(x|y) \\
&= \sum_{x,y} p(x,y) \log p(x) - \left(- \sum_{x,y} p(x,y) \log p(x|y) \right) \\
&= H(X) - H(X|Y)
\end{aligned}$$

By symmetry, $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$. Since $H(X,Y) = H(Y|X) + H(X)$, therefore $I(X;Y) = H(X) + H(Y) - H(X,Y)$.

3.2 The Optimal Baseline based on Ranked Features

In this section, an optimal baseline will be obtained using the ranked features based on mutual information. The past researches have proven that the mutual information can measure the similarity between the two elements regardless of their distribution, whether linear or nonlinear. When the mutual information is most significant, it means that these two elements are closest to each other. The optimal baseline can be obtained using the measurement of mutual information of all the features in the label class. When all the features are ranked according to the mutual information, the most significant feature is the most similar to the label class, and this feature is sufficient to represent the label class. When more significant features are selected, the compact subset of features will well present the label class. By measuring the performance of the ranked features in a classifier, the more ranked features added into the

classifier, the better is the performance. From this point of view, there will be a point or a few points on the number of ranked features which will indicate the highest performance in the classification.

In microarray data, due to the high dimensionality, the data set always contain irrelevant features and noise. It is essential to find a better baseline involving as many as relevant features rather than using all the features. Since the relevant features are ranked according to the mutual information, likewise the ranked features are ranked according to their relevancy to the label class. The performance of using these relevant features will outperform than using all the features in classification, hence a better baseline can be obtained using those ranked features. At the same time, the number of features, k that is needed to achieve this optimal baseline, can also be known. The number of features, k plays a vital role in a classification problem, as current research has no idea on how many features are needed to achieve the current baseline because the current baseline is obtained using all the features. Therefore, previous research in feature selection only focused on getting some selected feature which are less than all other features is reasonable. Thus, there is no standard guideline on the selected features that cannot exceed the specified number of features. Now with this new number of feature, k , any other feature selection method should not exceed excess k to achieve the same performance in the classification problem.

3.3 Algorithm on Ranked Features

Mutual information can measure the similarity between the features and class. When the mutual information score is more significant, it means that the feature is closer to the label class. The mutual information score for all the features between the label class can be computed using equation (1). The microarray data set was a $N \times M$ matrix where N represents the number of attributes, and M represents the number of samples. For a feature set $X = \{x_1, x_2, \dots, x_N\}$ of a data set D with N dimension and M sample, the mutual information for each x_i between the label class will be computed. The real experimental data will be normalised into $(-1,1)$ before calculating the mutual information score. The advantage of remapping the feature into three equal bins is that no data pre-processing is needed when the data has a missing value, as the calculation of mutual information only depends on the remapped frequency count and not based on the original value. Again, mutual information did not consider the nature of the relationship between the features and the label class; therefore, mutual information can be applied to the data either linear or nonlinear.

Each feature will be divided into three equal bins, and these three bins represent the expression of the microarray data in low, normal and high categories. The label class will be divided into p equal bins, where p is the number of classes of the label class. The features with three equal bins and the label class with p equal bins will then be remapped into a frequency count and form a probability mass function (pmf) and the joint probability mass function

(pdf) of each x_i and label class. The mutual information score for each feature and label class will be computed using the equation (1). Then, the features will be ranked according to the mutual information score and the graph of accuracy versus the cumulative of ranked features will be plotted. The highest accuracy from the graph will represent the optimal baseline, and the cutoff number of the feature can be obtained at the same time. Figure 3.2 shows the flowchart of finding the optimal baseline and the unique cutoff number of features.

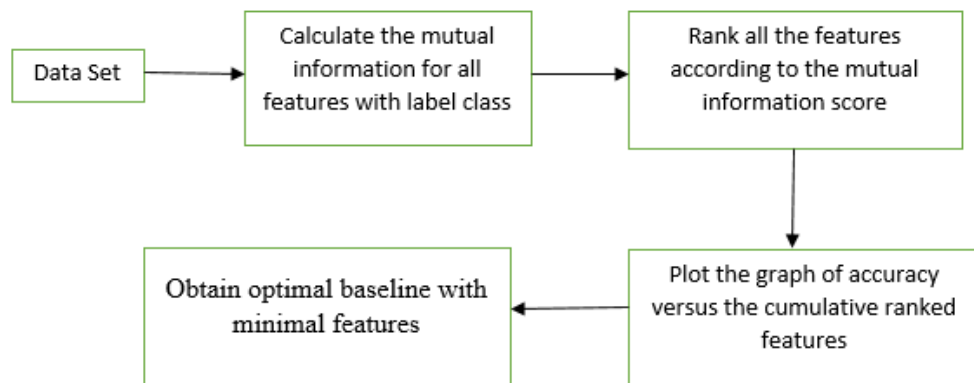


Figure 3.2: Flowchart of finding the optimal baseline and the unique cutoff number of features

The performance of the ranked feature measure using the support vector machine classifier and the accuracy of the research was defined as follows:

$$accuracy = 1 - \frac{false\ negative + false\ positive}{false\ negative + false\ positive + true\ positive + true\ negative}$$

or

$$accuracy = \frac{true\ positive + true\ negative}{false\ negative + false\ positive + true\ positive + true\ negative}$$

The next section shows the algorithm on calculating the mutual information score for each feature in a microarray data set.

The algorithm on calculating mutual information score:

Input:

A training sample D with a full feature set $X = \{x_1, x_2, \dots, x_N\}$ and the label class C with N dimension.

Output:

1. (Initialisation) Set $X \leftarrow$ "initial set of N features"
2. (Normalising) For $\forall x_i \in X$, normalise each x_i in $[0,1]$
3. (Remapping) For $\forall x_i \in X$, remapped each x_i into three equal bins and for remapped the label class C into p equal bins
4. (Forming) For $\forall x_i \in X$, formed a frequency count and a joint pdf the each x_i and label class C
5. (Calculating) For $\forall x_i \in X$, calculate the mutual information for each pair of x_i and label class C using the joint pdf
6. (Ranking) For $\forall x_i \in X$, ranked the mutual information score for each x_i in descending order
7. (Plotting) Plotted a graph of accuracy versus cumulative ranked features

8. (Identifying) Identify the highest accuracy and the number of features

Table 3.1 shows an example of a feature with microarray expression versus the label class. Assume that the label class “1” represents the tumour cell, and label class “0” represents the normal cell.

Table 3.1: The expression of feature respect to label class

Label	1	1	1	1	1	0	0	0	0	0
Feature	0.53	1.83	2.25	0.86	0.31	8.69	9.56	10.34	13.57	12.76

The feature and the label class in Table 3.1 will then be remapped into a frequency count in three equal bins for a feature and two equal bins for label class in Figure 3.3 which shows the 3-dimensional histogram plot. The probability mass functions (pmf) and the joint probability mass functions (pdf) of a feature and the label class are shown in Table 3.2 below.

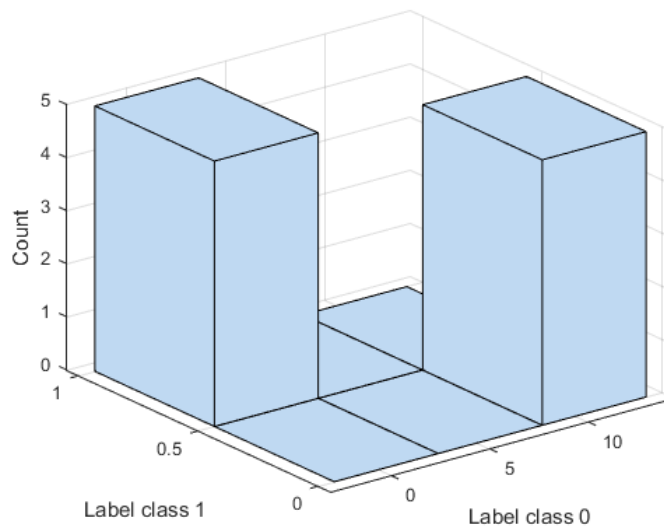


Figure 3.3: Joint distribution for a feature in the 3-dimensional histogram

Table 3.2: Joint probability mass function of a feature and the label class

Joint pmf of X and Y	Label Class	
	0	0.5
Feature	0	0
	0.5	0

Using the information in Table 3.2, the mutual information between the feature and label class was 1 bit; this value was derived using equation (1). The same calculation will be applied to all features in the microarray data set, and a set of mutual information score, MI , will be obtained in $N \times 1$ dimension. The mutual information score will be ranked in descending order. The first feature with the highest mutual information score was the feature which is most similar to the label class where this feature will be well presented in the label class. Again, when more features with high mutual information were added to the predictive model, the performance of the classification will become better. The performance of the predictive model, which is accuracy, will be plotted against the accumulative ranked features. From the graph, it can be observed that the first highest accuracy serves as the optimal baseline for this data set. At the same time, the number of features, k to obtain this highest accuracy can be obtained. Figure 3.4 shows an example of the accuracy graph versus the cumulative ranked features. The data set used in this example consists of 50 features, and the full features are ranked according to the mutual information score. From this data set, it is shown that the highest mutual information score of the feature gives an accuracy of 90.7%, and the baseline using all the features was 93.3%.

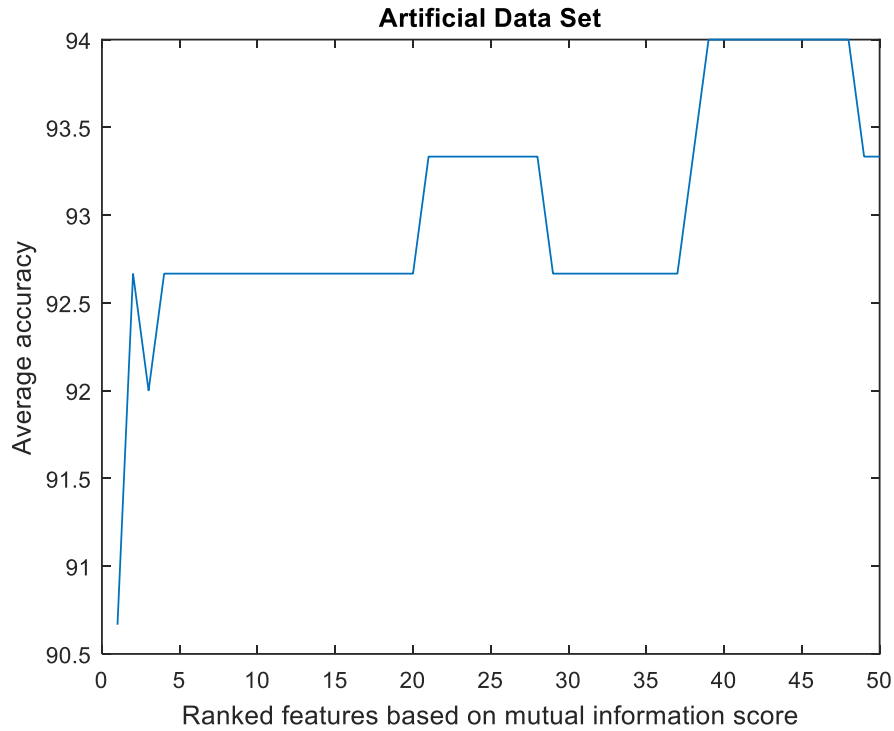


Figure 3.4: Average accuracy versus the ranked features based on the mutual information score

The proposed idea here is to obtain an optimal baseline based on the ranked features, and from the graph, the optimal baseline has the highest accuracy of 94% with 39 ranked features. Therefore, the optimal baseline will be 94% instead of the baseline using all the features, which are 93.4%. Not only can the optimal baseline be obtained from the ranked features using a mutual information score, but the number of features needed to achieve this optimal baseline is also known. This study tells us that as long as not more than 39 features can reach this optimal baseline, that is to say, the feature selection method need only less than 39 features to reach this optimal baseline. From here, we can deduce how best is this data set on classification, and we should not take more than 39 features to reach the optimal baseline. This idea allows us to compare the performance within the data set itself unlike the

previous research. We have no clear idea of how many features are needed to achieve the baseline because the previous baseline was obtained using all the features. Therefore, we not only have an optimal baseline but a unique cutoff number of features before we start to reduce the dimension of the data.

CHAPTER 4

DIMENSION REDUCTION WITH SUPPORT VECTOR MACHINE

4.1 Ranked Mutual Information with Support Vector Machine (rMI-SVM algorithm)

4.1.1 Dimension Reduction

The earlier chapter explained how to obtain the optimal baseline based on the ranked features by mutual information. This present chapter will describe how to reduce the high dimensional data set, such as microarray data that always consist of more than a thousand to ten thousand features. These microarray data always consist of redundant features, irrelevant features and noise. Therefore, it is essential to remove all these features and noise before building the predictive model. With the inclusion of these features it will increase the complexity of the predictive model and yield a low performance in classification resulting to overfitting which is another issue that requires attention. The proposed algorithm in this chapter will select the newly added features to provide new information to the predictive model. The idea adopted from mutual information captured the most relevant subset of features by ranking the features according to the mutual information score. Thus, the support vector machine (SVM) classifier plays a vital role to reduce the dimensionality of the data set.

In addition, the proposed algorithm will obtain a smaller number of features to build the predictive model, and is also able to show that the selected features give a better performance compared to the model when using full features (existing baseline). The number of features selected by the proposed algorithm also provides the same or better performance compared to the new proposed baseline in the previous chapter using ranked features. At the same time, the proposed algorithm will guarantee that the newly added feature will provide new information to the predictive model and obtain a strictly increasing accuracy graph versus the number of features. The features selected by the proposed algorithm will guarantee that the predictive model will only get better and better when more features are added to the predictive model.

4.1.2 Relevancy and Redundancy

In this study, the feature relevancy and feature redundancy will be defined as follows:

Definition 1: (Feature relevancy). Feature x_i is more relevant to the label class C than feature x_j if $I(x_i; C) > I(x_j; C)$.

Definition 2: (Feature redundancy). Feature x_i is a redundant feature to feature x_j with respect to the label class C , if $I(x_i; C) = I(x_j; C)$.

Mutual information is a powerful tool in finding the relevant features from a vast and high dimensional data set. However, from definition 2, with only the mutual information score, the relevant features and the redundant features cannot be differentiated. The mutual information score in the previous chapter is only able to rank the relevancy among the features with the label class but it failed to give any information between the features themselves. Therefore, removing the redundant features is an essential step to reduce the dimensionality of the high dimensional data set. This study uses the filter method with a support vector machine to remove the redundant features. Although the filter method is widespread in the feature selection method, nevertheless, it has only been studied during the past few decades. The previous researchers filtered the redundant features by using the mutual information score between the features (non-dynamic change methods) and subsequently use the conditional mutual information score between the already selected features and label class with the newly added feature (dynamic change methods).

In this study, the proposed algorithm used a support vector machine (SVM) to filter the redundant features taking into consideration the dynamic change between the newly added feature with the already selected features and label class. The proposed method can avoid overly burdensome calculations on the conditional mutual information, and can quickly remove the redundant features. The proposed method indicates that the first screening will significantly reduce the dimension of the data set, and with a small number of screening processes, smaller features can be obtained. The predictive model

using these smaller features can achieve the same performance, or better performance compared to the optimal baseline.

The first feature in the ranked features is the most relevant feature with respect to the class while the last feature in the ranked features will contain less amount of information needed in the predictive model. The feature with the highest mutual information score will set as a targeted feature in the selected subset, S . Towards the end, the selected subset, S , will gather the needed smaller features to build the predictive model. In general, the remaining ranked features after excluding the targeted feature can be categorised into two groups.

Group 1: Features that has high relevancy to the class but with low redundancy to the selected subset S .

Group 2: Features that has high relevancy to the class but with high redundancy to the selected subset S .

Features from Group 1 will provide better performance in the predictive model than the features from Group 2. In past researches, the researchers always use conditional mutual information to filter out the features from Group 2. The proposed algorithm is to find the features that are highly relevant to the class but with low redundancy to the selected subset S , which is the feature from Group 1.

4.1.3 Increment of the Information Content

The proposed algorithm will select the new candidate feature by observing the performance of the predictive model using a SVM classifier. The next candidate feature will be chosen according to the amount of new information added to the predictive model. When new information is added to the predictive model, the performance of the classification will become better. Conversely, if no new information is added to the predictive model, then the performance of the classification will remain the same or become worse. In general, when a candidate feature is added to the model, three different phenomena will happen:

- 1) First phenomenon, if the candidate feature x_j was an irrelevant feature or noise in respect to the subset S , then the performance of the model will decrease. Therefore, the performance of the predictive model with $S \cup \{x_j\}$ will be lower than the performance of the predictive model with S only.
- 2) Second phenomenon, if the candidate feature x_j was a relevant feature but is redundant, meaning that this candidate feature did not provide additional information content to the predictive model to the subset S , then the performance of the predictive model will remain unchanged. Therefore, the performance of the predictive model with $S \cup \{x_j\}$ will be the same with the performance of the predictive model with S only.

3) Third phenomenon, if the candidate feature x_j was relevant, and it provides an increment of information content to the subset S , then the performance of the predictive model will increase. Therefore, the performance of the predictive model with $S \cup \{x_j\}$ will be higher than the performance of the predictive model with S only.

These three phenomena can be shown in the graph of the performance of the predictive model using SVM versus the cumulative ranked features. Therefore, the next selected candidate feature will be the feature from the third phenomenon. The performance of the predictive model will be tested by adding the ranked candidature feature one by one. When the newly added candidate feature gives a better performance than the performance of the predictive model, before adding this candidate feature, then this candidate feature will be selected. The newly added candidate feature provides new information to the predictive model. When the newly added candidate feature provides the same performance as the performance of the predictive model before adding this candidate feature, then this candidate feature will be filtered out. The newly added candidate feature is a redundant feature with respect to the selected subset S ; therefore, no new information is added to the predictive model. When the newly added candidate feature gives a lower performance than the performance of the predictive model before adding this candidate feature, then this candidate feature will be filtered out. The newly added candidate feature might be an irrelevant feature or noise.

4.1.4 Minimal Features

In past researches, the performance of the model will become better when more relevant features are added to the predictive model. When more informative features are added to the predictive model, the predictive model will perform better. Based on the methods of past researches, the researchers have great freedom in choosing any number of features that are needed to build the predictive model. Usually, researchers will test the performance of the predictive model with a range of k , $5 \leq k \leq 50$, where k is the number of a feature used to build the predictive model. The optimum value k differs among the data set, and this will depend on the amount of information contained in the data set. For example, a subset of five features can give better classification performance compared to a subset of ten features, if the five features contained high information content compared to the ten features. Therefore, it does not always guarantee that more features in a predictive model will yield better performance, but in contrast, it will increase the complexity of the predictive model.

The proposed method in this study will search all the ranked features, until there are no more features added into the predictive model. Also, the proposed algorithm guarantees that the next added feature will improve the performance of the predictive model. As evidenced from past researches, when more features are added into the predictive model, the performance of the classification are not rising, but are gradually increasing with frustrated motion. This will lead to a significant problem when someone wants to choose

a certain number of elements. For example, a researcher can set a value for k from five to seven. At the same time, the accuracy of using five features was 90%, and the accuracy of using six features may become 88%, and when using seven features, the accuracy rose to 92%. In this situation, what is the optimum number of features as the minimum number of features can be five, but with seven features, the predictive model can achieve better performance.

The most exciting is the existence of the sixth feature which is either essential or is not important. Since it is not a relatively small number of selected features, and the performance is not the best, so why does this feature appeared in the subset of selected features. Is this feature redundant? If this feature is removed, will the performance of the predictive model get better? The proposed algorithm can alleviate this issue by getting a subset of features that give strictly increasing performance when more features are added, and obtained the minimal features. Additionally, a smaller number of features is more preferable and by reducing it helps the biologists to predict the disease accurately (Bir-Jmel et al., 2019).

4.2 Algorithm on the rMI-SVM algorithm

From the proposed algorithm, a smaller number of features can be obtained, and the performance of the predictive model is better than using the full features (baseline). The performance using these lower number of features also gives the same or better performance compared to the new proposed baseline using the ranked features in Chapter 3. At the same time, the

proposed algorithm will guarantee that the newly added feature will provide new information to the predictive model and obtain a strictly increasing accuracy graph versus the number of features. The features selected by the proposed algorithm will guarantee that the predictive model will only get better and better when more features are added to the predictive model.

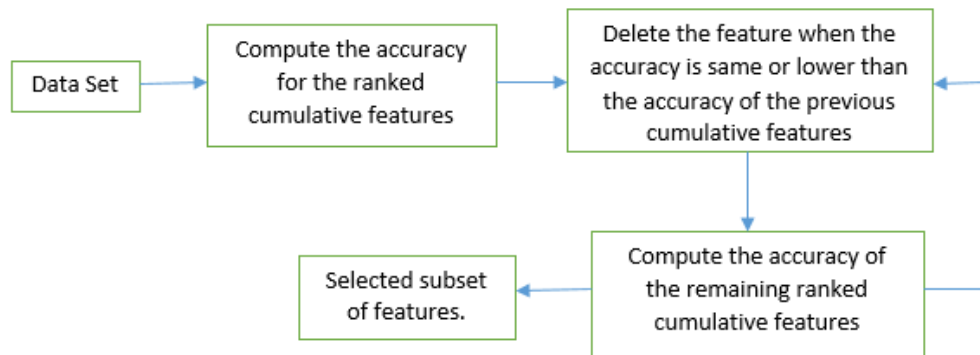


Figure 4.1: Flowchart of finding a smaller selected feature

Figure 4.1 shows the flowchart of finding a smaller selected feature. The accuracy for the ranked cumulative features has been calculated after which the feature will be deleted based on the comparison of the accuracy. A graph can be plotted using the accuracy versus the ranked cumulative features.

The next section shows the algorithm upon selecting the smaller features using the support vector machine.

The algorithm upon selecting the smaller features:

Input:

A training sample D with a full feature set $X = \{x_1, x_2, \dots, x_N\}$ and the label class C with N dimension.

Output:

1. (Initialisation) Set $X \leftarrow$ "initial set of N features"
2. (Normalising) For $\forall x_i \in X$, normalise each x_i in $[0,1]$
3. (Remapping) For $\forall x_i \in X$, remapped each x_i into three equal bins and for remapped the label class C into p equal bins
4. (Forming) For $\forall x_i \in X$, formed a frequency count and a joint pdf for each x_i and label class C
5. (Calculating) For $\forall x_i \in X$, calculate the mutual information for each pair of x_i and label class C using the joint pdf
6. (Ranking) For $\forall x_i \in X$, ranked the mutual information score for each x_i in descending order
7. (Plotting) Plotted a graph of accuracy versus cumulative ranked features
8. (Filtering) Deleted the feature when the accuracy is same or lower than the accuracy of the previous cumulative features
9. (Recompute) Compute the accuracy of the remaining ranked cumulative features from step 8
10. (Repeating) Repeat steps 8-9 until no more features are filtered out

For example, Figure 4.2 shows the average accuracy versus the ranked features based on the mutual information score. The third feature will be deleted because when the third feature is added into the predictive model, the

average accuracy is dropped. The third feature did not give any newly added information to the predictive model; because the third feature is noise which may worsen the predictive model. The fifth feature until the 20th feature will be removed because when these 16 features are added into the predictive model, the performance of the predictive model remains the same showing that these 16 features did not give newly added information to the predictive model; these 16 features are, therefore, deemed redundant. Based on Figure 4.2, only the first, second, fourth, 21st, and 39th will be selected in the first round. Then, the same process will be repeated for these five features, until no more features can be deleted based on the comparison of the accuracy.

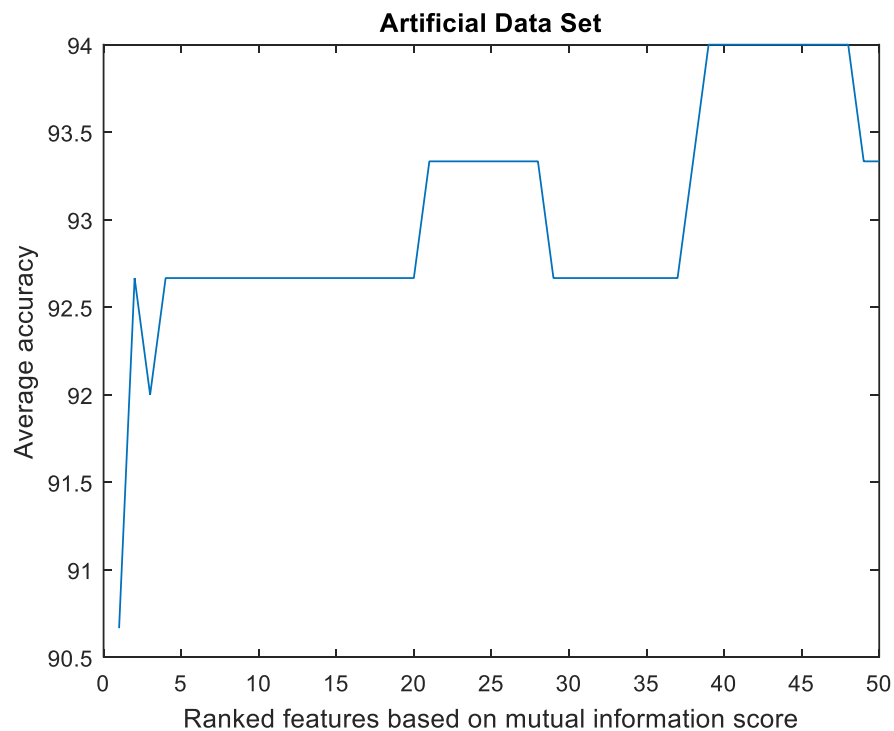


Figure 4.2: Average accuracy versus the ranked features based on the mutual information score

Figure 4.3 shows the average accuracy for a smaller selected feature from the example above. The example in Figure 4.2 uses data with 50 features. After removing those redundant features and noise, only two features remain. The number of features has reduced 96% of the original features, and it shows that there is an excellent dimensionality reduction. The average accuracy shows an increment from 93.3% (using full features- existing baseline) to 95.5% (using two selected features). The optimal baseline as proposed in Chapter 3 was 94% with 39 features, and after dimension reduction, the accuracy becomes 95.5% with two features only. As promised earlier, the proposed algorithm is able to get smaller features to achieve better accuracy compared to the accuracy using all the features and it is even better than the optimal baseline proposed in Chapter 3. Besides this, a strictly increasing graph as observed in Figure 4.3, where the newly added feature will provide new information to the predictive model. The features selected by the proposed algorithm show that the performance of the predictive model will only get better and better when more features are added to the predictive model.

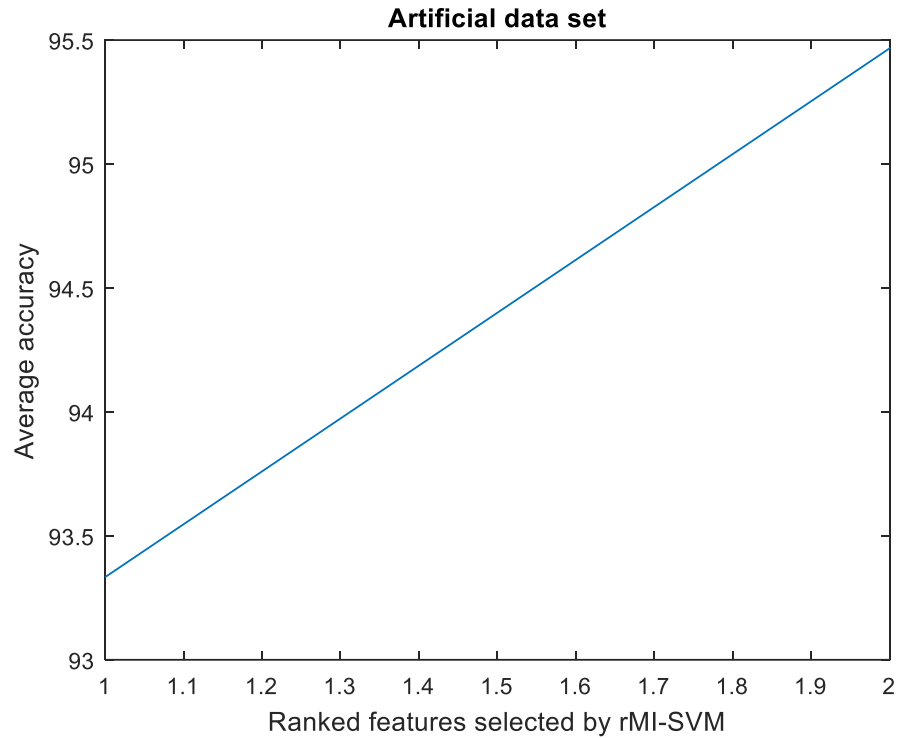


Figure 4.3: Average accuracy for a smaller selected feature

The proposed method rMI-SVM algorithm will select a smaller number of features by plotting the strictly increasing average accuracy graph, and the proposed method significantly reduces the dimension of the data set with minimal features. The new findings of this proposed method are the redundant features and noise can easily be detected in an accuracy performance graph versus the cumulative ranked features. The horizontal line in the accuracy graph indicates the redundant features because there is no input of new information to the predictive model, as a result, there is no increment in the accuracy. When noise is added to the predictive model, the accuracy of the predictive model will decrease. These redundant features and noise will then be deleted, and the same process will continue until all remaining features provide new information to the predictive model.

CHAPTER 5

EVALUATION OF rMI-SVM

5.1 Classifier

The classifier is an essential tool in data mining. The concept of classification is to learn a classification function and the classifier builds a classification model based on the existing data. The building model can then be applied to data prediction. According to Akadi et al., (2008); Che et al., (2017) there are several classifiers such as linear regression, logistic regression, support vector machine classifier, naïve Bayes classifier, k -nearest neighbour, decision trees classification, random forest, gradient boost, and so on. For this research, the support vector machine classifier will be used due to the high predictive accuracy performance compared to k - nearest neighbour classifier or decision tree classifier. To evaluate the classification error rate, the k - fold cross-validation will be used in this research.

There are several kernel functions in support vector machine classifier such as linear, polynomial, radial basis function (RBF) and sigmoid. In this research, the data set will be tested by four different types of kernel functions and from here a kernel function will be chosen that gives the highest performance in the predictive model. The four types of kernel functions are linear, quadratic, cubic and RBF. The purpose of testing the data set with these

four different kernel functions is to identify which kernel function is suitable for a given data set because the distribution of the data set was unknown until it has been tested.

5.2 Receiver Operating Characteristic Curve

The receiver operating characteristic curve has been used in the fields of biology, criminal psychology and recently it has been well developed in the areas of machine learning and data mining. In medicine, it is widely used in the diagnosis of diseases, and is also applied in empirical medical research, radiology and social science research. The receiver operating characteristic curve acts as alternative methods that are easy to operate and incorporate with the gold standard for clinical identification. For example, to identify a biopsy of cancer, the receiver operating characteristic curve can be used to replace the gold standard in classifying the tumour into cancerous or non-cancerous. The y-axis of the receiver operating characteristic curve represents the true positive rate, also known as sensitivity while the x-axis of the receiver operating characteristic curve represents the false-positive rate. The sensitivity refers to the probability that the result is correctly classified as positive, and the specificity refers to the probability that the result is correctly classified as negative. In medicine, sensitivity indicates the probability that a person with a disease is classified as positive and the specificity indicates the probability that a normal person is classified as negative.

		<u>True class</u>	
		p	n
<u>Hypothesized class</u>	Y	True Positives	False Positives
	N	False Negatives	True Negatives
Column totals:		P	N

Figure 5.1: Confusion matrix

Figure 4.4 shows the confusion matrix. The sensitivity, specificity, false-positive rate and false-negative rate were defined as follows:

$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

$$\text{Specificity} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}}$$

$$\text{False – positive Rate} = \frac{\text{False Positive}}{\text{True Negative} + \text{False Positive}}$$

$$\text{False – negative Rate} = \frac{\text{False Negative}}{\text{True Positive} + \text{False Negative}}$$

Precision and recall are two metrics which are widely used in the field of information retrieval and statistical classification to evaluate the quality of the results. The precision refers to the ratio of the number of related items retrieved from the total number of items retrieved, the recall rate refers to the

ratio of the number of related items retrieved from the number of related items in the system. In general, precision is the number of items accurately retrieved while recall is on how many accurate items are retrieved. The definitions of precision and recall are as follows:

$$\text{Precision} = \frac{\textit{True Positive}}{\textit{True Positive} + \textit{False Positive}}$$

$$\text{Recall} = \frac{\textit{True Positive}}{\textit{True Positive} + \textit{False Negative}}$$

The higher the precision and recall, the better it is, but these two are contradictory in some cases. In medicine, the recall is much more important than precision. The reason is that false-negative is more critical than the false-positive in diagnosis cases. False-positive did not give any harm to the patient without the disease because they only need to run through more test to confirm the absence of the disease. The true positive and false-negative are relevant elements in disease diagnosis cases. The receiver operating characteristics curve has an outstanding characteristic that it will remain unchanged when the distribution of the positive samples and negative samples change in the test. When the receiver operating characteristics curve was investigated, the diagonal line was taken as a reference line. If the receiver operating characteristics curve of the classifier just falls on the diagonal reference line, it means that the classifier does not discriminate the diagnosis of a disease. If the receiver operating characteristics curve moves to the upper left, the higher is the sensitivity of the classifier to the disease and the lower the false-positive rate, this means the discriminating power of the classifier is better. When the

point closest to the upper left corner (0,1), it becomes the point of least misclassification, its sensitivity is the largest, and the false-positive rate the smallest.

Generally, when measuring the quality of a classifier, in addition to the receiver operating characteristics curve, the area under curve can be used to discriminate the discriminating power of the receiver operating characteristics curve. The area under the curve value ranges from 0 to 1, and the larger, the better is the classifier. In medical diagnosis, the main task is to find out the disease of the patient, which is high in true positive rate. At the same time, the misdiagnosing of the patient without disease as with the disease that was low in false-positive rate. It is not difficult to establish that the true positive rate and the false-positive rate are mutually restrictive. If a doctor was sensitive to a symptom of a disease, then with a small symptom, the doctor will judge the patient with the disease. Therefore, the true positive rate should be high, but the false-positive rate will become higher accordingly. In the most extreme case, the doctor will treat all the patients with the disease; therefore, both the true positive rate and the false-positive rate will reach 1. In the case of the area under curve, which is greater than 0.5 and the closer the area under curve was to one, the better is the diagnostic effect. When the area under curve is equal to 0.5, it means that the diagnostic method is utterly ineffective and has no diagnostic effect.

5.3 Robustness

The sample size of the microarray data analysis is relatively small, therefore feature selection in small sample size might affect the feature selection by chance. The biomarker was an essential tool in diagnosis, prognosis or treatment in cancer and other medical diseases; therefore it is vital to have an algorithm that is able to figure out the necessary features and not feature selection by chance. In this research, the Z -score analysis (Li et al., 2001; Jirapech-Umpai and Aitken, 2005) is applied to determine the robustness of the algorithm in feature selection. The Z -score measures the significance of the occurrence of the feature selected. A feature with high Z -score value indicates that the feature was not chosen by chance. Besides that, an algorithm that selected a subset of features were among those selected features with a high Z -score value and were deemed to be more robust.

The Z -score of the feature defined as $Z = \frac{f_i - E(f_i)}{\sigma}$, where f_i was the frequency of the feature was selected, and σ was the standard deviation of f_i . Let N be the total number of features and $E(f_i)$ was the expected number of times feature i was selected and $f_{\bar{c}}$ be the average number of selected features, then the probability of feature i was selected was $\frac{f_{\bar{c}}}{N}$ and $E(f_i) = P(f_i)K$, $\sigma = \sqrt{P(f_i)(1 - P(f_i))K}$, where K is the number of replicates.

5.4 Comparison with other Relevant Work

Most of the methods were focused on the three essential criteria: relevancy, redundancy, and dynamic change on the increment of information content. Those previous methods were focused on maximising the relevancy between the features and label class and minimising the redundancy between the features simultaneously. This measurement can be in terms of independently as a dynamic change or dependently. The earlier research on feature selection methods based on information theory such as MIFS, MIFS-U, mRMR, was dependent on a parameter where this parameter was not well-defined and had no further information on how to determine the value of the parameter. Besides using the mutual information in finding the relevancy between the features and class, past research also combine mutual information with other methods or algorithm in searching for the redundant features. For example, MIFS-U was the evolution of the MIFS, and MIFS-U was combined with the Taguchi method in searching the new information, and while mRMR combines the filter method and wrapper method in obtaining the compact subset of features. The combine method or algorithm used by MIFS-U and mRMR might select the features with high relevancy and high redundancy at the same time.

Later on, NMIFS and NMIFS-ND became the enhanced edition of MIFS, MIFS-U and mRMR where these proposed methods did not involve any parameter in the algorithm. NMIFS used average normalised mutual information combined with genetic algorithm in finding the compact subset of

features. Similarly, NMIFS-ND used mutual information was combined with NSGA-II algorithm in finding the compact subset of features. These were the first generation of the feature selection methods that involve mutual information when combined with other techniques or algorithm in finding the compact subset of features.

The second generation of the feature selection methods starts to use the calculation on conditional mutual information or joint mutual information. JMI, JMIM, NJMIM, and MRI were using conditional mutual information. The joint mutual information can overcome the problem when two features to the label class had the same mutual information score. Therefore, the proposed second-generation methods used joint mutual information to filter redundant features. These methods consider the relevant features by calculating the mutual information between the features and label class. Next, the redundant features were filtered by using the joint mutual information between the candidate feature and the already selected features to the label class. That was the second generation method that considered the increment of information content into the predictive model.

In the year 2018, the third generation feature selection not only considered the relevancy, redundancy, increment of information content but also the dynamic change when the new candidate feature was added into the compact subset. MIFS, mRMR, JMI and MRI were the first and second-generation methods that did not consider the dynamic change when the new

candidate feature was added into the compact subset. DCSF was a proposed method that involved the dynamic change when the new candidate feature was added into the compact subset. DCSF used conditional mutual information between the feature and the label class to find the relevant feature which was different from the traditional proposed methods. However, DCSF needs to set a parameter as one of the first generation methods.

Overall, the previously proposed feature selection methods have slowly improved and strengthened from the first generation until the third generation. The past research have always used only the baseline which was generated using full features which was far away from the real result as this includes all the features which involved the irrelevant features, redundant features and noise. Besides this, the past research methods did not show the continually increasing performance when a new candidate feature was added into the predictive model and no guideline on how to select the minimum feature that was needed in the predictive model. The first proposed idea here to improve and strengthen the past feature selection method is to find a new and better baseline using ranked features by mutual information score. The proposed algorithm in Chapter 3 was able to obtain a better baseline that involves as many as relevant features that were ranked by mutual information score. Besides a better baseline can be obtained using the proposed algorithm, at the same time, the number of features that are needed to achieve this optimal baseline can be obtained. The number of features serves as a guideline on the maximum number of features that are required for any feature selection algorithm to get the new proposed baseline.

Secondly, the rMI-SVM algorithm was proposed in this research to reduce the dimensionality of the data set, and at the same time, a compact subset of features can be obtained. This compact subset guarantees that every newly added feature will provide new information to the predictive model. The rMI-SVM algorithm does not depend on any parameter as this method only use the information theory on mutual information. Therefore rMI-SVM algorithm was free from the trouble of setting the parameter in the algorithm. The rMI-SVM algorithm is a simple method that only involved the calculation on mutual information, unlike some of the past feature selection method such as MIFS-U and mRMR that requires a more complex calculation and time-consuming wrapped method to obtain the compact subset of features. The redundant features and noise can be easily detected from the performance graph versus the cumulative ranked features. When the newly added feature lowers down the performance of the predictive model, then this newly added feature is a noise. When the newly added feature, shows no improvement or remains unchanged in the performance of the predictive model, then this newly added feature is redundant.

The rMI-SVM algorithm considers the dynamic change of the newly added candidate feature by evaluating the performance of the predictive model using SVM classifier, and not as the previous feature selection method that used conditional mutual information. The rMI-SVM algorithm is able to select a compact subset of a feature that ensures each added new candidate feature will provide new information to the predictive model. Therefore, a strictly increasing average accuracy versus the number of features graph will be

obtained. Overall, the rMI-SVM algorithm considers the maximum relevancy of the features and filter the redundant features by taking into account the newly added candidate feature with the already selected subset of features to the label class.

5.5 Influence on the Classifier to the Predictive Model

Several classifiers are used in gene expression analysis, such as decision tree classifier, support vector machine classifier, and k -nearest neighbour classifier. The rMI algorithm combined with the SVM classifier is used in this model because the predictive accuracy is high, and the memory usage is acceptable. The rMI algorithm can combine with any classifiers as rMI is a classifier independent feature selection algorithm. The role of the classifier is to filter out the redundancy features based on the predictive model performance. The redundant features will give no additional information to the model and thus no improvement is shown in the predictive model. The rMI algorithm will only select the next candidate feature if new information is added to the predictive model. Although different classifiers may have a different level of accuracy, such as support vector machine classifier was high in predictive accuracy. In contrast, the accuracy of the k -nearest neighbour depends on the dimensional of the data set. Nevertheless, when a non-redundant informative feature was added to the model, all the classifiers will show an improvement in the accuracy.

5.6 Experimental Setup

The experiment was performed to study the efficiency and effectiveness of the proposed mutual information-based feature selection method, rMI-SVM algorithm. The experiments were conducted on a laptop with Processor Intel(R) Core(TM) i5-4210U CPU @ 1.70GHz, 2401 Mhz, 2 Core(s), 4 Logical Processor(s) and 8 GB of random access memory (RAM). The rMI-SVM algorithm was performed using MATLAB R2017b.

CHAPTER 6

RESULTS

6.1 Data Set

The data sets used in this research were microarray data sets, diseases data set, and a handwriting data set. Indeed, the microarray data used in this research were public microarray data where the data can be downloaded from the National Center for Biotechnology Information (NCBI) and UCI Machine Learning Repository, <http://archive.ics.uci.edu/ml/index.php>. The Gene Expression Omnibus (GEO) at NCBI was the largest fully public repository for molecular data and gene expression data. The GEO database was publicly accessible via the www at <http://www.ncbi.nlm.nih.gov/geo>. A large number of publications are published in several journals such as Expert Systems with Applications, BioMed Central (BMC) Bioinformatics, Journal of Machine Learning Research, International Journal of Computer Science and Network Security, Institute of Electrical and Electronics Engineers (IEEE), Nature Genetics, Bulletin of Mathematical Biology, Computers in Biology and Medicine, Journal of Bioinformatics and Computational Biology, Journal of Mathematical Biology, etc.

Also, the microarray data have been used by other authors and publishers in many journal publications. Ten data sets had been downloaded to

evaluate the performance of the rMI-SVM algorithm. The first six data set were the microarray data with high dimensional attributes, but a low number of sample size while the next four data set were the clinical data which relate to some diseases with low dimensional in attributes, but a high number of sample size and the last data set were handwriting recognition data set. The summary of the 10 data was shown in Table 6.1. There were six binary classification data set and four multiclass classification data set.

The reason for choosing these 10 data set was that the microarray data set is a high dimensional data set with low sample size and widely used in the published journal; the rMI-SVM algorithm can be applied here to evaluate its performance on dimension reduction.

Table 6.1: Summary of the downloaded data set

No.	Name	No. of attribute	No. of sample	Type of classification
1.	Colon cancer	1988	62	Binary
2.	Leukaemia	7128	72	Binary
3.	Prostate cancer	2135	102	Binary
4.	Lung cancer	1626	181	Binary
5.	Skin cancer	22215	15	3 classes
6.	Lymphoma	4026	96	9 classes
7.	Parkinson	22	195	Binary
8.	Breast cancer	30	569	Binary
9.	Lung cancer	56	326	3 classes
10.	Handwriting	649	2000	10 classes

The first data set was colon cancer data set with 1988 number of attributes and 62 samples. The colon cancer data set contains 40 tumour samples and 22 normal samples. Table 6.2 shows part of the ribosomal protein cluster (Alon, 1999). The second data set was leukaemia data set with 7128 attributes and 72 samples (Golub et al., 1999). The leukaemia data set contains 47 acute lymphocytic leukaemia (ALL) and 25 acute myeloid leukaemia (AML). The third data set was prostate cancer data set with 2135 attributes and 102 samples. The prostate cancer data set contains 52 tumour samples and 50 normal samples (Dessi et al., 2013).

Table 6.2: Part of the ribosomal protein cluster of colon cancer data set

Gene number	Sequence	Name
T63591	3' UTR	60S acidic ribosomal protein P0 (human)
R50158	3' UTR	<i>Mus musculus</i> L36 ribosomal protein*
T52642	3' UTR	Guanylate kinase homolog (vaccinia virus)
R85464	3' UTR	ATP synthase lipid-binding protein P2 precursor (human)
X55715	Gene	Human Hums3 mRNA for 40S ribosomal protein s3
T52185	3' UTR	P17074 40S ribosomal protein
T56934	3' UTR	<i>Homo sapiens</i> alpha NAC mRNA (transcriptional coactivator)
T47144	3' UTR	JN0549 ribosomal protein YL30
T72879	3' UTR	60S ribosomal protein L7A (human)
T57633	3' UTR	40S ribosomal protein S8 (human)
T58861	3' UTR	60S ribosomal protein L30E (<i>Kluyveromyces lactis</i>)
T52015	3' UTR	Elongation factor 1-gamma (human)
T57619	3' UTR	40S ribosomal protein S6 (<i>Nicotiana tabacum</i>)
T72938	3' UTR	Ribosomal protein L10*
R02593	3' UTR	60S acidic ribosomal protein P1 (<i>Polyorchis penicillatus</i>)
T48804	3' UTR	40S ribosomal protein S24 (human)
R01182	3' UTR	60S ribosomal protein L38 (human)
T61609	3' UTR	<i>H. sapiens</i> gene for ribosomal protein Sa, partial cds*
H77302	3' UTR	60S ribosomal protein (human)
U14971	Gene	Human ribosomal protein S9 mRNA, complete cds
H54676	3' UTR	60S ribosomal protein L18A (human)
R86975	3' UTR	40S ribosomal protein S28 (human)
T51560	3' UTR	40S ribosomal protein S16 (human)
H09263	3' UTR	Elongation factor 1-alpha 1 (<i>H. sapiens</i>)
T49423	3' UTR	Breast basic conserved protein 1 (human)
T63484	3' UTR	Human ornithine decarboxylase antizyme (Oaz) mRNA, complete cds
R02593	3' UTR	60S acidic ribosomal protein P1 (<i>P. penicillatus</i>)
R22197	3' UTR	60S ribosomal protein L32 (human)
T51496	3' UTR	60S ribosomal protein L37A (human)

The fourth data set was lung cancer data set with 1626 number of attributes and 181 samples. The lung cancer data set contains 31 malignant pleural mesotheliomas (MPM) and 150 adenocarcinomas (AD) (Podolsky et al., 2016). The fifth data set was skin cancer data set with 22215 number of attributes and 15 samples. The skin cancer data set contains six healthy people, four patient with actinic keratosis and five patient with squamous cell carcinoma (Nestor and Zarraga, 2012.). The sixth data set was lymphoma data set with 4026 number of attributes and 96 samples. There were nine classes in this data set. The samples were categorised into nine classes according to the category of the mRNA sample studied (Aguilar-Ruiz et al., 2004; Alomari et al., 2017).

The seventh data set was Parkinson data set with 22 number of attributes and 195 samples. The Parkinson data set contains 48 healthy samples and 147 Parkinson disease (PD) samples (Ramani and Sivagami, 2011). This data set consists of a series of biomedical speech measurements from 31 people, of which 23 had Parkinson disease. Each attribute in the data set represented a specific voice metric, and there were approximately six recordings per people. The attributes covered the measurements of average, maximum and minimum fundamental vocal frequency, several measurements on variation in fundamental frequency and amplitude, two measures of the ratio of noise to tonal elements in the voice, two nonlinear measurements of dynamical complexity, signal fractal scaling exponent and three nonlinear measurements in fundamental frequency variation.

The eighth data set was breast cancer data set with 30 number of attributes and 569 samples. The breast cancer data set contains 212 malignant samples and 357 benign samples (Salama et al., 2012). The data set was calculated from a digitised image of fine-needle aspiration of the breast lumps. It described the characteristics of the nucleus that are present in the image. For each cell nucleus, ten real-valued attributes were computed such as radius, perimeter, area, texture, smoothness, compactness and concavity, concave points, symmetry and fractal dimension of the nucleus. The ninth data was lung cancer data set with 56 number of attributes and 32 samples (Hong and Yang, 1991; Naseriparsa et al., 2013.). This data set was published in the Journal of Pattern Recognition (1991). Hong and Yang (Year) have used this data set on the optimal discriminant plane for a small sample size. This data set described the three types of pathological lung cancers. The 10th data set was a handwritten data set with 649 number of attributes and 2000 samples (Bins and Draper 2001). This data set consists of a number of handwritten features from zero to nine, and there were a total of ten classes in this data set.

6.2 Experiment Procedure

The optimal baseline will be computed for all data sets using the ranked features, as described in Chapter 3. This process will be repeated for 50 times to obtain the average mutual information score for each feature. Each time, a new training and the test set will be obtained. The optimal baseline with the number of features will be identified before using the rMI-SVM algorithm as described in Chapter 4 to reduce the dimensionality of the data

set and for choosing the compact subset of a feature to build the predictive model. All the performance of the compact subset of a feature will be evaluated by four different classifiers such as SVM, k -nearest neighbour, Naïve Bayes and Tree Decision with k -fold cross-validation. The rMI-SVM algorithm will be repeated five times to get the average accuracy. After this, the evaluation tools such as confusion matrix, ROC curve will be used to evaluate the performance of the chosen compact subset of a feature. The Z-score will be applied to the compact subset of a feature to show that the features are not selected by chance. Figure 6.1 shows the flowchart of the experimental procedure for obtaining the optimal baseline and the number of features, and Figure 6.2 shows the flowchart of the experimental procedure for dimension reduction on feature selection.

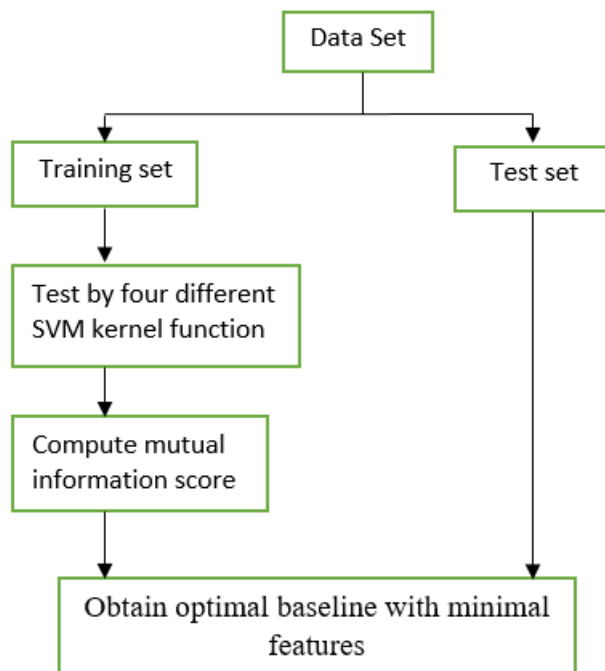


Figure 6.1: Flowchart of the experimental procedure for obtaining an optimal baseline and number of features

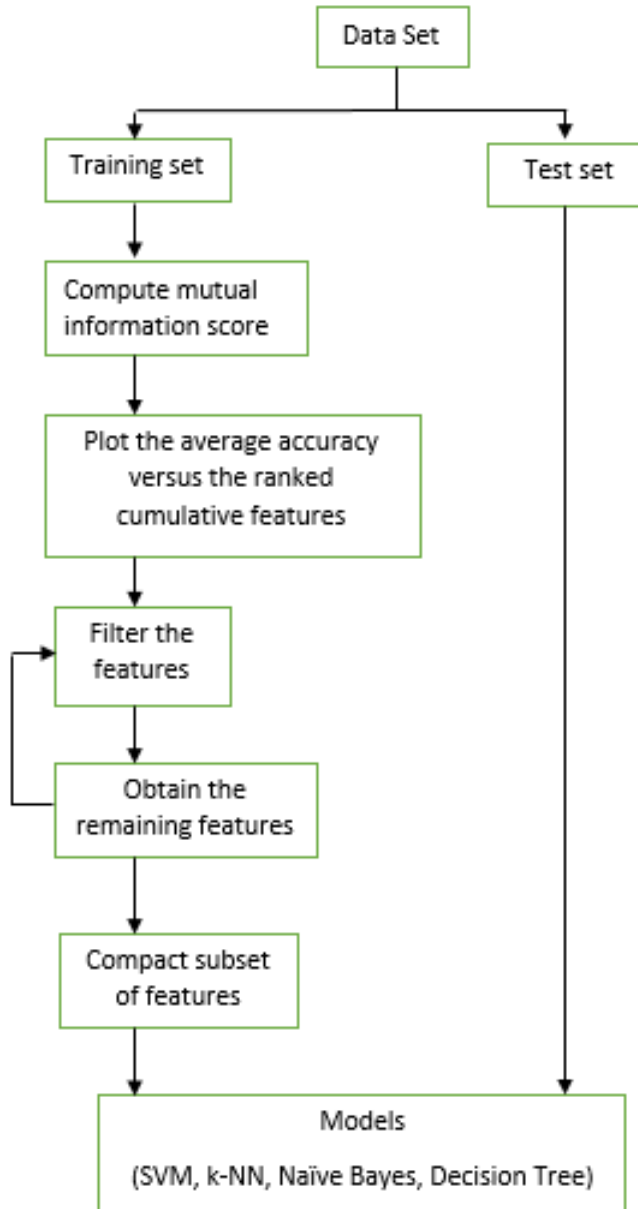


Figure 6.2: Flowchart of the experimental procedure for dimension reduction on feature selection

6.3 Evaluation on Four Different SVM Classifiers of Binary Data Set

Tables 6.3- 6.8 show the average accuracy of SVM cross-validation and average accuracy of SVM predictive model obtained from the four types of kernel function of SVM classifier of binary data set.

Table 6.3: Average accuracy of cross-validation and predictive model on four different kernel functions of classifier of colon cancer data set

Type of classifier	SVM Cross-validation	SVM Accuracy
SVM Linear	82.7273	80
SVM Quadratic	80	72.2222
SVM Cubic	60.9091	67.7778
SVM RBF	68.1818	68.8888

From Table 6.3, the colon cancer data set achieved the highest average accuracy of 80% by using the SVM classifier with linear kernel function, followed by SVM classifier with quadratic kernel function with the average accuracy of 72.22%. Similarly, the SVM classifier with radial basis function kernel function achieved the average accuracy of 68.89% and the lowest average accuracy was 67.78% by using SVM classifier with a cubic kernel function.

Table 6.4: Average accuracy of cross-validation and predictive model on four different kernel functions of classifier of a leukaemia data set

Type of classifier	SVM Cross-validation	SVM Accuracy
SVM Linear	97.6471	96.1905
SVM Quadratic	92.9412	82.8571
SVM Cubic	79.2157	70.4762
SVM RBF	89.8039	84.7619

From Table 6.4, the leukaemia cancer data set achieved the highest average accuracy of 96.19% by using the SVM classifier with linear kernel function, followed by SVM classifier with radial basis function kernel function with the average accuracy of 84.76%, SVM classifier with quadratic

kernel function with the average accuracy of 82.86% and the lowest average accuracy was 70.48% by using SVM classifier with a cubic kernel function.

Table 6.5: Average accuracy of cross-validation and predictive model on four different kernel functions of classifier of a prostate cancer data set

Type of classifier	SVM Cross-validation	SVM Accuracy
SVM Linear	91.3889	91.3333
SVM Quadratic	89.4444	90
SVM Cubic	90	70.4638
SVM RBF	83.3333	88

From Table 6.5, the prostate cancer data set achieved the highest average accuracy of 91.33% by using the SVM classifier with linear kernel function, followed by SVM classifier with quadratic kernel function with the average accuracy of 90%, SVM classifier with radial basis function kernel function with the average accuracy of 88% and the lowest average accuracy was 70.46% by using SVM classifier with a cubic kernel function.

Table 6.6: Average accuracy of cross-validation and predictive model on four different kernel functions of classifier of lung cancer data set

Type of classifier	SVM Cross-validation	SVM Accuracy
SVM Linear	99.0551	100
SVM Quadratic	99.2126	100
SVM Cubic	98.7402	94.4444
SVM RBF	97.9528	100

From Table 6.6, the lung cancer data set achieved the same average accuracy for the three classifiers: SVM classifier with linear kernel function, SVM classifier with quadratic kernel function and SVM classifier with radial basis function kernel function with an average accuracy of 100%. The lowest average accuracy was 94.44% by using SVM classifier with a cubic kernel function.

Table 6.7: Average accuracy of cross-validation and predictive model on four different kernel functions of classifier of Parkinson data set

Type of classifier	SVM Cross-validation	SVM Accuracy
SVM Linear	87.7372	80.3448
SVM Quadratic	87.5912	74.1379
SVM Cubic	89.0511	80
SVM RBF	89.0511	88.6207

From Table 6.7, the Parkinson data set achieved the highest average accuracy of 88.62% by using the SVM classifier with radial basis function kernel function, followed by SVM classifier with linear kernel function with the average accuracy of 80.34%, SVM classifier with cubic kernel function with the average accuracy of 80% and the lowest average accuracy was 74.14% by using SVM classifier with a quadratic kernel function.

Table 6.8: Average accuracy of cross-validation and predictive model on four different kernel functions of classifier of breast cancer data set

Type of classifier	SVM Cross-validation	SVM Accuracy
SVM Linear	97.3935	91.7647
SVM Quadratic	97.1930	85.5294
SVM Cubic	95.6892	86.4706
SVM RBF	95.8897	86.3529

From Table 6.8, the breast cancer data set achieved the highest average accuracy of 91.76% by using the SVM classifier with linear kernel function, followed by SVM classifier with cubic kernel function with the average accuracy of 86.47%, SVM classifier with radial basis function kernel function with the average accuracy of 86.35% and the lowest average accuracy was 85.53% by using SVM classifier with a quadratic kernel function. Therefore, based on the average accuracy analysis on SVM classifier with a different kernel function, the colon data set, leukaemia data set, prostate cancer data set, lung cancer data set, and breast cancer data set used SVM-Linear whereas the Parkinson data set used SVM-RBF in the algorithm later.

6.4 Optimal baseline of the Binary Data

The optimal baseline for binary data will be obtained using the algorithm provided in Chapter 3. Figures 6.3-6.8 show the average accuracy of the ranked features for colon cancer data set, the leukaemia data set, prostate cancer data set, lung cancer data set, Parkinson data set and breast cancer data set.

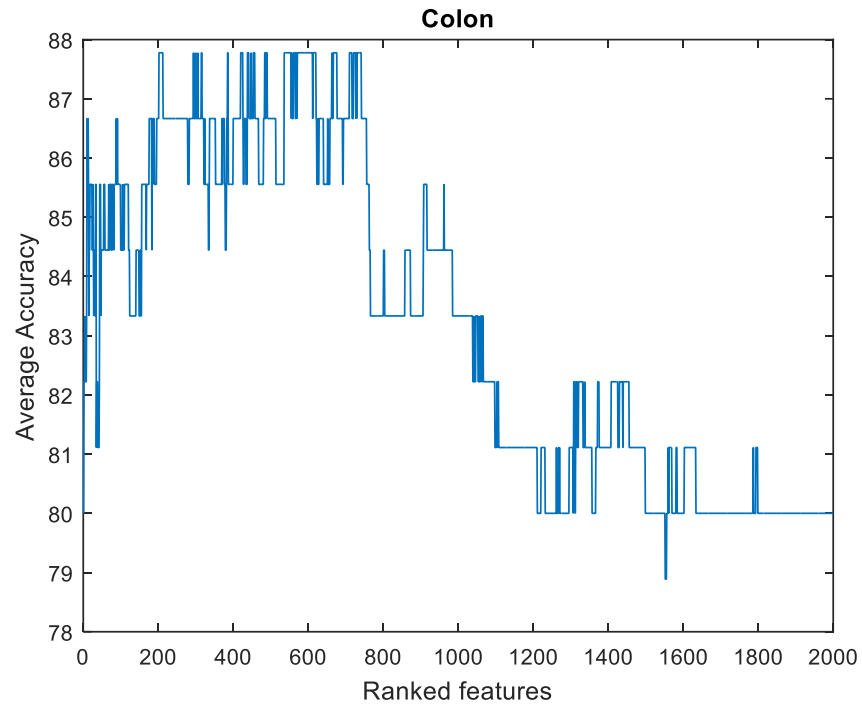


Figure 6.3: Average accuracy of the ranked features for colon cancer data set

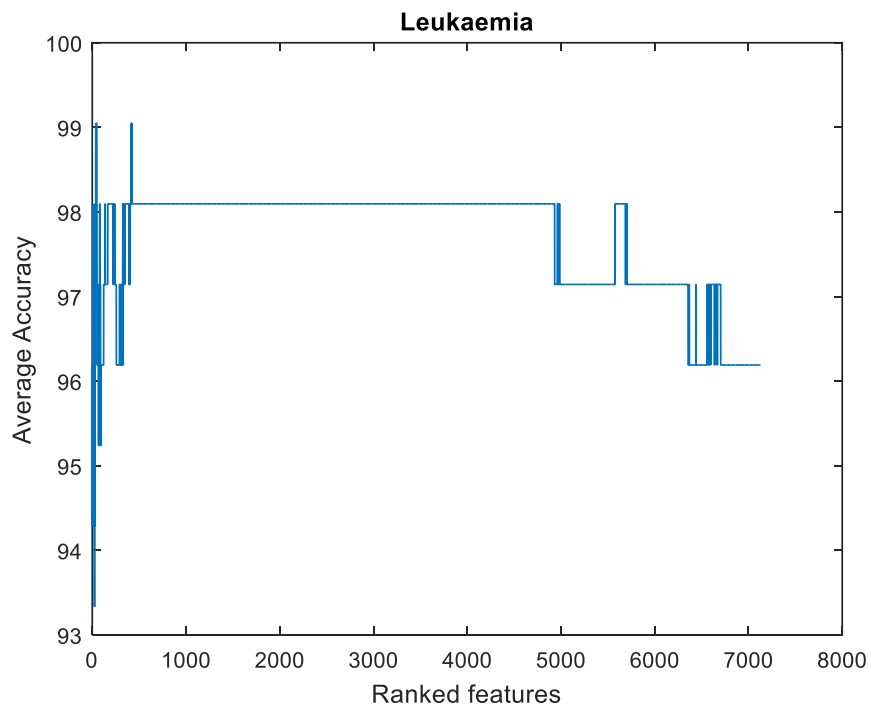


Figure 6.4: Average accuracy of the ranked features for the leukaemia data set

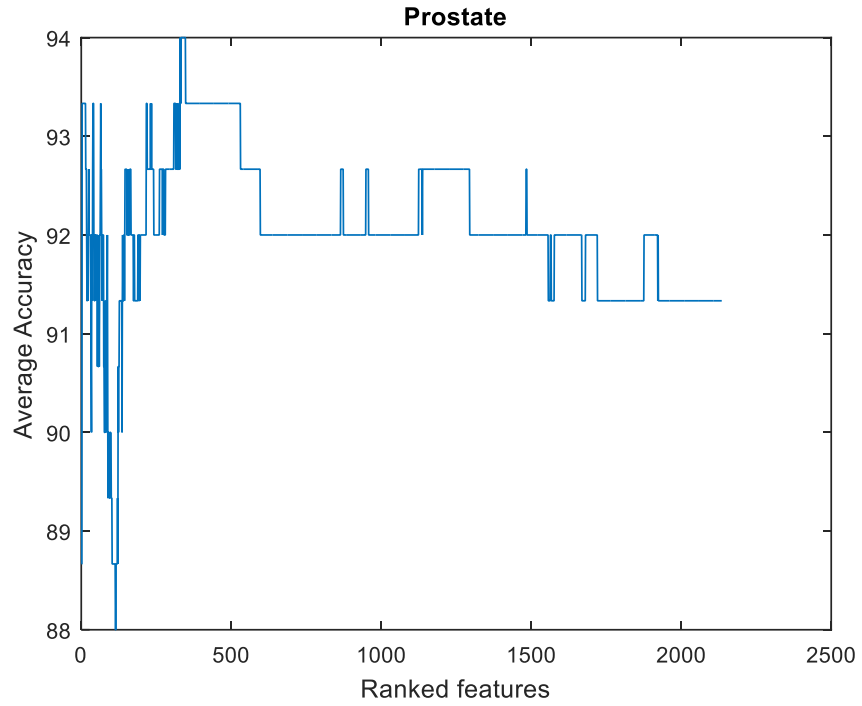


Figure 6.5: Average accuracy of the ranked features for prostate cancer data set

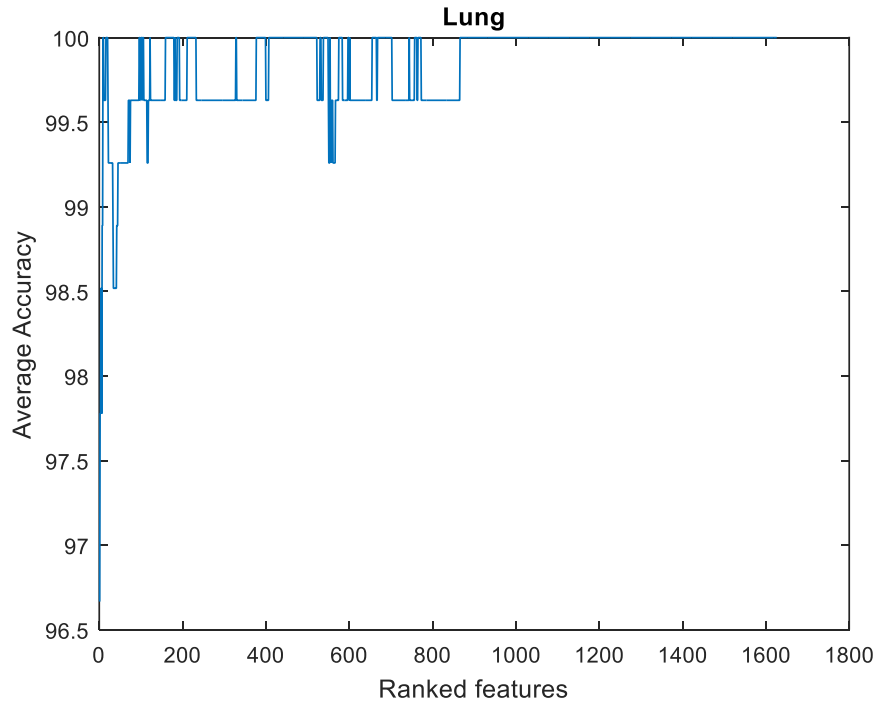


Figure 6.6: Average accuracy of the ranked features for lung cancer data set

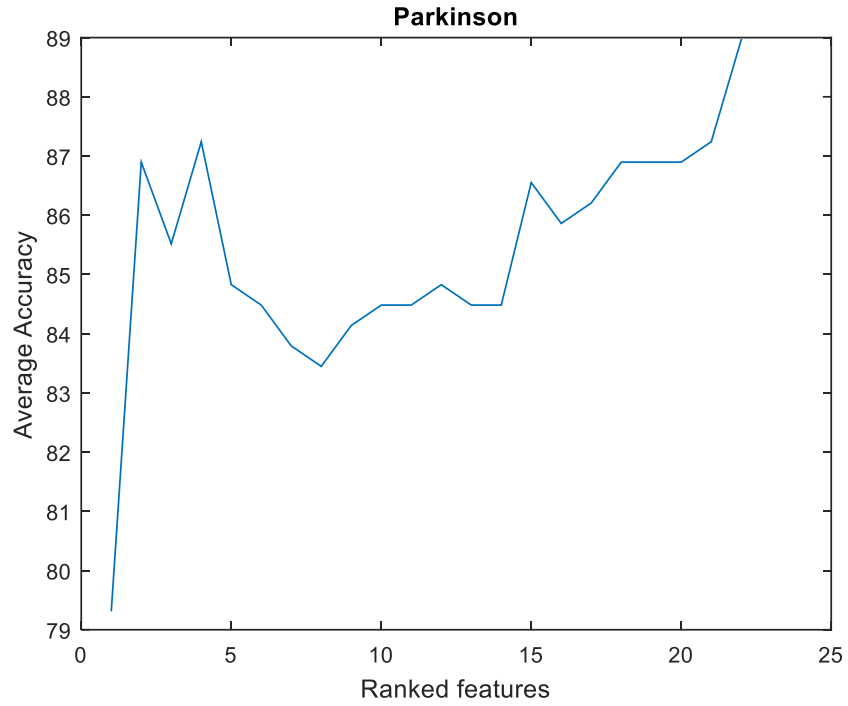


Figure 6.7: Average accuracy of the ranked features for Parkinson data set

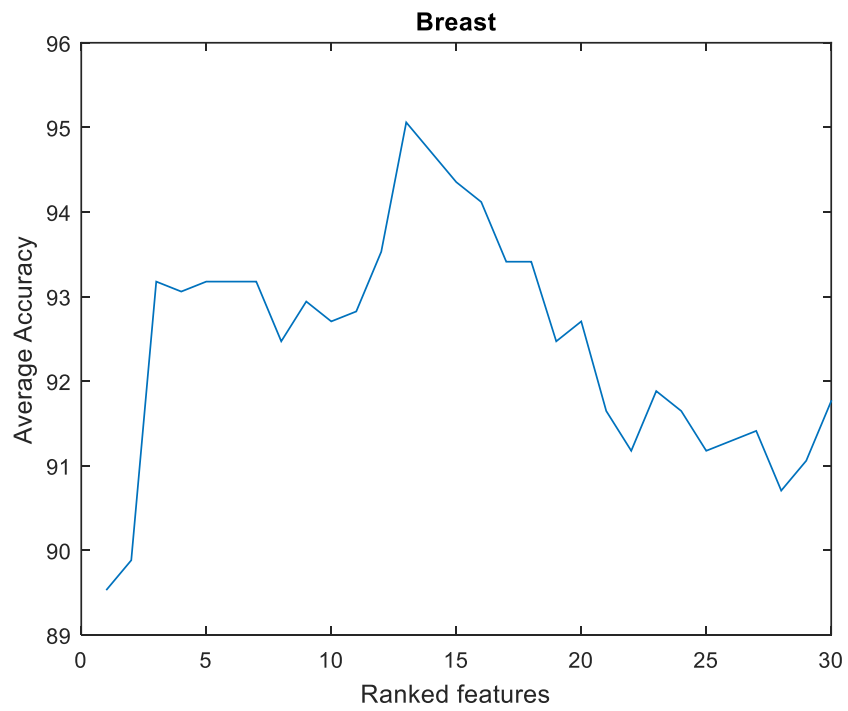


Figure 6.8: Average accuracy of the ranked features for breast cancer data set

From the figures above, the highest average accuracy is the optimal baseline of the data set and the number of features that obtained this optimal baseline were also obtained at the same time. From the figures, it was noticed that when using all the features as a baseline, it usually does not produce a good baseline, as these baselines are lower than the proposed optimal baselines. When all the features are included, meaning that at the same time, the redundancy and noise are included as well. Table 6.9 shows the baseline using the full features and the optimal baseline was obtained using the algorithm in Chapter 3 with the number of features to achieve this optimal baseline. The proposed algorithm shows that the optimal baselines are better than the baseline using the full attributes and the number of features required to obtain the optimal baseline are also lower than the full features. The features obtained using the proposed algorithm is a ranked feature where information contained in the features are also ranked from the most relevant to less relevant. Therefore, this number of features obtained from the proposed algorithm will provide better prediction power in the predictive model.

Table 6.9: The baseline using full features and the optimal baseline with the number of features

Data Set	Baseline	Full features	Optimal baseline	No. of features
Colon	80%	1988	87.78%	202
Leukaemia	96.19%	7128	99.05%	38
Prostate	91.33%	2135	94%	330
Lung	100%	1626	100%	9
Parkinson	88.62%	22	88.62%	22
Breast	91.76%	30	95.06%	13

The number of features obtained using the proposed algorithm plays a vital role in the features selection, and the features selection method should not take more than this number of features to achieve the same accuracy. Therefore, we can have a new guideline on what is the maximum number of features that are allowed in a feature selection. For the past research, they are using the full features to obtain the baseline, therefore no clear guideline on how many features are needed in a predictive model or more clearly what is the maximum number of features that allow a researcher to build a predictive model.

6.5 rMI-SVM algorithm for Binary Data Set

In this section, the rMI-SVM algorithm will be applied in the binary data set to reduce the dimension of the data set by filtering out the redundant features and noise. The redundant features and the noise can be easily detected from the average accuracy graph plotted using the ranked features. The ranked features are ranked according to the information contained, and the higher mutual information score indicated the features contain more information. When more ranked features are added to the predictive model, then more information are also added to the predictive model and the prediction power of the model will be better. A newly added feature to the predictive model, is the performance of the predictive model which will have three phenomena discussed in Chapter 4. When the newly added feature contains new information, then the performance of the predictive model will increase. When the newly added feature contains the same information with the already

selected features, then the performance of the predictive model will remain the same. When a noise is added to the predictive model, the performance of the predictive model will decrease. Therefore, based on the performance of the predictive model, the redundant features and noise can be detected and be filtered out.

Those redundant features, irrelevant features or noise will be filtered out by rMI-SVM algorithm to provide better performance using a lower number of features. Table 6.10 summarises the number of features selected by the rMI-SVM algorithm for six binary data set. Table 6.10 also shows the percentage of the dimension reduction of each binary data set. On average, the percentage of the dimension of the six binary data set has been reduced by about 90%, and the reduction is most significant in the microarray data set.

Table 6.10: Number of features selected by the rMI-SVM algorithm for six binary data set

Data set	No of features selected	Dimension to reduce
Colon	7	99.65%
Leukaemia	5	99.93%
Prostate	3	99.86%
Lung	4	99.75%
Parkinson	7	68.18%
Breast	11	63.33%

The performance of the predictive model using the number of features selected by the rMI-SVM algorithm was tested using four different classifiers, support vector machine classifier (SVM), k - nearest neighbour classifier

(KNN), naïve Bayes classifier (NB) and tree classification classifier (TC). On the other hand, the performance of the predictive model using the regression method and minimal-redundancy-maximal-relevance (mRMR) method on the same number of features was tested. SVM-CV and SVM-Acc represent the average accuracy of the cross-validation of the support vector machine classifier and the average accuracy from the predictive model of the support vector machine classifier. NB-CV and NB-Acc represent the average accuracy of the cross-validation of the naïve Bayes classifier and the average accuracy from the predictive model of the naïve Bayes classifier. KNN-CV and KNN-Acc represent the average accuracy of the cross-validation of the k -nearest neighbour classifier and the average accuracy from the predictive model of the k -nearest neighbour classifier. TC-CV and TC-Acc represent the average accuracy of the cross-validation of the tree classification classifier and the average accuracy from the predictive model of the tree classification classifier. The baseline indicated the cross-validation and average accuracy using full features.

Table 6.11 shows the 10- fold cross-validation, and average accuracy for four different classifiers using full features, seven features using the rMI-SVM algorithm, regression method and mRMR method for colon data set. The 10-fold cross-validation and average accuracy using full features achieved by SVM- linear kernel function classifier was 82.73% and 80%. The 10-fold cross-validation and average performance obtained by using tree classification classifier were 75% and 77.78%. The 10-fold cross-validation and average performance obtained by using naïve Bayes classifier were 56.36% and

74.44%, and the 10-fold cross-validation and average performance obtained by using *k*-nearest neighbour classifier were 73.64% and 73.33%.

The seven features selected by the rMI-SVM algorithm with linear kernel function gave the highest average accuracy with the SVM classifier-linear kernel function were compared with the regression method and mRMR method. The seven features selected by the rMI-SVM algorithm with linear kernel function gave the 10-fold cross-validation and average performance obtained by using SVM classifier - linear kernel function were 88.64% and 87.78%, the seven features selected by regression method gave the 10-fold cross-validation and average performance obtained by using SVM - linear kernel function classifier was 83.18% and 82.22%, and the seven features selected by mRMR method gave the 10-fold cross-validation and average performance obtained by using SVM- linear kernel function classifier was 85.45% and 81.11%.

The seven features selected by the rMI-SVM algorithm with linear kernel function gave the highest average accuracy with the naïve Bayes classifier compared to the regression method and mRMR method. The seven chosen features by the rMI-SVM algorithm with linear kernel function gave the 10-fold cross-validation, and average performance obtained by using naïve Bayes classifier was 88.64% and 86.67%. Whereas, the seven features selected by regression method gave the 10-fold cross-validation, and average performance obtained by using naïve Bayes classifier was 85.45% and 80%.

Finally, the seven chosen features by mRMR method gave the 10-fold cross-validation and average performance obtained by using naïve Bayes classifier was 85.45% and 78.89%.

The seven features selected by the rMI-SVM algorithm with linear kernel function gave the highest average accuracy with the k -nearest neighbour classifier was compared to the regression method and mRMR method. The seven chosen features by the rMI-SVM algorithm with linear kernel function gave the 10-fold cross-validation, and average performance obtained by using k -nearest neighbour classifier was 85.45% and 84.44%. Whereas the seven features selected by regression method gave the 10-fold cross-validation, and average performance obtained by using k -nearest neighbour classifier was 78.64% and 73.33%, and the seven chosen features by mRMR method gave the 10-fold cross-validation and average performance obtained by using k -nearest neighbour classifier was 79.09% and 80%.

The seven features selected by mRMR gave the highest average accuracy with the tree classification classifier were compared with the regression method and rMI-SVM algorithm with a linear kernel function. The seven features chosen by the rMI-SVM algorithm with linear kernel function gave the 10-fold cross-validation, and the average performance obtained by using tree classification classifier was 80% and 75.56%. Whereas, the seven features selected by regression method gave the 10-fold cross-validation, and the average performance obtained by using tree classification classifier was

78.64% and 74.44%. Finally, the seven chosen features by mRMR method gave the 10-fold cross-validation and the average performance obtained by using tree classification classifier was 80% and 85.56%.

Table 6.11: Ten-fold cross-validation and average accuracy for four different classifiers using full features, seven features using the rMI-SVM algorithm, Regression method and mRMR method of the colon data set

	SVM-CV	SVM-Acc	NB-CV	NB-Acc
Baseline	82.7273	80	56.3636	74.4444
rMI-SVM	88.6364	87.7778	88.6364	86.6667
Regression	83.1818	82.2222	85.4546	80
mRMR	85.4546	81.1111	85.4546	78.8889
	KNN-CV	KNN-Acc	TC-CV	TC-Acc
Baseline	73.6364	73.3333	75	77.7778
rMI-SVM	85.4546	84.4444	80	75.5556
Regression	78.6364	73.3333	78.6364	74.4444
mRMR	79.0909	80	80	85.5556

Table 6.12 shows the output of the confusion matrix and the ROC curve of the colon cancer data using full features, seven features selected by the regression method, mRMR, the rMI-SVM algorithm using SVM - linear kernel function classifier with 10-fold cross-validation. The negative “0” indicated the patient with colon cancer while the positive “1” indicated the normal person. When 2000 features were using the predictive model, the accuracy was 81.8%, the prediction speed was about 22 observations per second, and the training time was 7.727 second. The true negative count was 25, the false-positive count was three, the false-negative count was five, and the true positive count was 11; therefore, the recall was 0.89. The seven

features selected by the regression method, the accuracy was 79.5%, the prediction speed was about 1900 observation per second, and the training time was 0.75102 second. The true negative count was 26, the false-positive count was two, the false-negative count was seven, and the true positive count was nine; therefore, the recall was 0.93.

The seven features selected by the mRMR method, the accuracy was 84.1%, the prediction speed was about 2100 observations per second, and the training time was 0.80822 second. The true negative count was 25, the false-positive count was three, the false-negative count was four, and the true positive count was 12; therefore, the recall was 0.89. The seven features selected by the rMI-SVM algorithm, the accuracy was 86.4%, the prediction speed was about 1800 observation per second, and the training time was 0.65991 second. The true negative count was 27, the false-positive count was one, the false-negative count was five, and the true positive count was 11; therefore, the recall was 0.96. The seven features selected by rMI-SVM algorithm achieved the highest accuracy and fastest training time compared to the baseline using full features, regression method, and mRMR method. Therefore, the seven features selected by rMI-SVM algorithm provided a better prediction power.

Table 6.12: Output of the confusion matrix and the ROC curve of colon cancer data using seven features

	TP	TN	FP	FN	Accuracy
Full features	11	25	3	5	81.8%
Regression	9	26	2	7	79.5%
mRMR	12	25	3	4	84.1%
rM-SVM	11	27	1	5	86.4%
	Prediction speed		Training time		AUC
Full features	22 obs/sec		7.727 sec		0.87
Regression	1900 obs/sec		0.75102 sec		0.8
mRMR	2100 obs/sec		0.080822 sec		0.88
rM-SVM	1800 obs/sec		0.65991 sec		0.94

When full features were used in the predictive model, the area under curve was 0.87, but the seven features selected by the regression method, the area under curve was 0.8. The seven features chosen by the mRMR method, the area under curve was 0.88, and the seven features selected by the rMI-SVM algorithm, the area under curve was 0.94. The seven chosen features by rMI-SVM algorithm gave the highest area under curve value indicated that the prediction model built by these seven selected features was better when compared to the baseline using full features, regression method and mRMR method. Therefore, the classifier using the seven features chosen by the rMI-SVM algorithm was rightly predicted. In general, the predictive model that uses the seven features of the colon cancer data set that selected by rMI-SVM algorithm gave the highest accuracy in SVM classifier, naïve Bayes classifier and k -nearest neighbour classifier compared to the regression method and mRMR under the same classifiers and with the same number of features. The area under curve shows that the seven features selected by rMI-SVM

algorithm gives a better predictive model compared to the same number of features selected by the regression method and mRMR method, even when compared to the predictive model that was built using full features. The seven chosen features by rMI-SVM algorithm also achieved the highest recall value compared to the seven features selected by the mRMR method and regression method.

Table 6.13 shows the 10-fold cross-validation and average accuracy for four different classifiers using full features, five features used the rMI-SVM algorithm, regression method and mRMR method for leukaemia data set. By using full features, the 10-fold cross-validation and average performance achieved using SVM -linear kernel function classifier was 97.65% and 96.19%, and the 10-fold cross-validation and average performance obtained using *k*-nearest neighbour classifier were 90.2% and 94.29%. The 10-fold cross-validation and average performance obtained using tree classification classifiers were 88.24% and 88.57%, and the 10-fold cross-validation and average accuracy obtained using naïve Bayes classifier were 93.4% and 83.81%.

The five features selected by the rMI-SVM algorithm with linear kernel function gave the highest average accuracy with the SVM classifier-linear kernel function compared to the regression method and mRMR method. The five features selected by the rMI-SVM algorithm with linear kernel function gave the 10-fold cross-validation and average performance obtained

using SVM classifier - linear kernel function were 94.51% and 100%, the five features selected by regression method gave the 10-fold cross-validation and average performance obtained using SVM - linear kernel function classifier was 96.08% and 98.1%, and the five features selected by mRMR method gave the 10-fold cross-validation and average performance obtained using SVM - linear kernel function classifier was 98.04% and 94.29%.

The five features selected by the rMI-SVM algorithm with linear kernel function and regression method gave the highest average accuracy with the naïve Bayes classifier compared to the mRMR method. The five chosen features by the rMI-SVM algorithm with linear kernel function and regression method gave the 10-fold cross-validation, and average performance obtained using naïve Bayes classifier was 94.51% and 98.1%, and the five features selected by mRMR method gave the 10-fold cross-validation, and average performance obtained using naïve Bayes classifier were 95.69% and 93.33%.

The five features selected by the rMI-SVM algorithm with linear kernel function gave the highest average accuracy with the k -nearest neighbour classifier compared to the regression method and mRMR method. The five chosen features by the rMI-SVM algorithm with linear kernel function gave the 10-fold cross-validation, and average performance obtained by using k -nearest neighbour classifier was 98.04%, and 99.05%, the five features selected by regression method gave the 10-fold cross-validation, and average performance obtained by using k -nearest neighbour classifier was

96.47% and 98.1%, and the five chosen features by mRMR method gave the 10-fold cross-validation, and average performance obtained by using k -nearest neighbour classifier was 96.86% and 95.24%.

The five features selected by rMI-SVM algorithm gave the highest average accuracy with the tree classification classifier compared to the regression method and mRMR method. The five chosen features by the rMI-SVM algorithm with linear kernel function gave the 10-fold cross-validation, and average performance obtained by using tree classification classifier was 87.45%, and 94.29%, the five features selected by regression method gave the 10-fold cross-validation, and average performance obtained by using tree classification classifier were 90.98% and 89.52%, and the five chosen features by mRMR method gave the 10-fold cross-validation and average performance obtained by using tree classification classifier was 89.02% and 93.33%.

Table 6.13: Ten-fold cross-validation and average accuracy for four different classifiers using full features, five features using the rMI-SVM algorithm, Regression method and mRMR method of the leukaemia data set

	SVM-CV	SVM-Acc	NB-CV	NB-Acc
Baseline	97.6471	96.1905	93.3981	83.81
rMI-SVM	94.5098	100	94.5098	98.0952
Regression	96.0784	98.0952	94.5098	98.0952
mRMR	98.0392	94.2857	95.6863	93.3333
	KNN-CV	KNN-Acc	TC-CV	TC-Acc
Baseline	90.1961	94.2857	88.2353	88.5714
rMI-SVM	98.0392	99.0476	87.451	94.2857
Regression	96.4706	98.0952	90.9804	89.5238
mRMR	96.8628	96.8628	89.0196	93.3333

Table 6.14 shows the output of the confusion matrix and the ROC curve of the leukaemia data using full features, five of the features were selected by the regression method, mRMR, the rMI-SVM algorithm using SVM - linear kernel function classifier with 10-fold cross-validation. The negative “0” indicated the patient with acute myeloid leukaemia (AML) while the positive “1” indicated the patient with acute lymphocytic leukaemia (ALL). When 7128 features were used in the predictive model, the accuracy was 98%, the prediction speed was about 4.8 observation per second, and the training time was 36.35 second. The true negative count was 17, the false-positive count was one, the false-negative count was zero, and the true positive count was 33; therefore, the recall was one. The five features selected by the regression method, the accuracy was 94.1%, the prediction speed was about 1800 observation per second, and the training time was 0.72975 second. The true negative count was 16, the false-positive count was 2, the false-negative count was one, and the true positive count was 32; therefore, the recall was 0.97.

The five features when selected by mRMR method, the accuracy was 96.1%, the prediction speed was about 1800 observation per second, and the training time was 0.75703 second. The true negative count was 18, the false-positive count was zero, the false-negative count was two, and the true positive count was 31; therefore, the recall was 0.94. The five features selected by the rMI-SVM algorithm, the accuracy was 96.1%, the prediction speed was about 2200 observation per second, and the training time was 0.7383 second. The true negative count was 16, the false-positive count was two, the false-

negative count was zero, and the true positive count was 33; therefore, the recall was one. The classifier using full features gave the highest accuracy, however, using full features in a classifier was not a practical way in classifying as the training time, and the prediction speed will be much slower and time-consuming. The prediction speed was highest in using five features selected by the regression method and mRMR method, while the fastest training time was using five features selected by the regression method. Although the training was fastest when using five features selected by the regression method, the recall value by using the regression method was 0.97. The best recall value achieved by using full features in the classifier and five features selected by the rMI-SVM algorithm, again using full features was not a practical way in classification. Therefore, the classifier using the five features selected by the rMI-SVM algorithm was rightly predicted.

Table 6.14: Output of the confusion matrix and the ROC curve of leukaemia data using five features

	TP	TN	FP	FN	Accuracy
Full features	33	17	1	0	98.0%
Regression	32	16	2	1	94.1%
mRMR	31	18	0	2	96.1%
rM-SVM	33	16	2	0	96.1%
	Prediction speed		Training time		AUC
Full features	4.8 obs/sec		36.35 sec		1
Regression	1800 obs/sec		0.72975 sec		0.99
mRMR	1800 obs/sec		0.75703 sec		0.99
rM-SVM	2200 obs/sec		0.7383 sec		1

When full features using the predictive model, the area under curve was one of the five features selected by the regression method, hence the area under curve was 0.99. The five chosen features by mRMR method, the area under curve were 0.99, and the five features selected by the rMI-SVM algorithm, the area under curve was one. The highest area under curve was achieved by using full features in the classifier and the five chosen features by rMI-SVM algorithm. However, using full features in classification was not a practical way and the training time was prolonged. Therefore, the five features selected by rMI-SVM algorithm gave a better prediction model compared to the baseline using full features, regression method and mRMR method. Therefore, the classifier using the five features selected by the rMI-SVM algorithm was rightly predicted. In general, the prediction model built by using full features and five features selected by rMI-SVM algorithm gave a better prediction. However, a prediction model that contains many features will make the training time longer. When using five features selected from different methods among regression method, mRMR method and rMI-SVM algorithm, the rMI-SVM algorithm showed a better area under curve value among the other two methods.

Table 6.15 showed the 10-fold cross-validation and average accuracy for three different classifiers using full features, rMI-SVM algorithm, regression method and mRMR method for prostate cancer data set. By using full features, the 10-fold cross-validation and average performance obtained by using SVM - linear kernel function classifier was 91.39% and 91.33%, and the 10-fold cross-validation and average performance obtained by using k -

nearest neighbour classifier was 78.89% and 83.33%. The 10-fold cross-validation and average performance obtained by using tree classification classifier were 81.67% and 75.33%.

The three features selected by the rMI-SVM algorithm with linear kernel function and regression method gave the highest average accuracy with the SVM classifier- linear kernel function compared to the mRMR method. The three features selected by the rMI-SVM algorithm with linear kernel function gave the 10-fold cross-validation and the average performance obtained by using SVM classifier - linear kernel function were 92.22% and 93.33%. The three features selected by regression method gave the 10-fold cross-validation and average performance obtained using SVM - linear kernel function classifier was 91.94% and 93.33%. Lastly, the three features selected by mRMR method gave the 10-fold cross-validation and the average performance was obtained using SVM - linear kernel function classifier was 87.78% and 80%.

The three features selected by the rMI-SVM algorithm with linear kernel function and regression method gave the highest average accuracy with the k -nearest neighbour classifier compared to mRMR method. The three chosen features by the rMI-SVM algorithm with linear kernel function gave the 10-fold cross-validation, and average performance obtained by using k -nearest neighbour classifier was 91.39% and 88%, the three features selected by regression method gave the 10-fold cross-validation, and average

performance obtained by using k -nearest neighbour classifier was 90.56% and 88%, and the three chosen features by mRMR method gave the 10-fold cross-validation, and average performance obtained by using k -nearest neighbour classifier was 87.5% and 85.33%.

The three features selected by rMI-SVM algorithm and regression method gave the highest average accuracy with the tree classification classifier compared to the mRMR method. The three chosen features by the rMI-SVM algorithm with linear kernel function gave the 10-fold cross-validation, and the average performance obtained by using tree classification classifier was 88.33% and 90%. Whereas, the three features selected by regression method gave the 10-fold cross-validation, and the average performance obtained by using tree classification classifier was 89.17% and 90%, and the three chosen features by mRMR method gave the 10-fold cross-validation and average performance obtained by using tree classification classifier was 86.94% and 85.33%. The naïve Bayes classifier did not apply in this data set as this data set has zero variance. The rMI-SVM algorithm and regression method selected the same three features from prostate cancer data set; therefore, the average accuracy was the same by using the selection in rMI-SVM algorithm and regression method.

Table 6.15: Ten-fold cross-validation and average accuracy for four different classifiers using full features, three features using the rMI-SVM algorithm, Regression method and mRMR method of prostate cancer data set

	SVM-CV	SVM-Acc	NB-CV	NB-Acc
Baseline	91.3889	91.3333	nil	nil
rMI-SVM	92.2222	93.3333	nil	nil
Regression	91.9444	93.3333	nil	nil
mRMR	87.7778	88	nil	nil
	KNN-CV	KNN-Acc	TC-CV	TC-Acc
Baseline	78.8889	83.3333	81.6667	75.3333
rMI-SVM	91.3889	88	88.3333	90
Regression	90.5556	88	89.1667	90
mRMR	87.5	85.3333	86.9444	85.3333

Table 6.16 shows the output of the confusion matrix and the ROC curve of the prostate cancer data using full features. Three features were selected by the regression method, mRMR, the rMI-SVM algorithm using SVM - linear kernel function classifier with 10-fold cross-validation. The negative “0” indicated the patient with prostate cancer while the positive “1” indicated the normal person. When 2135 features were used in the predictive model, the accuracy was 88.9%, the prediction speed was about 34 observation per second, and the training time was 13.087 second. The true negative count was 31, the false-positive count was six, the false-negative count was two, and the true positive count was 33; therefore, the recall was 0.84. The three features selected by the regression method, showed the accuracy was 93.1%, the prediction speed was about 3400 observation per second, and the training time was 1.0846 second. The true negative count was

33, the false-positive count was four, the false-negative count was one, and the true positive count was 34; therefore, the recall was 0.89.

The three features selected by mRMR method, the accuracy was 88.9%, the prediction speed was about 3200 observation per second, and the training time was 0.7727 second. The true negative count was 35, the false-positive count was two, the false-negative count was six, and the true positive count was 29; therefore, the recall was 0.95. The three features selected by the rMI-SVM algorithm, the accuracy was 93.1%, the prediction speed was about 3400 observation per second, and the training time was 1.0846 second. The true negative count was 33, the false-positive count was four, the false-negative count was one, and the true positive count was 34; therefore, the recall was 0.971. The classifier using the three selected features by regression method and rMI-SVM algorithm gave the highest accuracy. The prediction speed and the training time were fastest in using three features selected by mRMR method. The prediction speed was fastest when using three features selected by regression method and rMI-SVM algorithm.

In comparison, the training time was fastest when using the three features selected by mRMR method. Still, the recall value by using the mRMR method was lower than the regression method and rMI-SVM algorithm. The highest accuracy and recall value were achieved by the classifier using the three selected features by regression method and rMI-SVM algorithm. The regression method and rMI-SVM algorithm chose the same three features.

Table 6.16: Output of the confusion matrix and the ROC curve of prostate cancer data using three features

	TP	TN	FP	FN	Accuracy
Full features	33	31	6	2	88.9%
Regression	34	33	4	1	93.1%
mRMR	29	35	2	6	88.9%
rM-SVM	34	33	4	1	93.1%

	Prediction speed	Training time	AUC
Full features	34 obs/sec	13.087 sec	0.94
Regression	3400 obs/sec	1.0846 sec	0.97
mRMR	3200 obs/sec	0.7727 sec	0.96
rM-SVM	3400 obs/sec	1.0846 sec	0.97

When full features were used in the predictive model, the area under curve was 0.94 and with the three features selected by the regression method, the area under curve was 0.97. The three chosen features by mRMR method, the area under curve were 0.96, and the three features selected by the rMI-SVM algorithm, the area under curve was 0.97. The three features that were selected by regression method and rMI-SVM algorithm achieved the highest area under curve as obtained in the classifier compared to the three chosen features by mRMR method and using full features in the classifier. Therefore, the classifier using the three features selected by the regression method and rMI-SVM algorithm was rightly predicted. In general, the three chosen features that used regression method and rMI-SVM algorithm performed well in the support vector classifier as the same features were selected between regression method and rMI-SVM algorithm compared to using full features and three features selected by mRMR method.

Table 6.17 shows the 10-fold cross-validation and average accuracy for three different classifiers using full features, rMI-SVM algorithm, regression method and mRMR method for lung cancer data set. By using full features, the 10-fold cross-validation and average performance were obtained by using SVM - linear kernel function classifier was 99.06% and 100%, and the 10-fold cross-validation and average performance were obtained by using k -nearest neighbour classifier was 98.58% and 99.63%. The 10-fold cross-validation and average performance were obtained by using tree classification classifier was 94.8% and 94.81%.

The four features selected by the rMI-SVM algorithm with linear kernel function gave the highest average accuracy with the SVM classifier-linear kernel function compared to the regression and mRMR method. The four features selected by the rMI-SVM algorithm with linear kernel function gave the 10-fold cross-validation, and average performance obtained by using SVM classifier - linear kernel function was 99.06% and 100%, whereas the four features selected by regression method, and mRMR method gave the 10-fold cross-validation and average performance obtained by using SVM - linear kernel function classifier was 98.74% and 99.26%.

The four features selected by the rMI-SVM algorithm with linear kernel function gave the highest average accuracy with the k -nearest neighbour classifier compared to the regression mRMR method. The four features chosen by the rMI-SVM algorithm with linear kernel function gave

the 10-fold cross-validation, and average performance obtained by using k -nearest neighbour classifier was 98.58% and 99.63%. Whereas, the four features selected by regression method gave the 10-fold cross-validation, and average performance obtained by using k -nearest neighbour classifier was 97.17% and 99.26%, and the four chosen features by mRMR method gave the 10-fold cross-validation, and average performance obtained by using k -nearest neighbour classifier was 97.8% and 98.89%.

The four features selected by rMI-SVM algorithm gave the highest average accuracy with the tree classification classifier compared to the regression method and mRMR method. The four chosen features by the rMI-SVM algorithm with linear kernel function gave the 10-fold cross-validation, and average performance obtained by using tree classification classifier was 95.59% and 97.04%, whereas the four features selected by regression method gave the 10-fold cross-validation, and average performance obtained by using tree classification classifier were 96.54% and 96.67%, and the four chosen features by mRMR method gave the 10-fold cross-validation and average performance obtained by using tree classification classifier was 96.69% and 95.93%. The naïve Bayes classifier did not apply in this data set as this data set has zero variance.

Table 6.17: Ten-fold cross-validation and average accuracy for four different classifiers using full features, four features using the rMI-SVM algorithm, Regression method and mRMR method of lung cancer data set

	SVM-CV	SVM-Acc	NB-CV	NB-Acc
Baseline	99.0551	100	nil	nil
rMI-SVM	99.0551	100	nil	nil
Regression	98.7402	99.2593	nil	nil
mRMR	98.7402	99.2593	nil	nil
	KNN-CV	KNN-Acc	TC-CV	TC-Acc
Baseline	98.5827	99.6296	94.8032	94.8148
rMI-SVM	98.5827	99.6296	95.5906	97.0370
Regression	97.1654	99.2593	96.5354	96.6667
mRMR	97.7953	98.8889	96.6929	95.9259

Table 6.18 shows the output of the confusion matrix and the ROC curve of the lung cancer data using full features. Four features were selected by the regression method, mRMR, the rMI-SVM algorithm using SVM - linear kernel function classifier with 10-fold cross-validation. The negative “0” indicated the malignant pleural mesothelioma (MPM) while the positive “1” indicated the adenocarcinoma (AD). When 1626 features are used in the predictive model, the accuracy was 99.2%, the prediction speed was about 130 observation per second, and the training time was 9.5591 second. The true negative count was 21, the false-positive count was one, the false-negative count was zero, and the true positive count was 105; therefore, the recall was one. The four features selected by the regression method, the accuracy was 99.2%, the prediction speed was about 6000 observation per second, and the training time was 0.91766 second. The true negative count was 21, the false-

positive count was one, the false-negative count was zero, and the true positive count was 105; therefore, the recall was one.

The accuracy of the four features selected by mRMR method was 99.2%, the prediction speed was about 3500 observation per second, and the training time was 0.84527 second. The true negative count was 21, the false-positive count was one, the false-negative count was zero, and the true positive count was 105; therefore, the recall was one. For the four features selected by the rMI-SVM algorithm, the accuracy was 99.2%, the prediction speed was about 3100 observation per second, and the training time was 0.89517 second. The true negative count was 21, the false-positive count was one, the false-negative count was zero, and the true positive count was 105; therefore, the recall was one. The classifier using the four selected features by regression method, mRMR method and rMI-SVM algorithm gave the same accuracy with the classifier using full features. The prediction speed was fastest when using the four features selected by the regression method while the fastest training was achieved by using the four features selected by mRMR method. The recall value was the same in the classifier using full features, four features selected by regression method, mRMR method and rMI-SVM algorithm. The mRMR method has three selected features that are similar with the features selected by the regression method. Therefore, the regression method, mRMR method and rMI-SVM algorithm selected four with almost the same features in the lung cancer data set.

Table 6.18: Output of the confusion matrix and the ROC curve of lung cancer data using four features

	TP	TN	FP	FN	Accuracy
Full features	105	21	1	0	99.2%
Regression	105	21	1	0	99.2%
mRMR	105	21	1	0	99.2%
rM-SVM	105	21	1	0	99.2%
	Prediction speed	Training time	AUC		
Full features	130 obs/sec	9.5591 sec	1		
Regression	6000 obs/sec	0.91766 sec	1		
mRMR	3500 obs/sec	0.84527 sec	1		
rM-SVM	3100 obs/sec	0.89517 sec	1		

When full features are using the predictive model, the area under curve was one, and the four features selected by the regression method, mRMR method and rMI-SVM algorithm, the area under curve was one. Since the four selected features were almost the same among the regression method, mRMR method and rMI-SVM algorithm, therefore the receiver operating characteristic curve and area under curve were the same. In general, the four selected features by using the regression method, mRMR and rMI-SVM algorithm were well performed in the support vector classifier because almost the same features were selected compared to using full features in the classifier.

Table 6.19 shows the 10-fold cross-validation and average accuracy for four different classifiers using full features, rMI-SVM algorithm, regression method and mRMR method for Parkinson data set. By using full features, the

10-fold cross-validation and average performance obtained by using k -nearest neighbour classifier were 93.14% and 90.69%, and the 10-fold cross-validation and average performance obtained by using SVM - Gaussian kernel function classifier was 89.05% and 88.62%. The 10-fold cross-validation and average performance obtained by using tree classification classifier were 84.67% and 78.28%, and the 10-fold cross-validation and average performance obtained by using naïve Bayes classifier was 70.8% and 65.86%.

The seven features selected by the rMI-SVM algorithm with Gaussian kernel function gave the highest average accuracy with the SVM classifier-Gaussian kernel function compared to the regression method and mRMR method. The seven features selected by the rMI-SVM algorithm with Gaussian kernel function gave the 10-fold cross-validation and average performance obtained by using SVM classifier - Gaussian kernel function were 88.61% and 90.69%, whereas the seven features selected by regression method gave the 10-fold cross-validation and average performance obtained by using SVM - Gaussian kernel function classifier was 85.99% and 84.83%, and the seven features selected by mRMR method gave the 10-fold cross-validation and average performance obtained by using SVM - Gaussian kernel function classifier was 87.59% and 86.21%.

The seven features selected by mRMR gave the highest average accuracy with the naïve Bayes classifier compared to the regression method and rMI-SVM algorithm. The seven chosen features by the rMI-SVM

algorithm with Gaussian kernel function gave the 10-fold cross-validation, and average performance obtained by using the naïve Bayes classifier was 71.97% and 72.07%, whereas the seven features selected by regression method gave the 10-fold cross-validation, and average performance obtained by using naïve Bayes classifier was 78.25% and 73.79%, and the seven chosen features by mRMR method gave the 10-fold cross-validation and average performance obtained by using naïve Bayes classifier was 75.62% and 74.48%.

The seven features selected by the rMI-SVM algorithm with Gaussian kernel function gave the highest average accuracy with the k -nearest neighbour classifier compared to the regression method and mRMR method. The seven features chosen by the rMI-SVM algorithm with Gaussian kernel function gave the 10-fold cross-validation, and average performance obtained by using k -nearest neighbour classifier was 90.37% and 86.21%, whereas the seven features selected by regression method gave the 10-fold cross-validation, and average performance obtained by using k -nearest neighbour classifier was 91.67% and 81.03%, and the seven chosen features by mRMR method gave the 10-fold cross-validation and average performance obtained by using k -nearest neighbour classifier was 88.61% and 83.79%.

The seven features selected by mRMR method gave the highest average accuracy with the tree classification classifier compared with the regression method and rMI-SVM algorithm. The seven features chosen by the rMI-SVM algorithm with Gaussian kernel function gave the 10-fold cross-

validation, and the average performance obtained by using tree classification classifier was 87.01% and 81.03%, whereas the seven features selected by regression method gave the 10-fold cross-validation and average performance obtained by using tree classification classifier was 83.07% and 80.34% and the seven chosen features by mRMR method gave the 10-fold cross-validation and average performance obtained by using tree classification classifier was 86.72% and 82.41%.

Table 6.19: Ten-fold cross-validation and average accuracy for four different classifiers using full features, seven features using the rMI-SVM algorithm, Regression method and mRMR method of Parkinson data set

	SVM-CV	SVM-Acc	NB-CV	NB-Acc
Baseline	89.0511	88.6207	70.8029	65.8621
rMI-SVM	88.6131	90.6897	71.9708	72.069
Regression	85.9854	84.8276	78.2482	73.7931
mRMR	87.5912	86.2069	75.6204	74.4828
	KNN-CV	KNN-Acc	TC-CV	TC-Acc
Baseline	93.1387	90.6897	84.6715	78.2759
rMI-SVM	90.365	86.2069	87.0073	81.0345
Regression	91.6788	81.0345	83.0657	80.3448
mRMR	88.6131	83.7931	86.7153	82.4138

Table 6.20 shows the output of the confusion matrix and the ROC curve of the Parkinson data using full features, seven features were selected by the regression method, mRMR, rMI-SVM algorithm using SVM with Gaussian kernel function classifier with 10-fold cross-validation. The negative “0” indicated the patient with Parkinson disease while the positive “1” indicated the normal person. When full features using the predictive model,

the accuracy was 85.4%, the prediction speed was about 4600 observation per second, and the training time was 0.69047 second. The true negative count was 16, the false-positive count was 18, the false-negative count was two, and the true positive count was 101; therefore, the recall was 0.47. The seven features selected by the regression method, the accuracy was 82.5%, the prediction speed was about 4600 observation per second, and the training time was 0.80413 second. The true negative count was 13; the false-positive count was 21, the false-negative count was three, and the true positive count was 100; therefore, the recall was 0.38.

The seven features selected by mRMR method, the accuracy was 85.4%, the prediction speed was about 5800 observation per second, and the training time was 0.7306 second. The true negative count was 15, the false-positive count was 19, the false-negative count was one, and the true positive count was 102; therefore, the recall was 0.44. The seven features selected by the rMI-SVM algorithm, the accuracy was 86.9%, the prediction speed was about 5400 observation per second, and the training time was 0.73833 second. The true negative count was 16, the false-positive count was 18, the false-negative count was zero, and the true positive count was 103; therefore, the recall was 0.47. The classifier using seven features selected by rMI-SVM algorithm gave the highest accuracy. The prediction speed was fastest using seven features selected by mRMR, and training time was fastest using full features. However, using full features in a classifier was not a practical way of doing classification. The best recall value was achieved using seven features selected by the rMI-SVM algorithm and using full features in the classifier.

However, using full features in the prediction model was not a practical way. The prediction speed of the rMI-SVM algorithm was slightly slower than the prediction speed when using the seven features selected by mRMR.

Table 6.20: Output of the confusion matrix and the ROC curve of Parkinson data using seven features

	TP	TN	FP	FN	Accuracy
Full features	101	16	18	2	85.4%
Regression	100	13	21	3	82.5%
mRMR	102	15	19	1	85.4%
rM-SVM	103	16	18	0	86.9%

	Prediction speed	Training time	AUC
Full features	4600 obs/sec	0.69074 sec	0.89
Regression	4600 obs/sec	0.80413 sec	0.89
mRMR	5800 obs/sec	0.7306 sec	0.85
rM-SVM	5400 obs/sec	0.73833 sec	0.92

When full features were using the predictive model, the area under curve was 0.89 and the seven features selected by the regression method, the area under curve was 0.89. For the seven features selected by mRMR method, the area under curve was 0.85, and the seven features chosen by rMI-SVM algorithm, the area under curve was 0.92. The highest area under curve achieved by using seven features was selected by rMI-SVM algorithm in the classifier. Therefore, the classifier using the seven features chosen by the rMI-SVM algorithm was rightly predicted. In general, the prediction model built by seven features selected by rMI-SVM algorithm gave a better prediction with the area under curve was 0.92 compared to the prediction model built by

seven features selected by regression method or mRMR method or when using full features.

Table 6.21 shows the 10-fold cross-validation and average accuracy for four different classifiers using full features, rMI-SVM algorithm, regression method and mRMR method for breast cancer data set. By using full features, the 10-fold cross-validation and average performance obtained by using k -nearest neighbour classifier was 95.14% and 92.35%, whereas the 10-fold cross-validation and average accuracy obtained by using SVM- linear kernel function classifier was 97.39% and 91.76%. The 10-fold cross-validation and average performance obtained by using tree classification classifier was 93.83% and 89.29%, and the 10-fold cross-validation and average performance obtained by using naïve Bayes classifier was 93.58% and 84.24%.

The eleven features selected by the rMI-SVM algorithm with linear kernel function gave the highest average accuracy with the SVM classifier- linear kernel function compared to the regression method and mRMR method. The eleven features selected by the rMI-SVM algorithm with linear kernel function gave the 10-fold cross-validation and the average performance obtained by using SVM classifier - linear kernel function was 97.59% and 96.59%, whereas the eleven features selected by regression method gave the 10-fold cross-validation and the average performance obtained by using SVM - linear kernel function classifier was 94.24% and 92.82% and the eleven features selected by mRMR method gave the 10-fold cross-validation and

average performance obtained by using SVM - linear kernel function classifier was 97.24% and 93.29%.

The eleven features selected by the rMI-SVM algorithm with linear kernel function give the highest average accuracy with the naïve Bayes classifier compared to the regression method and mRMR method. The eleven features chosen by the rMI-SVM algorithm with linear kernel function gave the 10-fold cross-validation, and the average performance obtained by using naïve Bayes classifier was 94.19% and 91.53%, whereas the eleven features selected by regression method gave the 10-fold cross-validation, and the average performance obtained by using naïve Bayes classifier was 94.19% and 90.12%, and the eleven features selected by mRMR method gave the 10-fold cross-validation and average performance obtained by using naïve Bayes classifier was 95.49% and 90.82%.

The eleven features selected by mRMR method gave the highest average accuracy with the k -nearest neighbour classifier compared to the regression method and rMI-SVM algorithm. The eleven chosen features by the rMI-SVM algorithm with linear kernel function gave the 10-fold cross-validation, and average performance obtained by using k -nearest neighbour classifier was 95.13% and 93.53%, the eleven features selected by regression method gave the 10-fold cross-validation, and average performance obtained by using k -nearest neighbour classifier was 93.93% and 91.53% and the eleven features selected by mRMR method gave the 10-fold cross-validation

and average performance obtained by using k -nearest neighbour classifier was 96.29% and 93.76%.

The eleven features selected by rMI-SVM algorithm gave the highest average accuracy with the tree classification classifier compared to the regression method and mRMR. The eleven features chosen by the rMI-SVM algorithm with linear kernel function gave the 10-fold cross-validation, and average performance obtained by using tree classification classifier was 94.09% and 89.18%, whereas the eleven features selected by regression method gave the 10-fold cross-validation, and the average performance obtained by using tree classification classifier was 92.63% and 88.94%, and the eleven features selected by mRMR method gave the 10-fold cross-validation and the average performance obtained by using tree classification classifier was 94.64% and 88.71%.

Table 6.21: Ten-fold cross-validation and average accuracy for four different classifiers using full features, eleven features using the rMI-SVM algorithm, Regression method and mRMR method of breast cancer data set

	SVM-CV	SVM-Acc	NB-CV	NB-Acc
Baseline	97.3935	91.7647	93.584	84.2353
rMI-SVM	97.594	96.5882	94.1855	91.5294
Regression	94.2356	92.8235	94.1855	90.1177
mRMR	97.2431	93.2941	95.4887	90.8235
	KNN-CV	KNN-Acc	TC-CV	TC-Acc
Baseline	95.1378	92.3529	93.8346	89.2941
rMI-SVM	95.1378	93.5294	94.0852	89.1765
Regression	93.9348	91.5294	92.6316	88.9412
mRMR	96.2907	93.7647	94.6366	88.7059

Table 6.22 shows the output of the confusion matrix and the ROC curve of the breast cancer data using all features, eleven features were selected by the regression method, mRMR, the rMI-SVM algorithm using SVM - linear kernel function classifier with 10-fold cross-validation. The negative “0” indicated the malignant while the positive “1” indicated the benign. When full features used the predictive model, the accuracy was 93.7%, the prediction speed was about 15000 observation per second, and the training time was 0.77129 second. The true negative count was 138, the false-positive count was eleven, the false-negative count was 14, and the true positive count was 236; therefore, the recall was 0.93. For the eleven features selected by the regression method, the accuracy was 94%, the prediction speed was about 12000 observation per second, and the training time was 0.79358 second. The true negative count was 134, the false-positive count was 15, the false-

negative count was nine, and the true positive count was 241; therefore, the recall was 0.9.

For the eleven features selected by mRMR method, the accuracy was 96.2%, the prediction speed was about 15000 observation per second, and the training time was 0.75591 second. The true negative count was 139, the false-positive count was 10, the false-negative count was five, and the true positive count was 245; therefore, the recall was 0.93. For the eleven features selected by the rMI-SVM algorithm, the accuracy was 97%, the prediction speed was about 17000 observation per second, and the training time was 0.76167 second. The true negative count was 142, the false-positive count was seven, the false-negative count was five, and the true positive count was 245; therefore, the recall was 0.95. The classifier using eleven features selected by rMI-SVM algorithm gave the highest accuracy. The prediction speed was fastest in classifier using eleven features selected by rMI-SVM algorithm while the training time was fastest in classifier using eleven features selected by mRMR method. The best recall value achieved using eleven features was selected by mRMR method and rMI-SVM algorithm in the classifier.

Table 6.22: Output of the confusion matrix and the ROC curve of breast cancer data using eleven features

	TP	TN	FP	FN	Accuracy
All features	236	138	11	14	93.7%
Regression	241	134	15	9	94.0%
mRMR	245	139	10	5	96.2%
rM-SVM	245	142	7	5	97%

	Prediction speed	Training time	AUC
All features	15000 obs/sec	0.77129 sec	0.93
Regression	12000 obs/sec	0.79358 sec	0.98
mRMR	15000 obs/sec	0.75591 sec	0.99
rM-SVM	17000 obs/sec	0.76167 sec	0.99

When full features used the predictive model, the area under curve was 0.93 and the eleven features selected by the regression method, the area under curve was 0.98. The eleven features chosen by mRMR method, the area under curve was 0.99, and the eleven features selected by the rMI-SVM algorithm, the area under curve was 0.99. The highest area under curve achieved by the classifier using eleven features was selected by mRMR method and rMI-SVM algorithm. Therefore, the classifier using the eleven features selected by mRMR method and the rMI-SVM algorithm was rightly predicted. In general, the prediction model built by eleven features selected by mRMR method and rMI-SVM algorithm gave a better prediction compared to the prediction model built by full features and eleven features selected by the regression method.

6.6 Evaluation on Four Different Classifiers for Multiclass Data Set

Tables 6.23- 6.26 show the average accuracy of SVM cross-validation and average accuracy of SVM predictive model obtained from the four types of kernel function of SVM classifier of the multiclass data set.

Table 6.23: Average accuracy of 2-fold cross-validation and predictive model on four different kernel functions of classifier of skin cancer data set

Type of classifier	SVM Cross-validation	SVM Accuracy
SVM Linear	60	66.6667
SVM Quadratic	20	33.3333
SVM Cubic	25	40
SVM RBF	41.6667	40

From Table 6.23, the 2-fold cross-validation was used in the SVM classifier with different kernel functions. The skin cancer data set achieved the highest average accuracy of 66.67% by using the SVM classifier with linear kernel function, followed by SVM classifier with cubic kernel function and with radial basis function kernel function gave an average accuracy of 40%, and the lowest average accuracy was 33.33% by using SVM classifier with a quadratic kernel function.

Table 6.24: Average accuracy of 2-fold cross-validation and predictive model on four different kernel functions of classifier of the lymphoma data set

Type of classifier	SVM Cross-validation	SVM Accuracy
SVM Linear	83.6111	98.3333
SVM Quadratic	53.0556	80
SVM Cubic	25	70
SVM RBF	58.0556	67.5

From Table 6.24, the 2-fold cross-validation was used in the SVM classifier with different kernel functions. The lymphoma data set achieved the highest average accuracy of 98.33% by using the SVM classifier with linear kernel function, followed by SVM classifier with quadratic kernel function with the average accuracy of 80%, SVM classifier with cubic kernel function with the average accuracy of 70% and the lowest average accuracy was 67.5% by using SVM classifier with radial basis function kernel function.

Table 6.25: Average accuracy of 3-fold cross-validation and predictive model on four different kernel functions of classifier of lung cancer data set

Type of classifier	SVM Cross-validation	SVM Accuracy
SVM Linear	51.6667	67.5
SVM Quadratic	37.5	45
SVM Cubic	37.5	55
SVM RBF	35.8333	35

From Table 6.25, the 3-fold cross-validation was used in the SVM classifier with different kernel functions. The lung cancer data set achieved the highest average accuracy of 67.5% by using the SVM classifier with linear

kernel function, followed by SVM classifier with cubic kernel function with the average accuracy of 55%, SVM classifier with quadratic kernel function with the average accuracy of 45% and the lowest average accuracy was 35% by using SVM classifier with radial basis function kernel function.

Table 6.26: Average accuracy of 10-fold cross-validation and predictive model on four different kernel functions of classifier of the handwriting data set

Type of classifier	SVM Cross-validation	SVM Accuracy
SVM Linear	98.2286	98.2
SVM Quadratic	94.3714	95.6
SVM Cubic	24.1	29.8667
SVM RBF	97.8	97.8

From Table 6.26, the 10-fold cross-validation was used in SVM classifier with different kernel functions. The handwriting data set achieved the highest average accuracy of 98.2% by using the SVM classifier with linear kernel function, followed by SVM classifier with radial basis function kernel function with the average accuracy of 97.8%, SVM classifier with quadratic kernel function with the average accuracy of 95.6% and the lowest average accuracy was 29.87% by using SVM classifier with a cubic kernel function. Therefore, based on the average accuracy analysis on SVM classifier with a different kernel function, the skin cancer data set, lymphoma data set, lung cancer data set, and handwriting data set use SVM-linear in the rMI-SVM algorithm later.

6.7 Optimal Baseline of the Multiclass Data

The optimal baseline for multiclass data will be obtained using the algorithm provided in Chapter 3. Figures 6.9-6.12 show the average accuracy of the ranked features for skin cancer data set, the lymphoma data set, lung cancer data set, and handwriting data set.

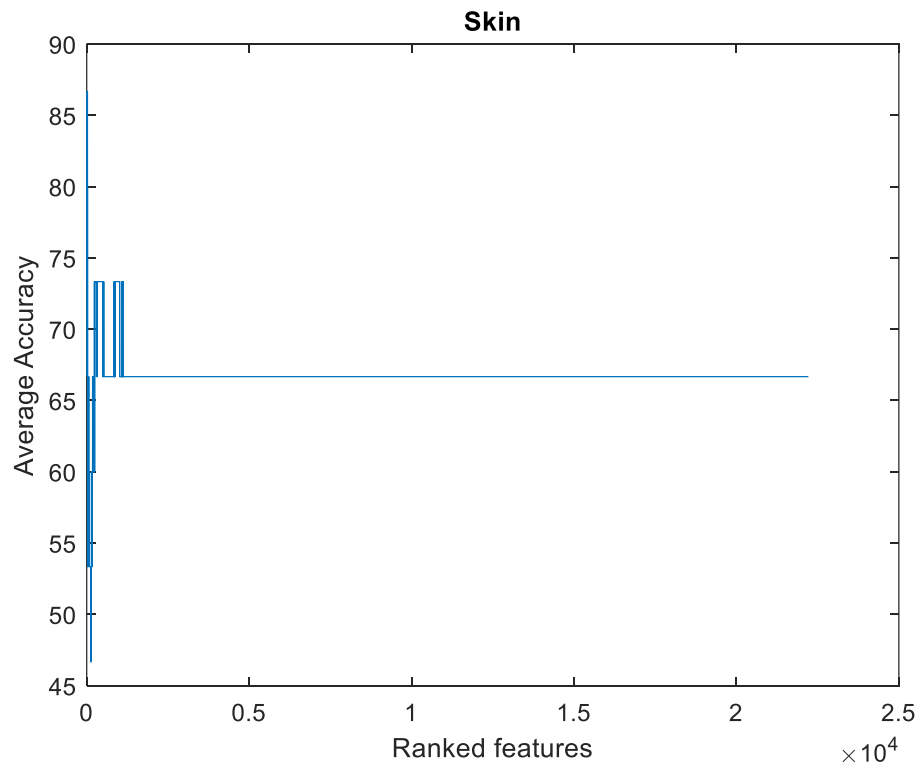


Figure 6.9: Average accuracy of the ranked features for skin cancer data set

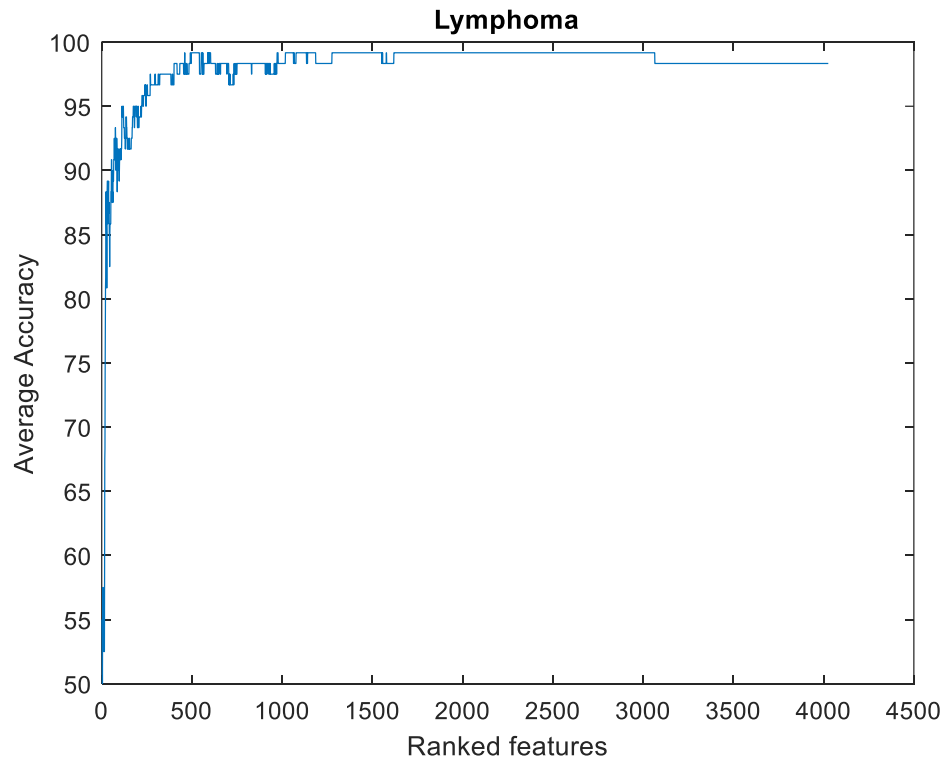


Figure 6.10: Average accuracy of the ranked features for the lymphoma data set

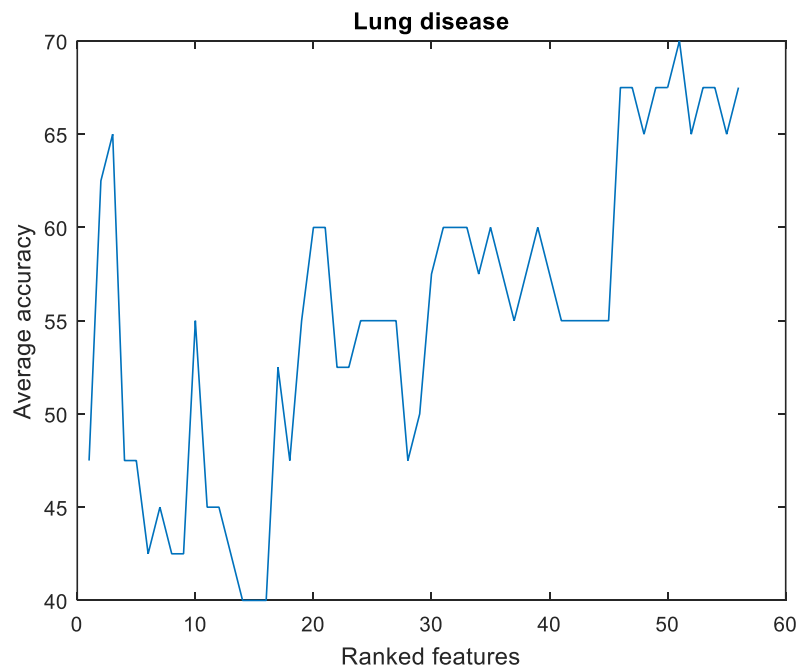


Figure 6.11: Average accuracy of the ranked features for lung cancer data set

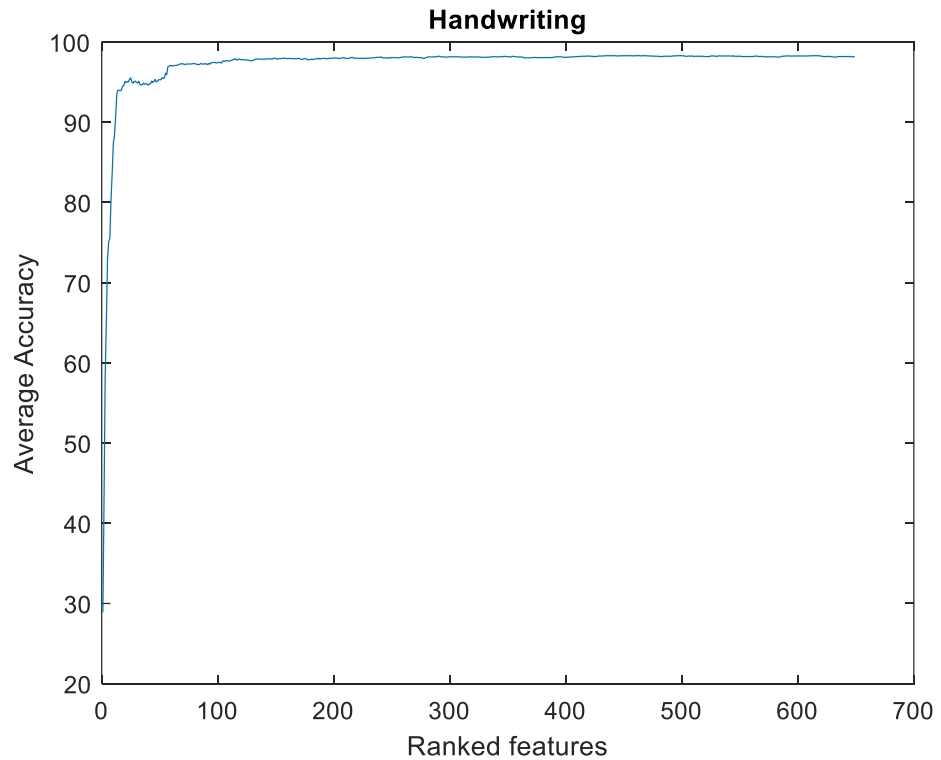


Figure 6.12: Average accuracy of the ranked features for the handwriting data set

From the figures above, the highest average accuracy is the optimal baseline of the data set and the number of features that obtained this optimal baseline were also obtained at the same time. From the figures, it was noticed that when using all the features as a baseline, it usually does not produce a good baseline, as these baselines are lower than the proposed optimal baselines. When all the features are included, the redundant features and noise will be added at the same time. Table 6.27 shows the baseline using full features and the optimal baseline obtained using the algorithm in Chapter 3 with the number of features to achieve this optimal baseline. The proposed algorithm shows that the optimal baselines are better than the baseline using full features and the number of features required to obtain the optimal baseline

are also lower than the full features. The features obtained using the proposed algorithm is a ranked feature where information contained in the features are also ranked from the most relevant to less relevant. Therefore, the number of features obtained from the proposed algorithm will provide better prediction power in the predictive model.

Table 6.27: The baseline using full features and the optimal baseline with the number of features for the multiclass data set

Data Set	Baseline	Full features	Optimal baseline	No. of features
Skin	66.67%	22215	86.67%	2
Lymphoma	98.33%	4026	99.17%	460
Lung	67.5%	56	70%	51
Handwriting	98.2%	649	98.3%	434

The number of features obtained using the proposed algorithm plays a vital role in features selection, and a features selection method should not take more than this number of features to achieve the same accuracy. Therefore, we can have a new guideline on what is the maximum number of features allowed in a features selection. In the previous research, they used the full features to obtain the baseline, as there is no clear guideline to show how many features are needed in a predictive model or to be more specific what is the maximum number of features that a researcher is allowed in order to build a predictive model.

6.8 rMI-SVM algorithm for Multiclass Data Set

In this section, the rMI-SVM algorithm will be applied in the multiclass data set to reduce the dimension of the data set by filtering out the redundant features and noise. The redundant features and noise can be easily detected from the average accuracy graph plotted using the ranked features. The ranked features are ranked according to the information contained, and the higher mutual information score indicated that the features contain more information. When more ranked features are added to the predictive model, the prediction power of the model will then be better. A newly added feature to the predictive model, is that the performance of the predictive model will have the three phenomena as discussed in Chapter 4. When the newly added feature contains new information, then the performance of the predictive model will increase. When the newly added feature contains the same information with the already selected features, then the performance of the predictive model will remain the same. When a noise added to the predictive model, the performance of the predictive model will decrease. Therefore, based on the performance of the predictive model, the redundant features and noise can be detected and filtered out.

Those redundant features, irrelevant features or noise will be filtered out by rMI-SVM algorithm to provide better performance using a lower number of features. Table 6.28 summarises the number of features selected by the rMI-SVM algorithm for six binary data set. Also, Table 6.28 shows the percentage of the dimension reduction of each binary data set. On average, the

percentage of the dimension of the six binary data set has been reduced by about 96%, and the reduction is most significant in the microarray data set.

Table 6.28: Number of features selected by the rMI-SVM algorithm for six binary data set

Data set	No of features selected	Dimension to reduce
Skin	4	99.98%
Lymphoma	22	99.45%
Lung	5	91.07%
Handwriting	50	92.3%

The performance of the predictive model using the number of features selected by the rMI-SVM algorithm has been tested using SVM. On the other hand, the performance of the predictive model using regression method on the same number of features was also tested. The k -fold cross-validation of the support vector machine classifier and the average accuracy of the predictive model was shown in the next section. SVM-CV and SVM-Acc will represent the average accuracy of the cross-validation of the SVM classifier and the average accuracy from the predictive model of the SVM classifier. The baseline indicated the cross-validation and average accuracy using full features.

Table 6.29 shows the 2-fold cross-validation and average accuracy for the SVM classifier using full features, rMI-SVM algorithm and regression method. The four features selected by the rMI-SVM algorithm with linear kernel function gave the highest average accuracy with the SVM classifier-

linear kernel function compared to the regression method and classifier using full features. By using full features, the 2-fold cross-validation and average accuracy achieved by using SVM - linear kernel function classifier were 63.33% and 66.67%. The four features selected by the rMI-SVM algorithm with linear kernel function gave the 2-fold cross-validation and average accuracy achieved by using SVM classifier - linear kernel function were 88.33% and 100%. The four features selected by the regression method gave the 2-fold cross-validation, and average accuracy achieved by using SVM - linear kernel function classifier was 60% and 86.67%.

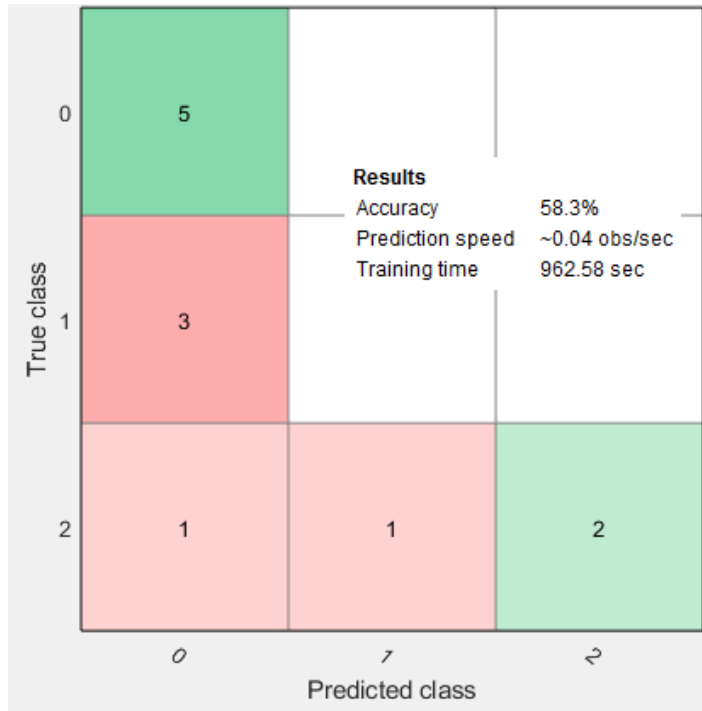
Table 6.29: Two-fold cross-validation and average accuracy for SVM classifier using full features, rMI-SVM algorithm, and regression method for skin cancer data set

	SVM-CV	SVM-Acc
Baseline	63.3333	66.6667
rMI-SVM	83.3333	100
Regression	60	86.6667

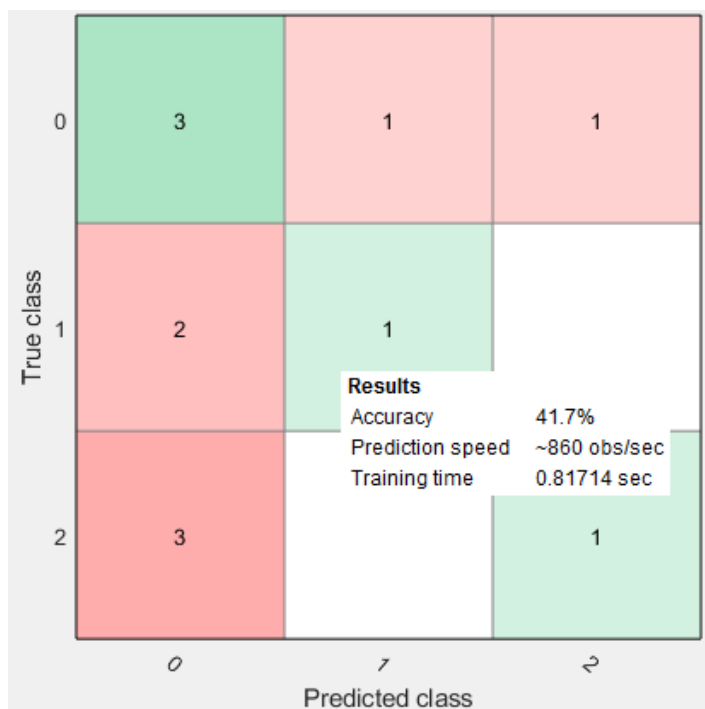
The confusion matrix and the receiver operating characteristic curve (ROC curve) were built to show the distribution of the true class versus the predicted class for the full features. Four of the features selected by regression method and rMI-SVM algorithm with linear kernel function used the SVM classifier with 2-fold cross-validation. Figures 6.13 (a)- (c) shows the confusion matrix of the skin cancer data using (a) full features, (b) four features selected by regression method, (c) four features selected by the rMI-SVM algorithm using SVM - linear kernel function classifier with 2-fold

cross-validation. The value “0” indicated the normal person; while value “1” indicated the actinic keratosis and value “2” indicated the squamous cell carcinoma.

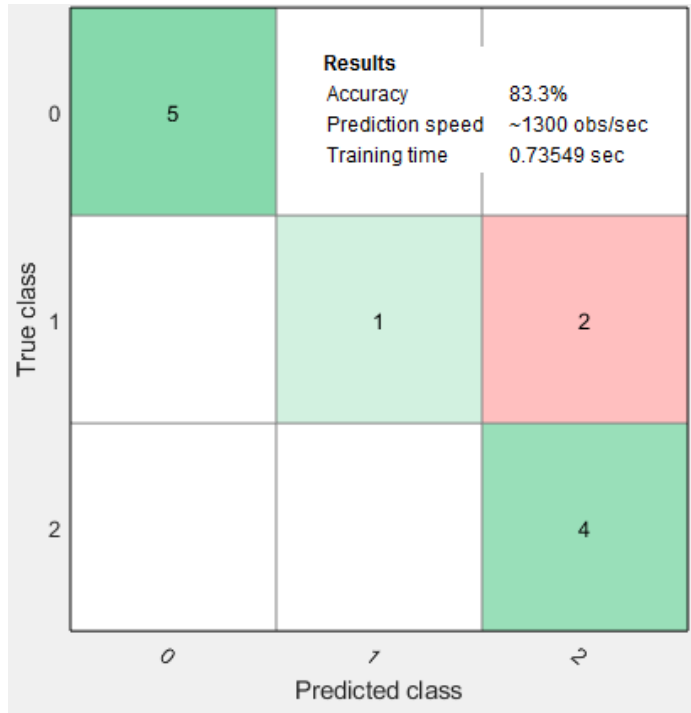
When full features used the predictive model, the accuracy was 58.3%, the prediction speed was about 0.04 observation per second, and the training time was 962.58 second. There were seven subjects which were classified correct while five other subjects were wrongly classified. The four features selected by the regression method, the accuracy was 41.7%, the prediction speed was about 860 observation per second, and the training time was 0.81714 second. Five subjects were classified correct while seven other subjects were wrongly classified. The four features selected by the rMI-SVM algorithm, the accuracy was 83.3%, the prediction speed was about 1300 observation per second, and the training time was 0.73549 second. There were ten subjects classified correct while two other subjects were wrongly classified.. The four features selected by rMI-SVM algorithm achieved the highest accuracy, fastest prediction speed and shortest training time compared to the baseline using all features and regression method. Therefore, the four features selected by rMI-SVM algorithm provided a better prediction power.



(a) Full features



(b) Four features selected by regression method



(c) Four features selected by the rMI-SVM algorithm

Figure 6.13: Confusion matrix for skin cancer data with 2-fold cross-validation

Table 6.30 shows the output of the ROC curve of the skin cancer data using full features. Four of the features were selected by the regression method, rMI-SVM algorithm using SVM - linear kernel function classifier with 2-fold cross-validation. When full features use the predictive model, the area under curve was 0.97, and the four features selected by the regression method, the area under curve was 0.71. The four features chosen by rMI-SVM algorithm, the area under curve was one. The four features selected by rMI-SVM algorithm gave the highest area under curve value indicated that the prediction model built by these four selected features was better compared to the baseline using full features and regression method. Therefore, the classifier using the four features selected by the rMI-SVM algorithm was rightly predicted. In general, the predictive model using the four features of the skin

cancer data set selected by rMI-SVM algorithm gave the highest accuracy in SVM classifier compared to the regression method with the same number of features. The area under curve shows that the four features selected by rMI-SVM algorithm gave a better predictive model compared to the same number of features chosen by regression method even when compared to the predictive model that was built using full features.

Table 6.30: Output of the ROC curve of skin cancer data using four features

	AUC
Full features	0.97
Regression	0.71
rM-SVM	1

Table 6.31 shows the 2-fold cross-validation and average accuracy for the SVM classifier using all features, rMI-SVM algorithm and regression method. The 22 features selected by the rMI-SVM algorithm with linear kernel function gave the highest average accuracy with the SVM classifier-linear kernel function compared to the regression method that used all the features. By using full features, the 2-fold cross-validation and average accuracy achieved by using SVM - linear kernel function classifier was 81.39% and 98.33%. The 22 features selected by the rMI-SVM algorithm with linear kernel function gave the 2-fold cross-validation and the average accuracy achieved by using SVM classifier - linear kernel function was 87.22% and 99.17%. The 22 features selected by the regression method gave the 2-fold

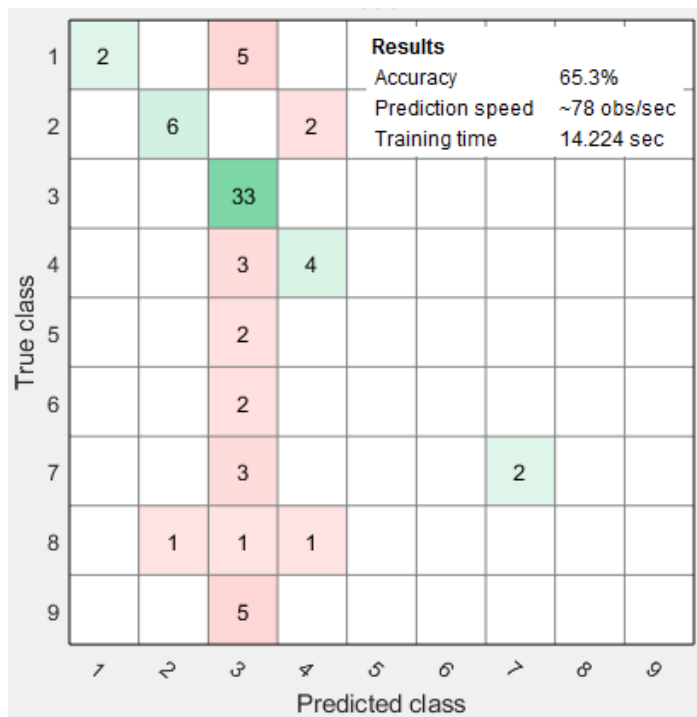
cross-validation, and average accuracy achieved by using SVM - linear kernel function classifier was 60.56% and 55%.

Table 6.31: Two-fold cross-validation and average accuracy for SVM classifier using full features, rMI-SVM algorithm, and regression method for lymphoma data set

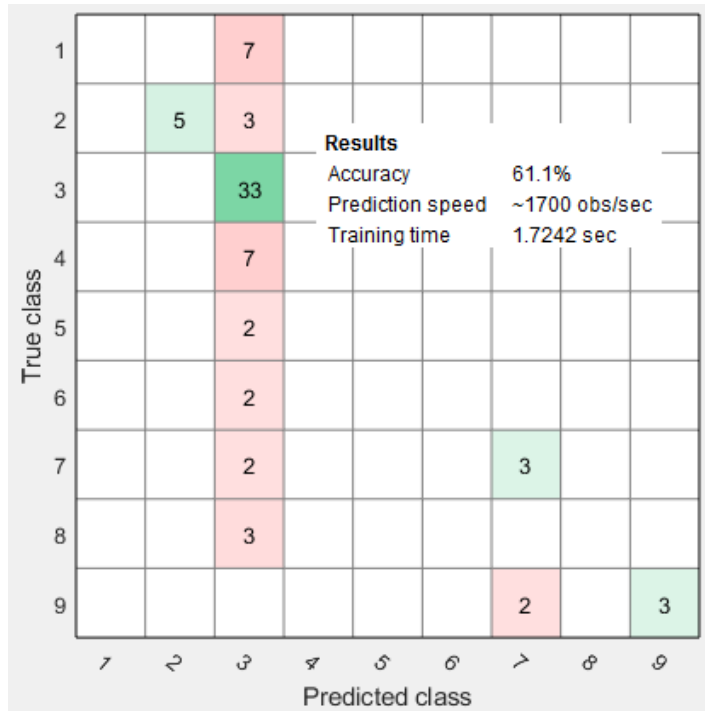
	SVM-CV	SVM-Acc
Baseline	81.3889	98.3333
rMI-SVM	87.2222	99.1667
Regression	60.5556	55

The confusion matrix and the receiver operating characteristic curve (ROC curve) were built to illustrate the distribution of the true class versus the predicted class for the full variables. The 22 features selected by regression method and rMI-SVM algorithm with linear kernel function used SVM classifier with 2-fold cross-validation. Figures 6.14 (a)- (c) shows the confusion matrix of the lymphoma data set using (a) full features, (b) 22 features selected by regression method, (c) 22 features selected by the rMI-SVM algorithm using SVM - linear kernel function classifier with 2-fold cross-validation. When full features used the predictive model, the accuracy was 65.3%, the prediction speed was about 78 observation per second, and the training time was 14.224 second. There were 47 subjects which were classified correct while 25 other subjects were wrongly classified. The 22 features selected by the regression method, the accuracy was 61.1%, the prediction speed was about 1700 observation per second, and the training time was 1.7242 second. There were 44 subjects which were classified correct

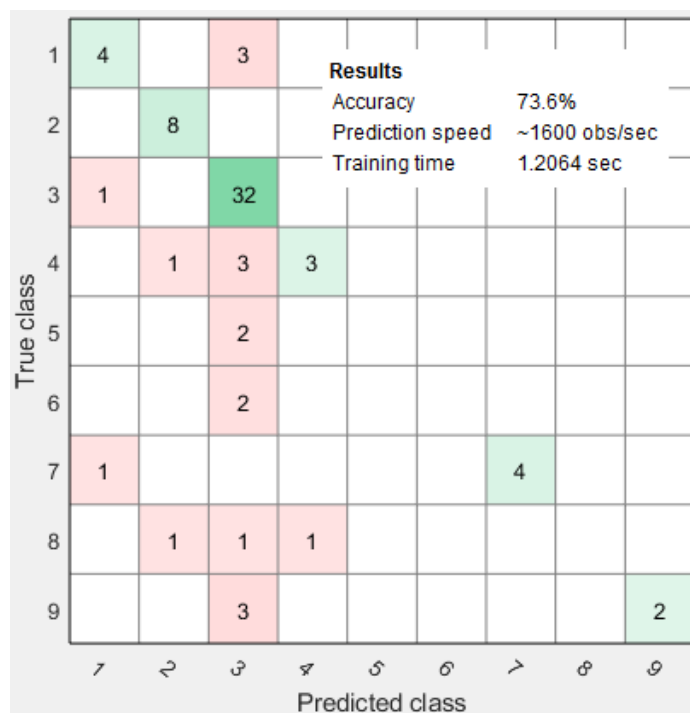
while 28 other subjects were wrongly classified. The 22 features selected by the rMI-SVM algorithm, the accuracy was 73.6%, the prediction speed was about 1600 observation per second, and the training time was 1.2064 second. There were 53 subjects classified correct while 19 other subjects were wrongly classified. The 22 features selected by rMI-SVM algorithm achieved the highest accuracy and shortest training time compared to the baseline using full features and regression method. Therefore, the 22 features selected by rMI-SVM algorithm provided a better prediction power.



(a) Full features



(b) 22 features selected by the regression method



(c) 22 features selected by the rMI-SVM algorithm

Figure 6.14: Confusion matrix for lymphoma data set with 2-fold cross-validation

Table 6.32 shows the output of the ROC curve of the lymphoma data using full features, 22 features were selected by the regression method, rMI-SVM algorithm, using SVM - linear kernel function classifier with 2-fold cross-validation. When full features used the predictive model, the area under curve was 0.93 and the 22 features selected by the regression method, the area under curve was 0.74. The 22 features selected by the rMI-SVM algorithm, the area under curve was 0.97. The 22 features selected by rMI-SVM algorithm gave the highest area under curve value indicated that the prediction model built by these 22 selected features was better compared to the baseline using full features and regression method. Therefore, the classifier using the 22 features selected by the rMI-SVM algorithm was rightly predicted. In general, the predictive model that used the 22 features of the lymphoma data set selected by rMI-SVM algorithm gave the highest accuracy to support the vector machine classifier compared to the regression method with the same number of features. The area under curve shows that the 22 features selected by rMI-SVM algorithm gave a better predictive model compared to the same number of features chosen by regression method which was compared to the predictive model that was built using full features.

Table 6.32: Output of the ROC curve of lymphoma data using 22 features

	AUC
All features	0.93
Regression	0.74
rM-SVM	0.97

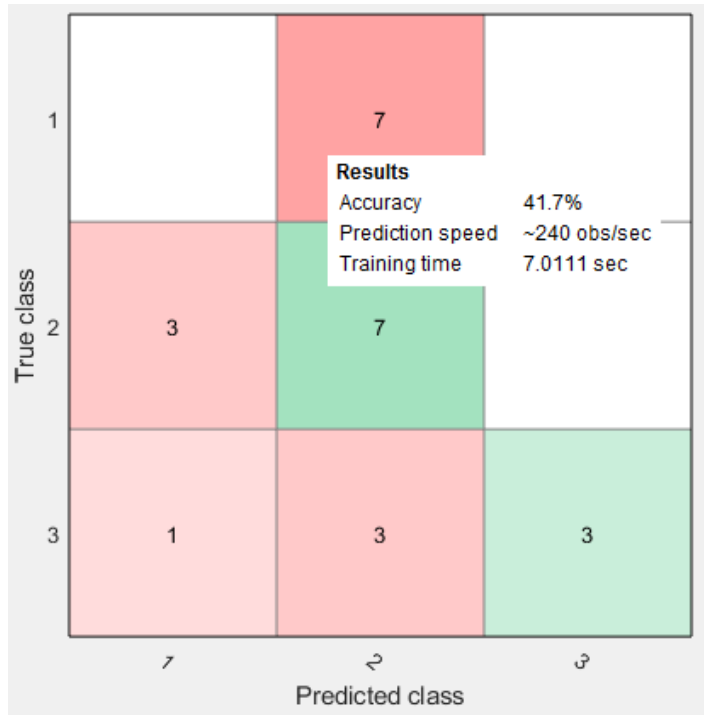
Table 6.33 shows the 3-fold cross-validation and average accuracy for the SVM classifier using all the features, rMI-SVM algorithm and regression method. The five features selected by the rMI-SVM algorithm with linear kernel function gave the highest average accuracy with the SVM classifier-linear kernel function compared to the regression method and using all the features. By using full features, the 3-fold cross-validation and average accuracy achieved by using SVM - linear kernel function classifier was 41.17% and 67.5%. The five features selected by the rMI-SVM algorithm with linear kernel function gave the 3-fold cross-validation and average accuracy achieved by using SVM classifier - linear kernel function was 64.17% and 77.5%. The five features selected by the regression method gave the 3-fold cross-validation, and the average accuracy achieved by using SVM - linear kernel function classifier was 56.67% and 67.5%.

Table 6.33: Three-fold cross-validation and average accuracy for SVM classifier using full features, rMI-SVM algorithm, and regression method for lung cancer data set

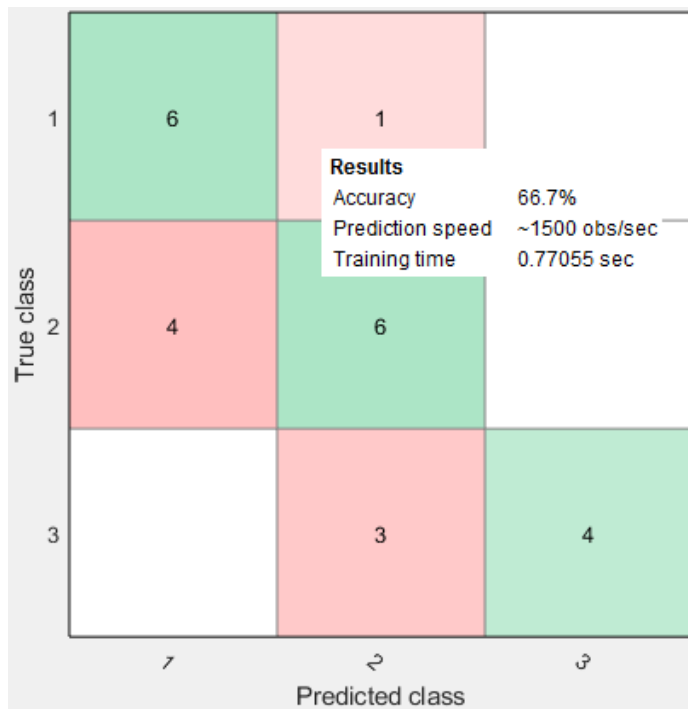
	SVM-CV	SVM-Acc
Baseline	44.1667	67.5
rMI-SVM	64.1667	77.5
Regression	56.6667	67.5

The confusion matrix and the receiver operating characteristic curve (ROC curve) were built to illustrate the distribution of the true class versus the predicted class for the full features. The five features selected by regression method and rMI-SVM algorithm with linear kernel function used the SVM

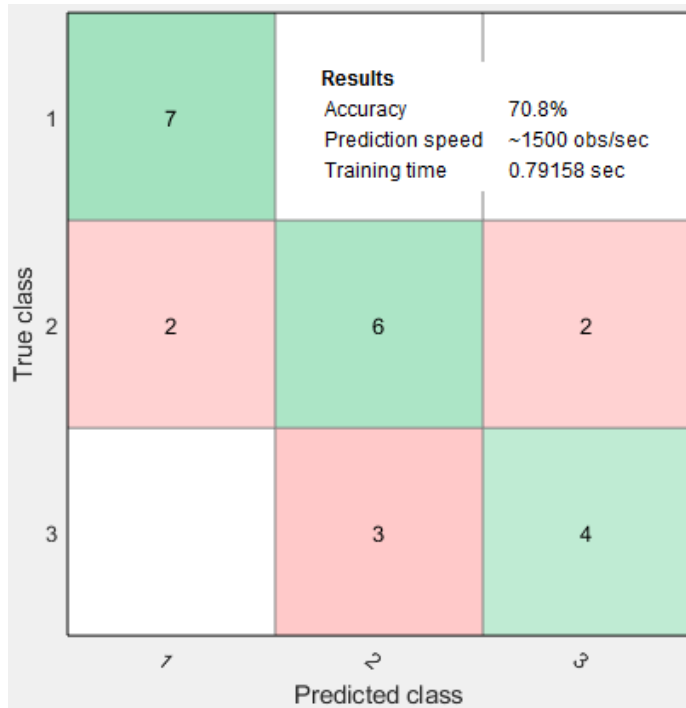
classifier with 3-fold cross-validation. Figures 6.15 (a)-(c) shows the confusion matrix of the lung cancer data set using (a) full features, (b) five features selected by regression method, (c) five features chosen by the rMI-SVM algorithm using SVM - linear kernel function classifier with 3-fold cross-validation. When full features are using the predictive model, the accuracy was 41.7%, the prediction speed was about 240 observation per second, and the training time was 7.0111 second. There were ten subjects which were classified correct while 14 other subjects were wrongly classified. The five features selected by the regression method, the accuracy was 66.7%, the prediction speed was about 1500 observation per second, and the training time was 0.77055 second. There were 16 subjects which were classified correctly while eight other subjects were wrongly classified. The five features selected by the rMI-SVM algorithm, the accuracy was 70.8%, the prediction speed was about 1500 observation per second, and the training time was 0.79158 second. There were 17 subjects which were classified correct while seven other subjects were classified wrongly. The five features selected by rMI-SVM algorithm achieved the highest accuracy compared to the baseline using all features and regression method. Therefore, the five features selected by rMI-SVM algorithm provided a better prediction power.



(a) Full features



(b) Five features selected by regression method



(c) Five features selected by the rMI-SVM algorithm

Figure 6.15: Confusion matrix for lung cancer data set with 3-fold cross-validation

Table 6.34 shows the output of the ROC curve of the lung cancer data using full features, five features were selected by the regression method, rMI-SVM algorithm using SVM - linear kernel function classifier with 3-fold cross-validation. When full features used the predictive model, the area under curve was 0.53 and the five features selected by the regression method, the area under curve was 0.89. The five features chosen by rMI-SVM algorithm, the area under curve was 0.97. The five features selected by rMI-SVM algorithm gave the highest area under curve value indicated that the prediction model built by these five selected features was better compared to the baseline using all features and regression method. Therefore, the classifier using the five features chosen by the rMI-SVM algorithm was rightly predicted. In general, the predictive model that used the five features of the lung cancer data

set selected by rMI-SVM algorithm gave the highest accuracy in SVM classifier compared to the regression method with the same number of features. The area under curve shows that the five features selected by rMI-SVM algorithm give a better predictive model compared to the same number of features chosen by regression method even when compared with the predictive model that was built using full features.

Table 6.34: Output of the ROC curve of lung cancer data using four features

	AUC
All features	0.53
Regression	0.89
rM-SVM	0.97

Table 6.35 shows the 10-fold cross-validation and average accuracy for the SVM classifier using full features, rMI-SVM algorithm and regression method. The 50 features selected by the rMI-SVM algorithm with linear kernel function gave the highest average accuracy with the SVM classifier-linear kernel function compared to the regression method using all the features. By using full features, the 10 –fold cross-validation and average accuracy achieved by using SVM - linear kernel function classifier was 98.23% and 98.2%. The 50 features selected by the rMI-SVM algorithm with linear kernel function gave the 10-fold cross-validation and average accuracy achieved by using SVM classifier - linear kernel function was 98.37% and 98.8%. The 50 features selected by the regression method gave the 10-fold cross-validation,

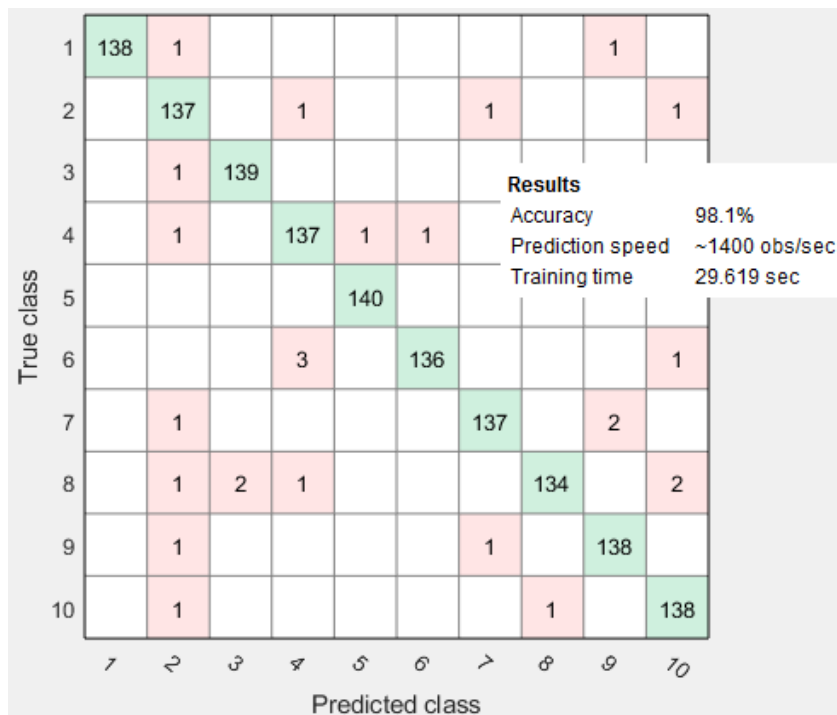
and average accuracy achieved by using SVM - linear kernel function classifier was 96.89% and 96.5%.

Table 6.35 Ten-fold cross-validation and average accuracy for support vector machine classifier using full features, rMI-SVM algorithm, and regression method for handwriting data set

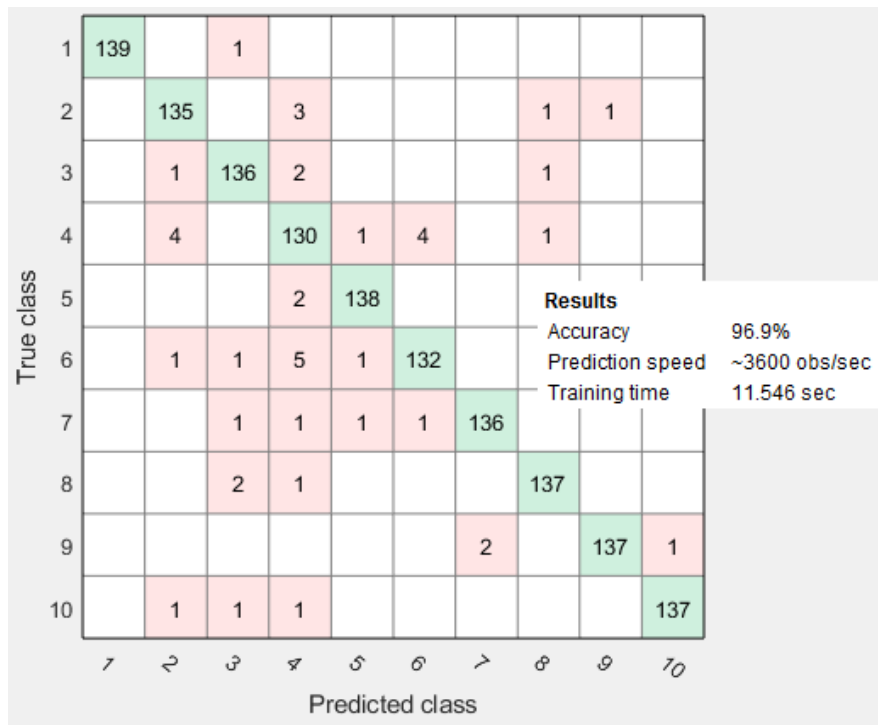
	SVM-CV	SVM-Acc
Baseline	98.2286	98.2
rMI-SVM	98.3714	98.8
Regression	95.8857	96.5

The confusion matrix and the receiver operating characteristic curve (ROC curve) were built to illustrate the distribution of the true class versus the predicted class for the 649 variables. The 50 features selected by regression method and rMI-SVM algorithm with linear kernel function used SVM classifier with 10-fold cross-validation. Figure 6.16 (a)-(c) shows the confusion matrix of the handwriting data set using (a) full features, (b) 50 features selected by regression method, (c) 50 features selected by the rMI-SVM algorithm using SVM - linear kernel function classifier with 10-fold cross-validation. When full features used the predictive model, the accuracy was 98.1%, the prediction speed was about 1400 observation per second, and the training time was 29.619 second. There were 1374 subjects which were classified correct while 26 other subjects were wrongly classified. The 50 features selected by the regression method, the accuracy was 96.9%, the prediction speed was about 3600 observation per second, and the training time was 11.546 second. There were 1357 subjects which were classified correct while 43 other subjects were wrongly classified. The 50 features selected by

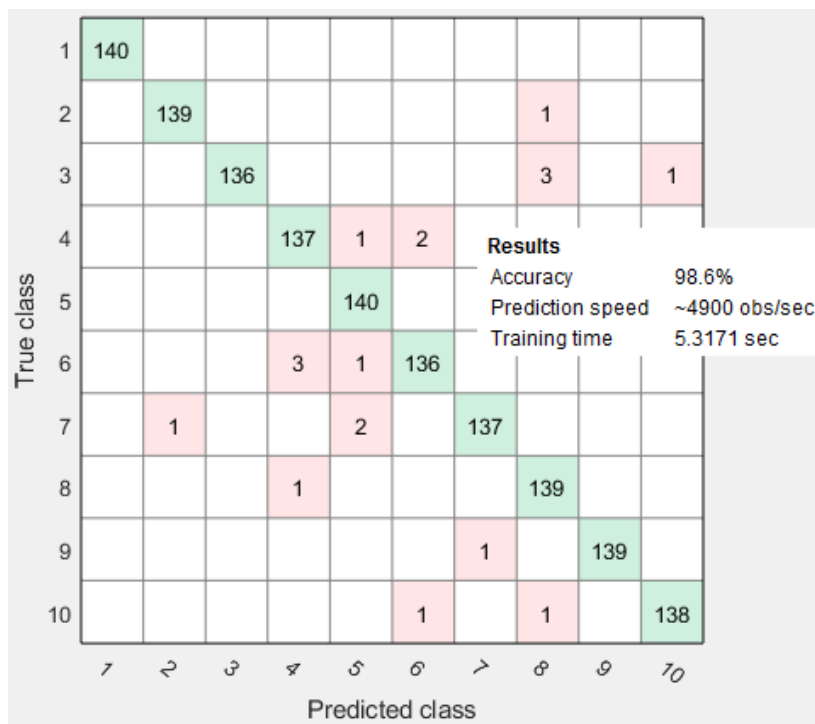
the rMI-SVM algorithm, the accuracy was 98.6%, the prediction speed was about 4900 observation per second, and the training time was 5.3171 second. There were 1381 subjects which were classified correct while only 19 subjects were wrongly classified. The 50 features selected by rMI-SVM algorithm achieved the highest accuracy, fastest prediction speed and shortest training time compared to the baseline using all features and regression method. Therefore, the 50 features selected by rMI-SVM algorithm provided a better prediction power.



(a) Full features



(b) 50 features selected by the regression method



(c) 50 features selected by the rMI-SVM algorithm

Figure 6.16: Confusion matrix for handwriting data set with 10-fold cross-validation

Table 6.36 shows the output of the ROC curve of the handwriting data using full features. The 50 features selected by the regression method, rMI-SVM algorithm used the SVM - linear kernel function classifier with 10-fold cross-validation. When full features used the predictive model, the area under curve was one, and the 50 features selected by the regression method, the area under curve was one. The 50 features selected by the rMI-SVM algorithm, the area under curve was 1. The 50 features selected by rMI-SVM algorithm gave a better and true positive in the current classifier which indicated that the prediction model built by these 50 selected features was better compared to the baseline using full features and regression method. Therefore, the classifier using the 50 features selected by the rMI-SVM algorithm was rightly predicted. In general, the predictive model that use the 50 features of the handwriting data set selected by rMI-SVM algorithm gave the highest accuracy in SVM classifier compared to the regression method with the same number of features. The area under curve shows that the 50 features selected by rMI-SVM algorithm gave a better predictive model compared to the same number of features chosen by regression method and even when compared to the predictive model that was built using full features.

Table 6.36: Output of the ROC curve of handwriting data using 50 features

	AUC
All features	1
Regression	1
rM-SVM	1

This chapter shows the result of the optimal baseline and dimension reduction using the rMI-SVM algorithm for the ten data sets. Some of the data sets are microarray data with high dimensional but low sample size; these data sets have been chosen to evaluate the performance of the proposed algorithm. The optimal baselines were obtained using the mutual information score to show that this optimal baseline is better than the existing baseline used by the researcher. The current baseline is using the full features which will contain the irrelevant features and noise in the predictive model. The built predictive model using full features has less predictive power; therefore, the baseline using all the features is less reliable. The proposed algorithm in Chapter 3 by excluding the irrelevant features in the predictive model will provide a better baseline and also obtain the number of features needed for this baseline. The number of features obtained from the proposed algorithm serves as a guideline on features selection as the number of cut off features that is needed in building a predictive model.

Besides this, the proposed algorithm in Chapter 4 has been evaluated using the ten chosen data sets. It shows that the dimension of the data has been reduced tremendously, up to 90% of the original dimension. After reduction of dimension, the performance of the predictive model built by the selected features using rMI-SVM algorithm provides a better than the existing baseline, regression method and mRMR. The sensitivity test through the confusion matrix and ROC curve also shows that the features selected by the proposed algorithm perform better than using full features and the features chosen by the regression method and mRMR.

CHAPTER 7

DISCUSSION

7.1 Evaluation of the features selected by the rMI-SVM algorithm

The features selected using various feature selection methods face the same problem, as to whether the chosen features are coincidence or whether the chosen features are represented in the predictive model. In this chapter, the features selected by the proposed algorithm will be evaluated using the Z-score. The Z-score has been discussed in Chapter 5.3, since the sample size in the microarray is relatively small. Therefore, Z-score that helps to evaluate the features selected by the proposed algorithm is representative and not chosen by chance. The representative features will give high value in the Z-score; but, the Z-score only highlights those relevant features and at the same time also highlights the redundant features. The redundant features are represented in the predictive model. When more redundant features are included in the predictive model, it does not guarantee that the performance will become better but it will increase the complexity of the predictive model. Therefore, the Z-score will show all the relevant features that are not selected by chance. Later, the features chosen by the proposed algorithm will compare the performance of the predictive model with some existing feature selection methods as discussed in Chapter 2.

7.2 Z-score Analysis

7.2.1 Z-score Analysis for Binary Data Set

Figure 7.1 shows the Z-score versus the features for colon cancer data set with $K=50$. When K value increases, more features with higher Z-score will show in the figure. From Figure 7.1, there were 185 features with the value of Z-score which indicated that a predictive model that contains these features with a high value of Z-score would be more robust.

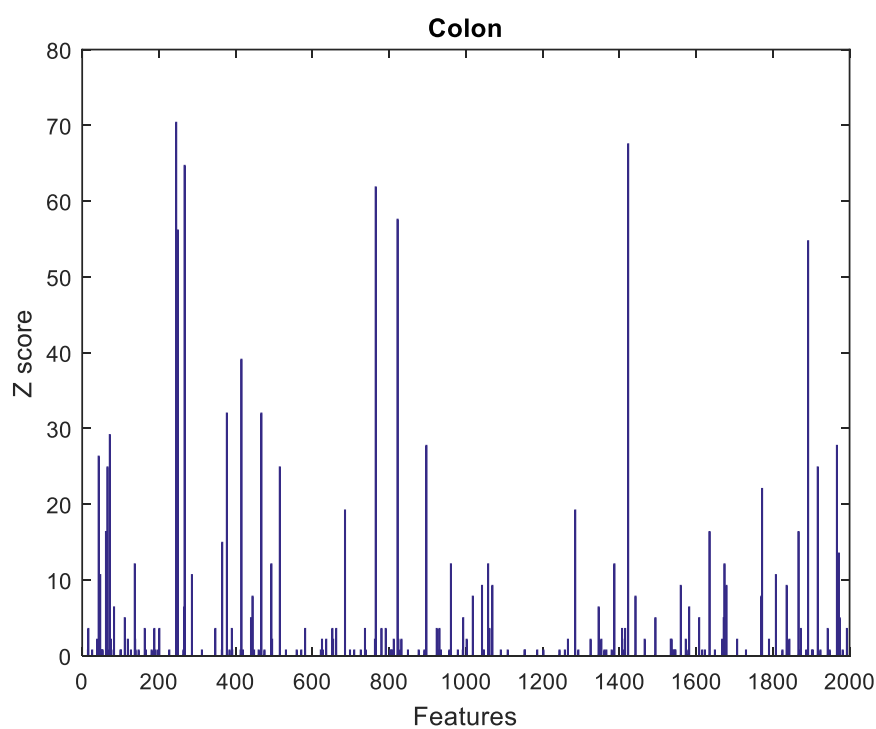


Figure 7.1: Z-score versus the features for colon cancer data set

Table 7.1 shows the top 20 ranked features with the Z-score which were selected by regression method, mRMR method and rMI-SVM algorithm. The regression method chose the top six ranked features, whereas the mRMR

method chose only three features from the top 20 ranked features and the rMI-SVM algorithm only selected six features from the top 20 ranked features. Although the rMI-SVM algorithm only selected six features from the top 20 ranked features, nevertheless, the predictive model built by the rMI-SVM algorithm using seven features gave the highest average accuracy compared to the regression method and mRMR method.

Although the *Z*-score was calculated based on the specific feature but it did not take a count on the dynamic change on the already selected features with the newly added feature. The *Z*-score is able to figure out the significance of the occurrence of a selected feature. This means that those features with high *Z*-score were not chosen by chance and contain more information needed in the predictive model. However, the features that provide high information and not selected by chance might be redundant to other features. Therefore, those redundant features will be filtered out by the rMI-SVM algorithm. From Table 7.1, the regression method did not take into account the dynamic change on the already selected features with the newly added feature in the predictive model.

Table 7.1: The top 20 ranked features with Z-score selected by regression method, mRMR method and rMI-SVM algorithm for colon cancer data set

Z score	Feature	Regression	mRMR	rMI-SVM
70.35624	245	v		v
67.51356	1423	v		v
64.67088	267	v		
61.82821	765	v	v	
57.56419	822	v		
56.14286	249	v	v	v
54.72152	1892			
39.0868	415			
39	20			
31.98011	377			v
31.98011	467			v
29.13743	72			v
27.71609	897			
27.71609	1967			
26.29476	43			
24.87342	66			
24.87342	515			
24.87342	1917			
24	34			
22.03074	1772		v	

Figure 7.2 shows the Z-score versus the features for leukaemia data set with $K=50$. When the K value increases, more features with higher Z-score will show in the figure. From Figure 7.2, there were 91 features with the value of Z-score which indicated that a predictive model that contains more features with a high value of Z-score would be more robust.

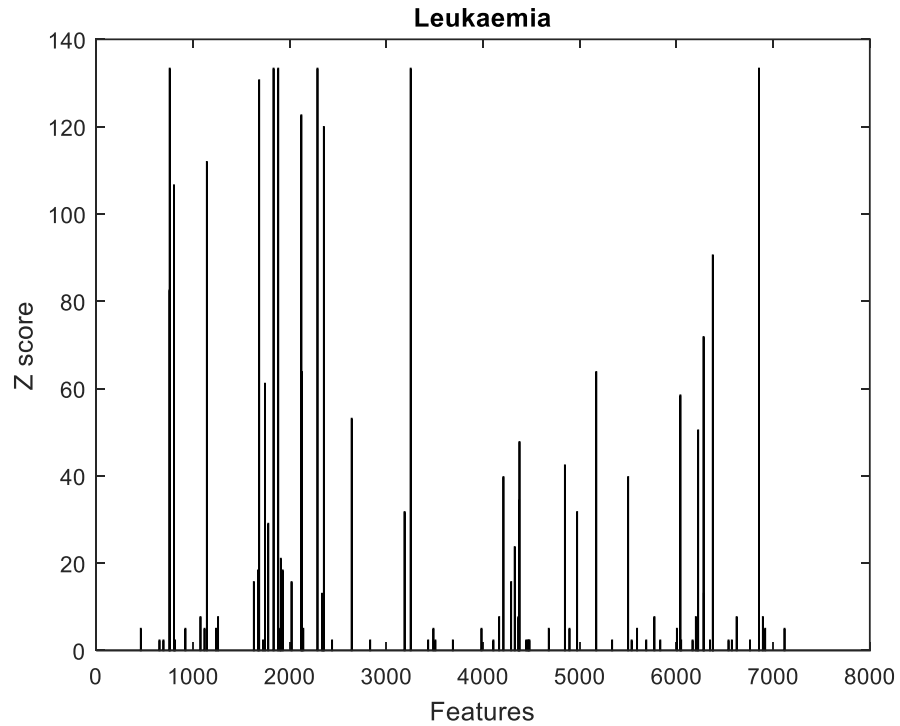


Figure 7.2: Z-score versus the features for the leukaemia data set

Table 7.2 shows the top 30 ranked features with the Z-score selected by regression method, mRMR method and rMI-SVM algorithm. The regression method selected five ranked features from the top 30 ranked features, whereas the mRMR method chose only three features from the top 30 ranked features and the rMI-SVM algorithm selected five features from the top 30 ranked features. Although the rMI-SVM algorithm did not choose the top five features, nevertheless, the predictive model built by the rMI-SVM algorithm using five features gave the highest average accuracy, compared to the regression method and mRMR method. From Table 7.2, the regression method did not take into account the dynamic change on the already selected features with the newly added feature in the predictive model.

Table 7.2: The top 30 ranked features with Z-score selected by regression method, mRMR method and rMI-SVM algorithm for leukaemia data set

Z score	Feature	Regression	mRMR	rMI-SVM
133.3042	760			
133.3042	1834	v	v	
133.3042	1882	v		
133.3042	2288	v		
133.3042	3252	v		v
133.3042	6854	v		
130.6306	1685		v	
122.6098	2121			
119.9362	2354		v	
111.9155	1144			
106.5683	804			
90.5268	6376			
82.50605	758			
71.81171	6281			
63.79096	2128			v
63.79096	5171			
61.11737	1745			
58.44379	6041			
53.09662	2642			
50.42303	6225			
47.74945	4377			v
42.40228	4847			
39.72869	4211			
39.72869	5501			
34.38152	4373			
31.70794	3189			v
31.70794	4973			
29.03435	1779			v
23.68718	4328			
21.0136	1909			

Figure 7.3 shows the Z-score versus the features for prostate data set with $K=50$. When the K value increases, more features with higher Z-score will be shown in the figure. From Figure 7.3, there were 101 features with value of Z-score which indicated that a predictive model that contains more features with high value of Z-score will be more robust.

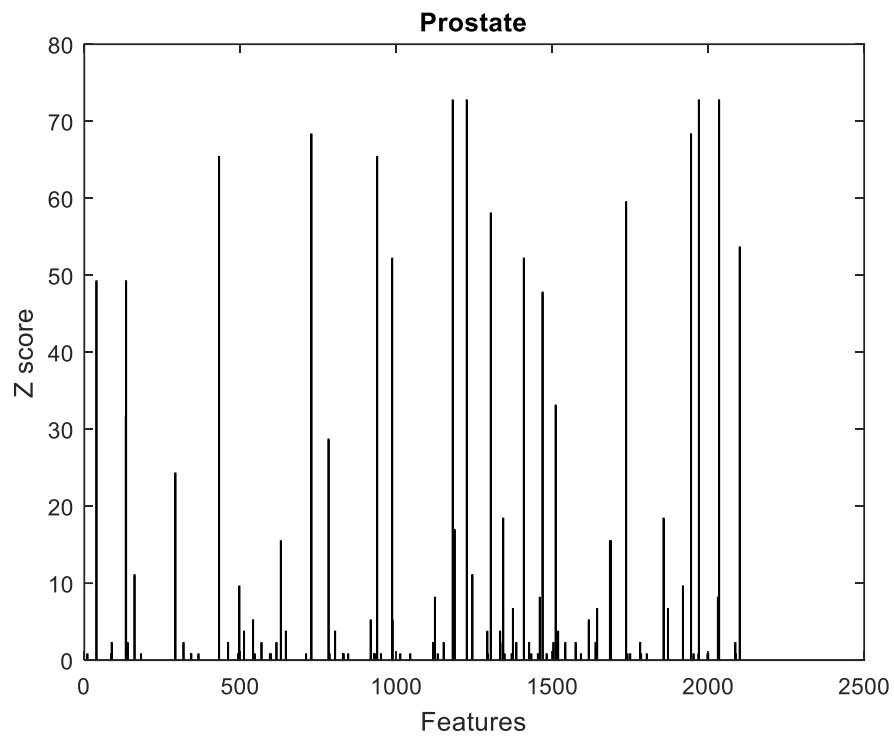


Figure 7.3: Z-score versus the features for the prostate data set

Table 7.3 shows the top 10 ranked features with the Z-score selected by regression method, mRMR method and rMI-SVM algorithm. The regression method selected three ranked features from the top 10 ranked features, whereas the mRMR method chose only two features from the top 10 ranked features and the rMI-SVM algorithm selected three features from the top 10 ranked features. The regression method and rMI-SVM algorithm had

chosen the same three features; therefore, the average accuracy was the same for the regression method, and rMI-SVM algorithm compared to the mRMR method.

Table 7.3: The top 10 ranked features with Z-score selected by regression method, mRMR method and rMI-SVM algorithm for prostate data set

Z score	Feature	Regression	mRMR	rMI-SVM
72.7152	1181			
72.7152	1226	v		v
72.7152	1970	v	v	v
72.7152	2035			
68.31103	728			
68.31103	1945			
65.37492	432	v	v	v
65.37492	939			
59.50269	1737			
58.03464	1303			

Figure 7.4 shows the Z-score versus the features for colon cancer data set with $K=50$. When the K value increases, more features with higher Z-score will be shown in the figure. From Figure 7.4, there were 61 features with the value of Z-score which indicated that a predictive model that contains more features with a high value of Z-score would be more robust.

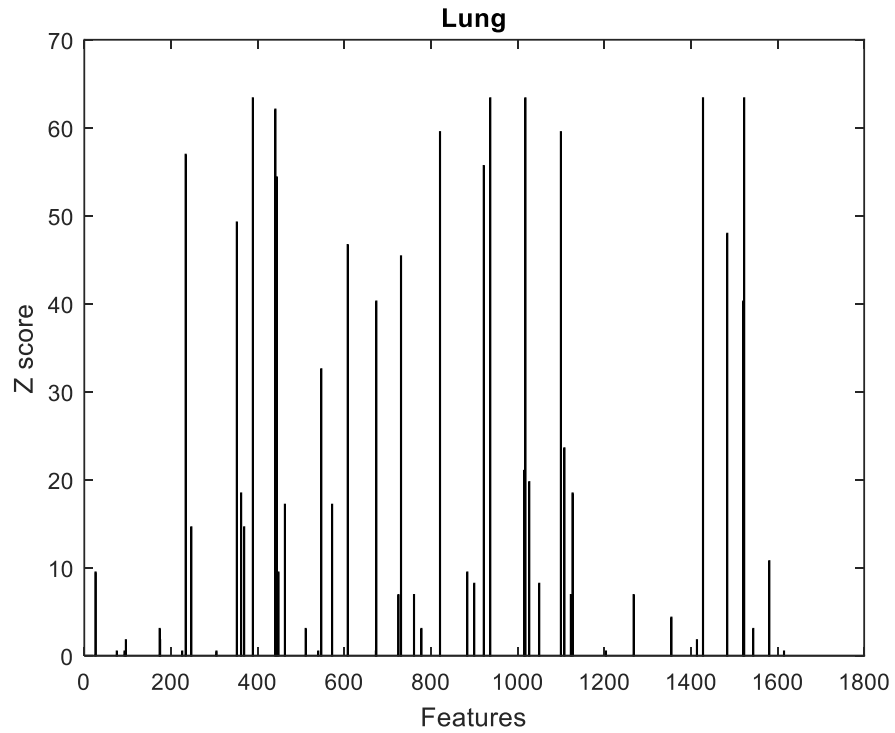


Figure 7.4: Z-score versus the features for lung cancer data set

Table 7.4 shows the top 20 ranked features with the Z-score selected by regression method, mRMR method and rMI-SVM algorithm. The regression method selected four ranked features from the top 20 ranked features, whereas the mRMR method chose only three features from the top 20 ranked features and the rMI-SVM algorithm selected only four features from the top 20 ranked features. The four chosen features by rMI-SVM algorithm gave the highest average accuracy compared to the regression method and mRMR method.

Table 7.4: The top 20 ranked features with Z-score selected by regression method, mRMR method and rMI-SVM algorithm for lung cancer data set

Z score	Feature	Regression	mRMR	rMI-SVM
123	67			
119	105			
110	17			
103	9			
103	102			
103	100			
101	94			
96	94			
64	7			
63.36403	389			
63.36403	937			
63.36403	1018	v		v
63.36403	1428	v	v	v
63.36403	1523			
62.08097	441	v	v	
59.51484	821	v	v	v
59.51484	1100			
56.94872	234			v
55.66565	922			
54.38259	444			

Figure 7.5 shows the Z-score versus the features for Parkinson data set with $K=50$. When the K value increases, more features with higher Z-score will be shown in the figure. From Figure 7.5, there were 13 features with the value of Z-score which indicated that a predictive model that contains more features with a high value of Z-score would be more robust.

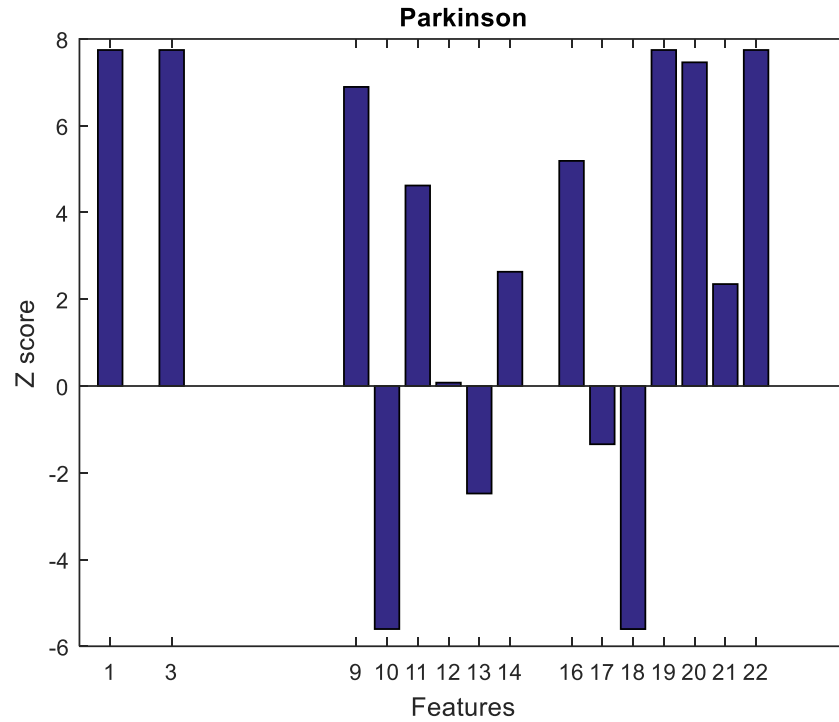


Figure 7.5: Z-score versus the features for Parkinson data set

Table 7.5 shows the top 10 ranked features with the Z-score selected by regression method, mRMR method and rMI-SVM algorithm. The regression method selected six ranked features from the top 10 ranked features, whereas the mRMR method chose only three features from the top 10 ranked features and the rMI-SVM algorithm only selected four features from the top 10 ranked features. Although the rMI-SVM algorithm only selected four features from the top 10 ranked features, nevertheless, the predictive model built by the rMI-SVM algorithm using seven features gave the highest average accuracy compared to the regression method and mRMR method.

Table 7.5: The top 10 ranked features with Z-score selected by regression method, mRMR method and rMI-SVM algorithm for Parkinson data set

Z score	Feature	Regression	mRMR	rMI-SVM
7.745967	1			v
7.745967	3	v	v	v
7.745967	19	v	v	v
7.745967	22	v		v
7.461948	20	v	v	
6.89391	9	v		
5.189798	16			
4.62176	11			
2.633629	14			
2.34961	21	v		

Figure 7.6 shows the Z-score versus the features for breast cancer data set with $K=50$. When the K value increases, more features with higher Z-score will be shown in the figure. From Figure 7.6, there were 12 features with the value of Z-score which indicated that a predictive model that contains more features with a high value of Z-score would be more robust.

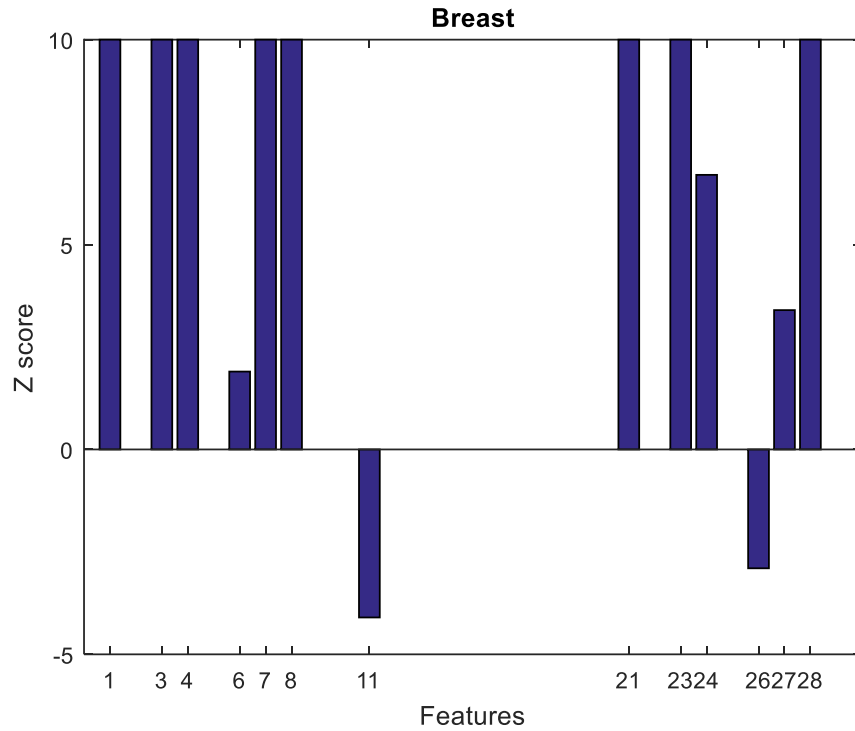


Figure 7.6: Z-score versus the features for breast cancer data set

Table 7.6 shows the top 10 ranked features with the Z-score selected by regression method, mRMR method and rMI-SVM algorithm. The regression method selected all the top 10 ranked features, whereas the mRMR method only picked eight features from the top 10 ranked features, and the rMI-SVM algorithm only selected seven features from the top 10 ranked features. Although the rMI-SVM algorithm only selected seven features from the top 10 ranked features, nevertheless, the predictive model built by the rMI-SVM algorithm using eleven features gave the highest average accuracy compared to the regression method and mRMR method.

The Z-score is able to figure out the significance of the occurrence of a selected feature. This means that those features with high Z-score were not

chosen by chance and contain more information needed in the predictive model. However, the features that provide high information and not selected by chance might be redundant to other features; therefore, those redundant features will be filtered out by the rMI-SVM algorithm. From Table 7.6, the regression method did not take into account the dynamic change on the already selected features with the newly added feature in the predictive model.

Table 7.6: The top 10 ranked features with Z-score selected by regression method, mRMR method and rMI-SVM algorithm for breast cancer data set

Z score	Feature	Regression	mRMR	rMI-SVM
10	1	v		
10	4	v	v	
10	7	v	v	v
10	8	v	v	v
10	21	v	v	v
10	23	v	v	v
10	28	v	v	v
6.7	24	v	v	v
3.4	27	v	v	v
1.9	6	v		

7.2.2 Z-score Analysis for Multiclass Data Set

Figure 7.7 shows the Z-score versus the features for skin cancer data set with $K=50$. When the K value increases, more features with higher Z-score will be shown in the figure. From Figure 7.7, there were 313 features with the

value of Z -score which indicated that a predictive model that contains more features with a high value of Z -score would be more robust.

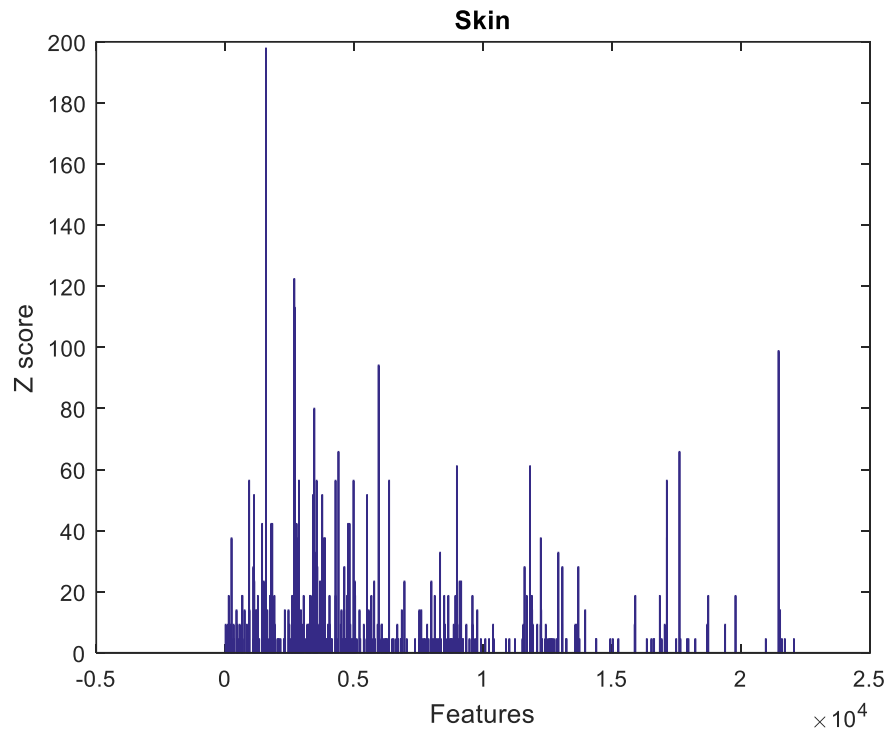


Figure 7.7: Z -score versus the features for skin cancer data set

Table 7.7 shows the top 20 ranked features with the Z -score selected by regression method and rMI-SVM algorithm. The regression method did not select any of the features in the top 20 ranked features, whereas the rMI-SVM algorithm selected two features from the top 20 ranked features. Although the rMI-SVM algorithm chose only two features from the top 20 ranked features, the predictive model built by the rMI-SVM algorithm using four features gave the highest average accuracy compared to the regression method because the Z -score was calculated based on the particular feature but did not take into account on the dynamic change on the already selected features with the newly

added feature. The features that contain high information and not chosen by chance might be redundant to other features; therefore, those redundant features will be filtered out using the rMI-SVM algorithm.

Table 7.7: The top 20 ranked features with Z-score selected by regression method and rMI-SVM algorithm for skin cancer data set

Z score	Feature	Regression	rMI-SVM
197.8346	1582		
122.3882	2680		
112.9574	2706		
98.81118	21462		v
94.09578	5952		
79.94958	3456		
65.80337	4394		
65.80337	17618		
61.08797	8986		
61.08797	11825		
56.37256	931		
56.37256	2862		
56.37256	3546		
56.37256	4280		
56.37256	4979		
56.37256	6359		
56.37256	17128		v
51.65716	1120		
51.65716	3408		
51.65716	3764		

Figure 7.8 shows the Z-score versus the features for lymphoma data set with $K=50$. When the K value increases, more features with higher Z-score

will be shown in the figure. From Figure 7.8, there were 191 features with the value of Z -score which indicated that a predictive model that contains more features with a high value of Z -score would be more robust.

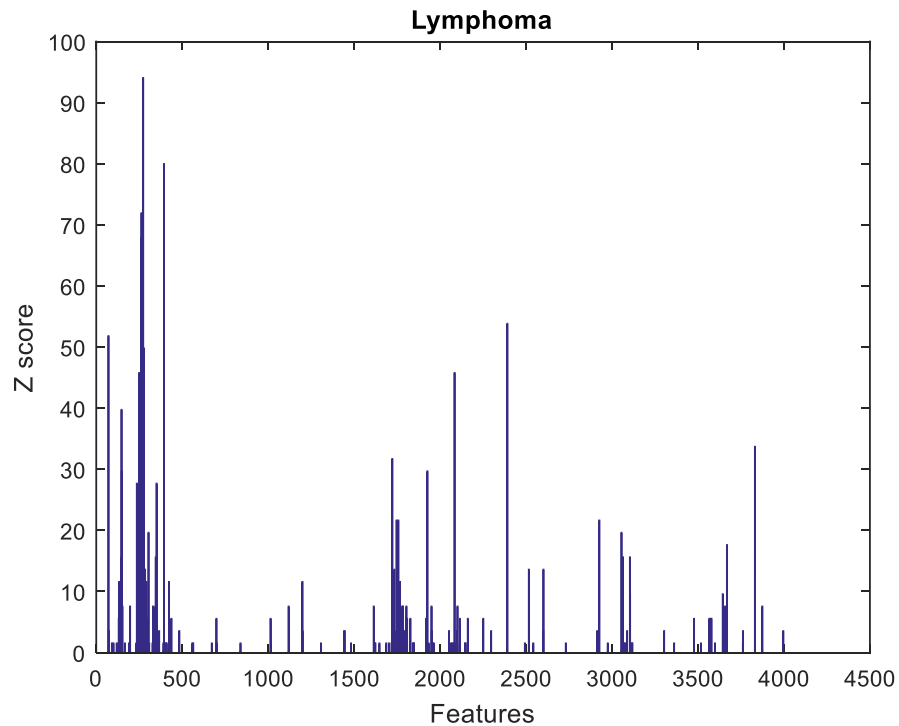


Figure 7.8: Z -score versus the features for the lymphoma data set

Table 7.8 shows the top 20 ranked features with the Z -score selected by regression method and rMI-SVM algorithm. The regression method picked six features from the top 20 ranked features, whereas the rMI-SVM algorithm selected seven features from the top 20 ranked features. The predictive model built by the rMI-SVM algorithm using 22 features gave the highest average accuracy compared to the regression method because the Z -score was calculated based on the specific feature but did not take a count on the dynamic change on the already selected features with the newly added feature.

Table 7.8: The top 20 ranked features with Z-score selected by regression method and rMI-SVM algorithm for lymphoma data set

Z score	Feature	Regression	rMI-SVM
94.0405	273	v	v
79.96005	394	v	
71.91408	263	v	
67.8911	262	v	
67.8911	266		v
65.87961	265		
53.81066	2391		
51.79917	71		v
49.78767	277		
45.76469	250		
45.76469	2084		v
39.73021	147	v	v
33.69574	271		
33.69574	3831		v
31.68425	276		
31.68425	1722		
29.67275	148	v	
29.67275	274		
29.67275	1926		
27.66126	236		v

Figure 7.9 shows the Z-score versus the features for lymphoma data set with $K=50$. When the K value increases, more features with higher Z-score will be shown in the figure. From Figure 7.9, there were 37 features with the value of Z-score which indicated that a predictive model that contains more features with a high value of Z-score would be more robust.

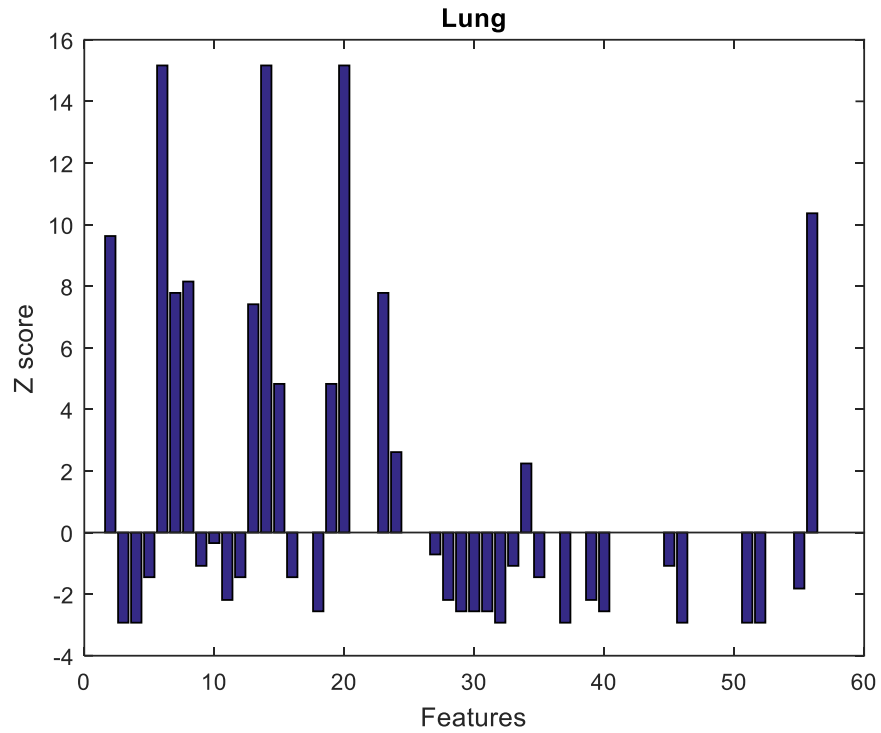


Figure 7.9: Z-score versus the features for lung cancer data set

Table 7.9 shows the top 15 ranked features with the Z-score selected by regression method and rMI-SVM algorithm. The regression method picked five features from the top 15 ranked features, whereas the rMI-SVM algorithm selected four features from the top 15 ranked features. The predictive model built by the rMI-SVM algorithm using five features gave the highest average accuracy compared to the regression method.

Table 7.9: The top 15 ranked features with Z-score selected by regression method and rMI-SVM algorithm for lung cancer data set

Z score	Feature	Regression	rMI-SVM
15.16575	6		v
15.16575	14	v	v
15.16575	20	v	v
10.36546	56		
9.626955	2		
8.149943	8	v	
7.78069	23	v	
7.411437	13		
4.826665	15		
4.826665	19	v	
2.611147	24		
2.241894	34		
-0.34288	10		v
-0.71213	27		
-1.08138	9		

Figure 7.10 shows the Z-score versus the features for handwriting data set with $K=50$. When the K value increases, more features with higher Z-score will be shown in the figure. From Figure 7.10, there were 38 features with the value of Z-score which indicated that a predictive model that contains more features with a high value of Z-score would be more robust.

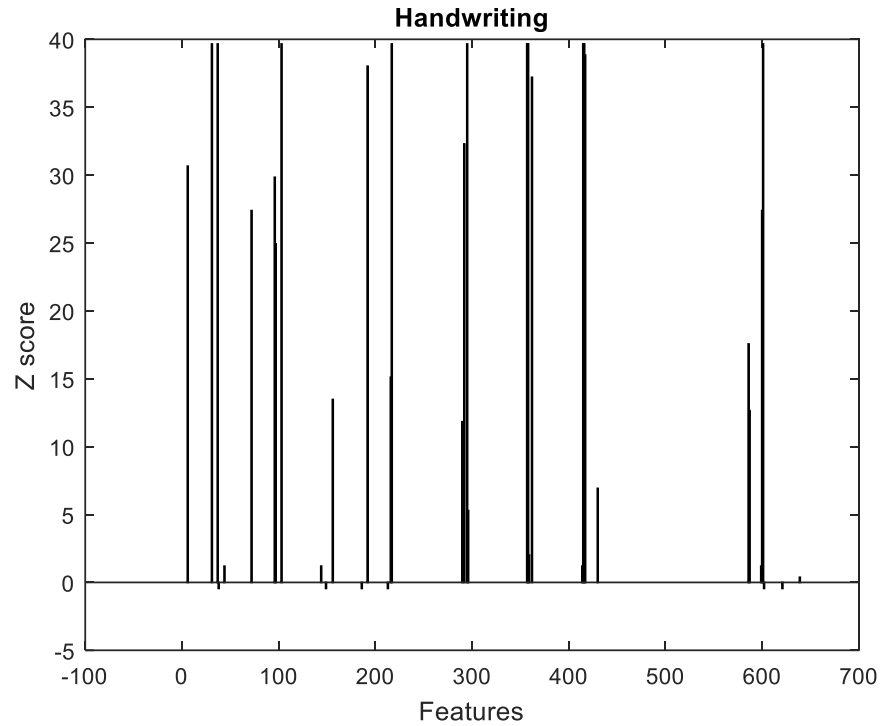


Figure 7.10: Z-score versus the features for the handwriting data set

Table 7.10 shows the top 20 ranked features with the Z-score selected by regression method and rMI-SVM algorithm. The regression method selected 12 features from the top 20 ranked features, whereas the rMI-SVM algorithm selected 17 features from the top 20 ranked features. The predictive model built by the rMI-SVM algorithm using 50 features gave the highest average accuracy compared to the regression method.

Table 7.10: The top 20 ranked features with Z-score selected by regression method and rMI-SVM algorithm for handwriting data set

Z score	Feature	Regression	rMI-SVM
39.65476	31		v
39.65476	37		v
39.65476	103	v	v
39.65476	217	v	v
39.65476	295		v
39.65476	357	v	v
39.65476	358		v
39.65476	415	v	v
39.65476	416	v	v
39.65476	601	v	v
38.83645	417	v	v
38.01813	192		v
37.19982	362		v
32.28994	292	v	v
30.65332	6		v
29.83501	96	v	
27.38007	72		v
27.38007	600	v	
24.92513	97	v	v
17.56031	586	v	

7.3 Cross-validation

7.3.1 Cross-validation for Binary Data Set

Overfitting is a widespread issue in machine learning or classification. When a statistical model or predictive model uses too many parameters, it is easy to cause overfitting. In machine learning, the machine becomes too entangled to construct a predictive model with a low error rate, and it covers

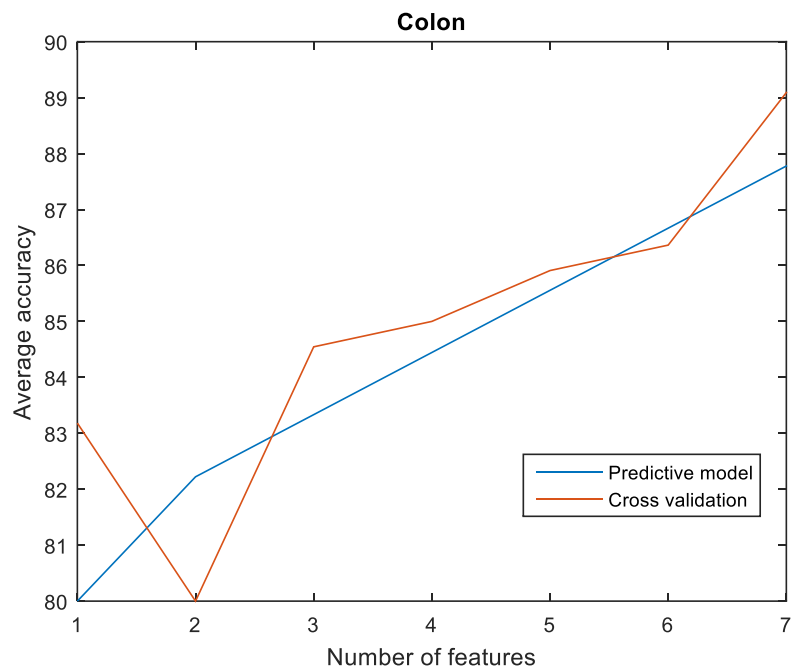
many data points so the error rate will be lower. In reality, when using this predictive model, it is not possible to predict the other data besides the training data. The easy way is to increase the sample size to overcome the overfitting. However, microarray data analysis usually involves a small sample size. Therefore, it is essential to determine the predictive model that was built as not overfitting.

A simple way to determine whether the predictive model was overfitting or not, an average accuracy of cross-validation and average accuracy of the predictive model versus the selected features was plotted. The average accuracy of the cross-validation increase and the average accuracy of the predictive model increase simultaneously when more features are added. If the average accuracy of cross-validation is growing while the average accuracy of the predictive model starts to decrease, overfitting occurs at the point where the average accuracy of the predictive model starts to drop. Figures 7.11 (a)-(f) show the average accuracy of the predictive model and cross-validation versus the selected features by the rMI-SVM algorithm for colon cancer data set, leukaemia data set, prostate cancer data set, lung cancer data set, Parkinson data set and breast data set. The rMI-SVM algorithm had selected seven features for colon cancer data set; the average accuracy of the predictive model shows a strictly increasing graph when the number of features increases. The average accuracy of the cross-validation shows there is not much difference from the average accuracy of the predictive model. Therefore, the seven features selected by rMI-SVM algorithm in the predictive model did not show an overfitting issue.

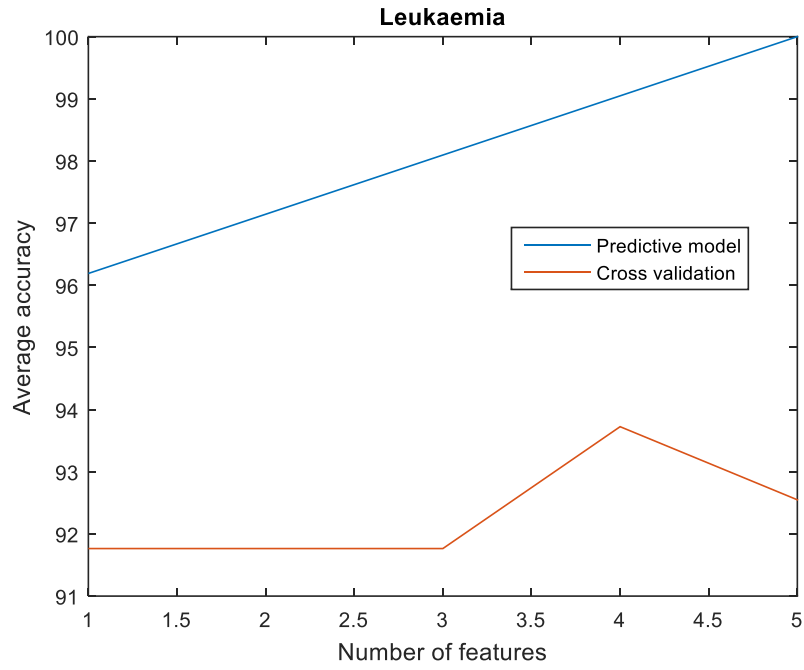
Five features were selected by the rMI-SVM algorithm for leukaemia data set, the average accuracy of the predictive model shows a strictly increasing graph when the number of features increases. The average accuracy of the cross-validation was lower than the average accuracy of the predictive model as the number of features increase. Therefore, the five features selected by rMI-SVM algorithm in the predictive model did not show an overfitting issue. The rMI-SVM algorithm had chosen three features for the prostate data set; the average accuracy of the predictive model shows a strictly increasing graph when the number of features increases. The average accuracy of the cross-validation was lower than the average accuracy of the predictive model as the number of features increase. Therefore, the three features selected by rMI-SVM algorithm in the predictive model did not show an overfitting issue. The rMI-SVM algorithm had chosen four features for the lung data set, the average accuracy of the predictive model shows a strictly increasing graph when the number of features increases. The average accuracy of the cross-validation was lower than the average accuracy of the predictive model as the number of features increase. Therefore, the four features selected by rMI-SVM algorithm in the predictive model did not show an overfitting issue.

The rMI-SVM algorithm had selected seven features for Parkinson data set, the average accuracy of the predictive model shows a strictly increasing graph when the number of features increases. The average accuracy of the cross-validation shows there is not much difference from the average accuracy of the predictive model. Therefore, the seven features selected by rMI-SVM algorithm in the predictive model did not show an overfitting issue.

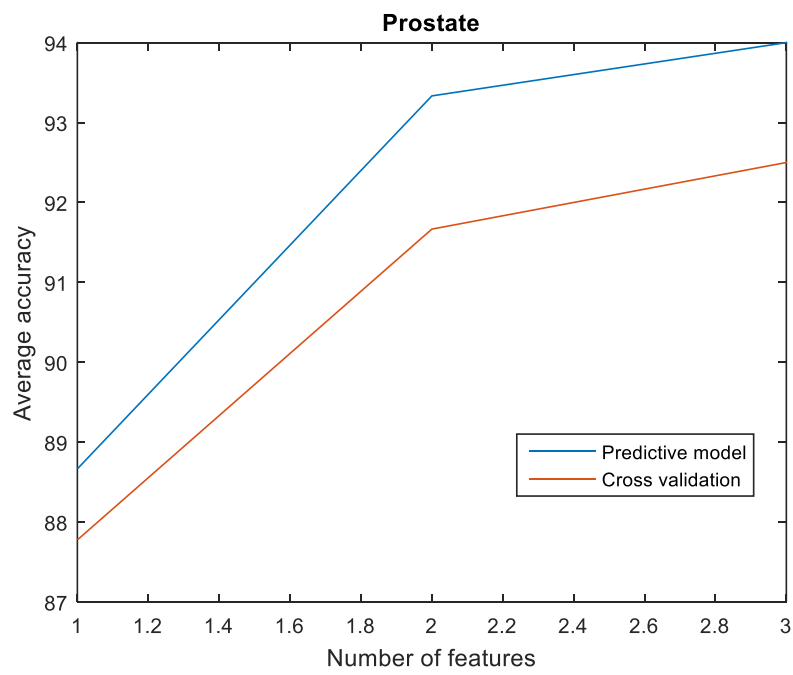
The rMI-SVM algorithm had chosen eleven features for breast cancer data set; the average accuracy of the predictive model shows a strictly increasing graph when the number of features increases. The average accuracy of the cross-validation was lower than the average accuracy of the predictive model as the number of features increase. Therefore, the eleven features selected by rMI-SVM algorithm in the predictive model did not show an overfitting issue.



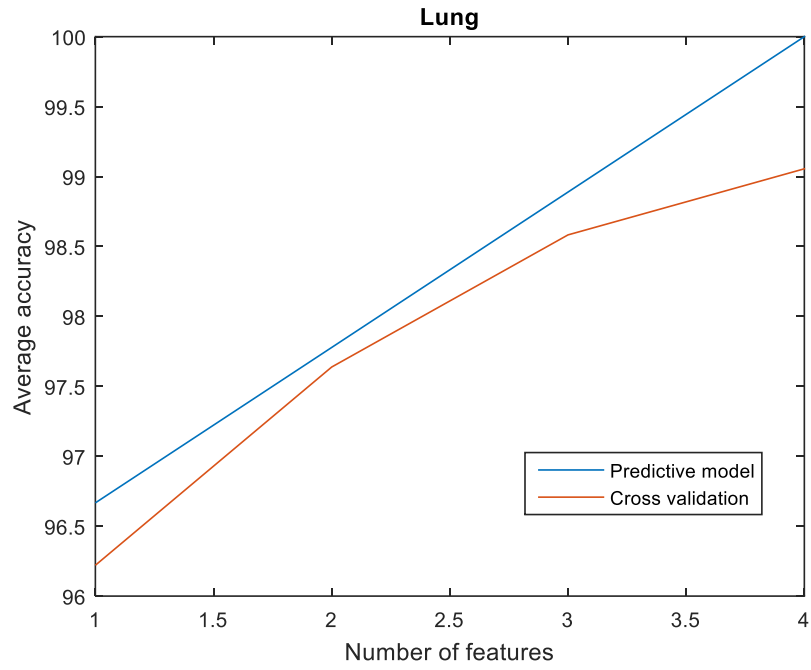
(a) colon cancer data set



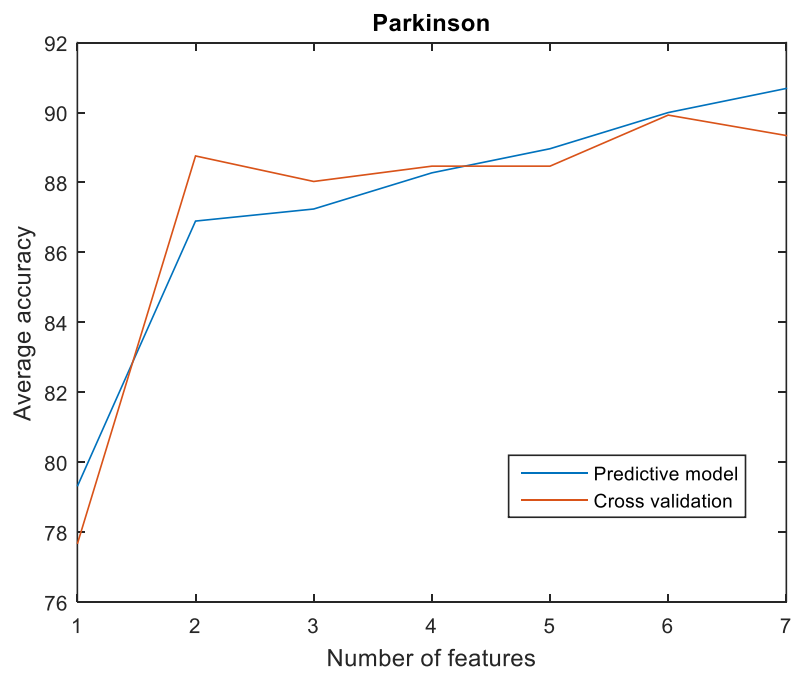
(b) leukaemia data set



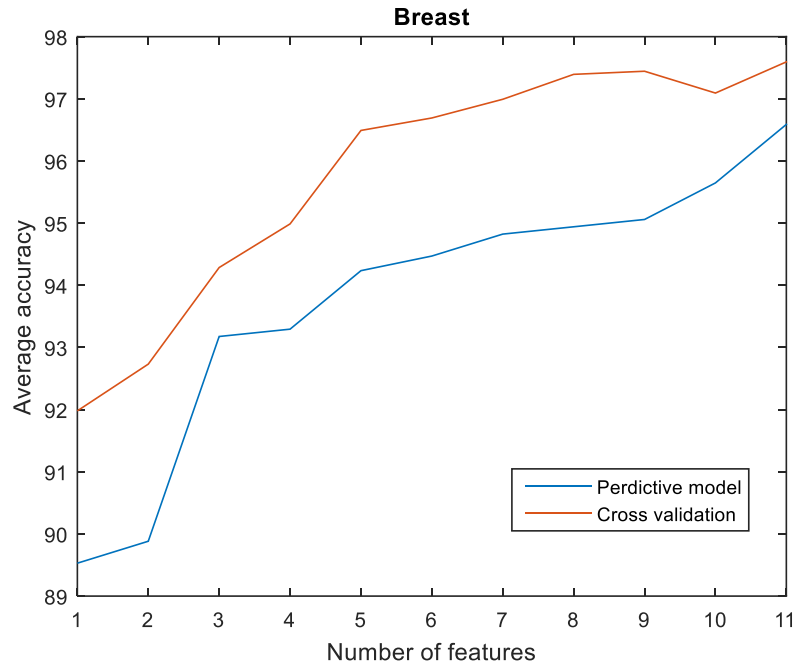
(c) prostate data set



(d) lung cancer data set



(e) Parkinson data set



(f) breast cancer data set

Figure 7.11: Average accuracy of the predictive model and cross-validation versus the selected features by the rMI-SVM algorithm for the binary data set

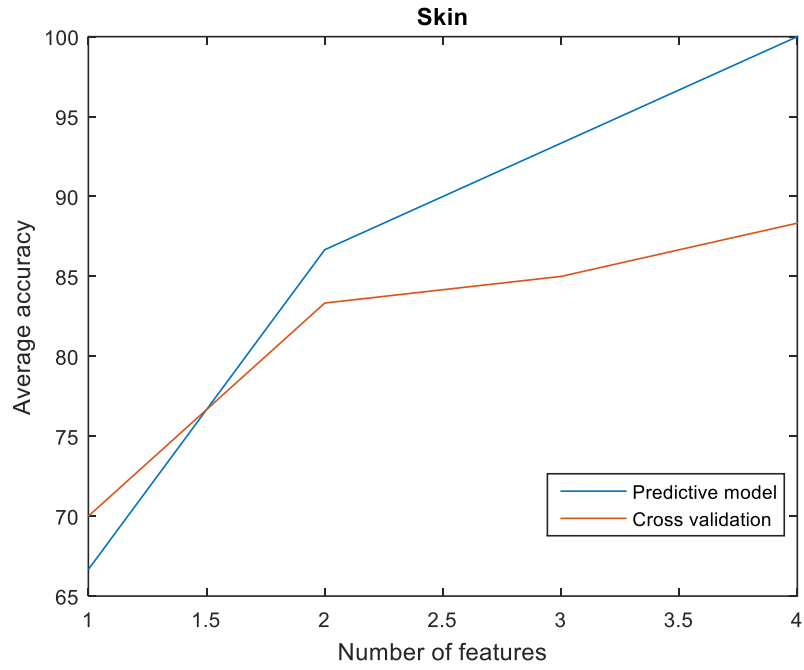
Overall, the Figures 7.11(a)-(f) show that the features selected by rMI-SVM algorithm did not show an overfitting issue in the predictive model for the colon cancer data set, leukaemia data set, prostate cancer data set, lung cancer data set, Parkinson data set and breast data set.

7.3.2 Cross-validation for Multiclass Data Set

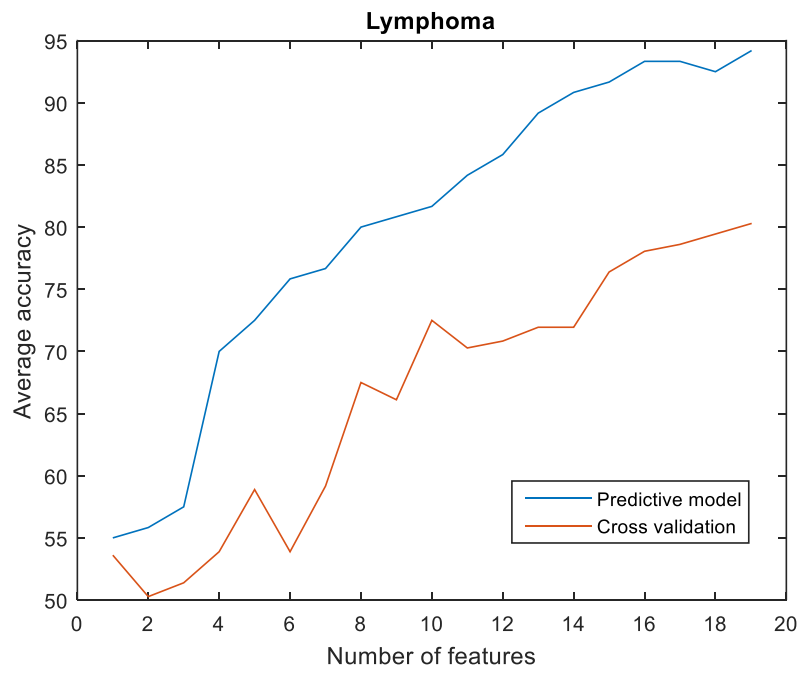
Figures 7.12 (a) - (d) show the average accuracy of the predictive model and cross-validation versus the selected features by the rMI-SVM algorithm for skin cancer data set, lymphoma data set, lung cancer data set and handwriting data set. The rMI-SVM algorithm had selected four features for

skin cancer data set; the average accuracy of the predictive model shows a strictly increasing graph when the number of features increases. The average accuracy of the cross-validation shows there is not much difference from the average accuracy of the predictive model. Therefore, the four features selected by rMI-SVM algorithm in the predictive model did not show an overfitting issue.

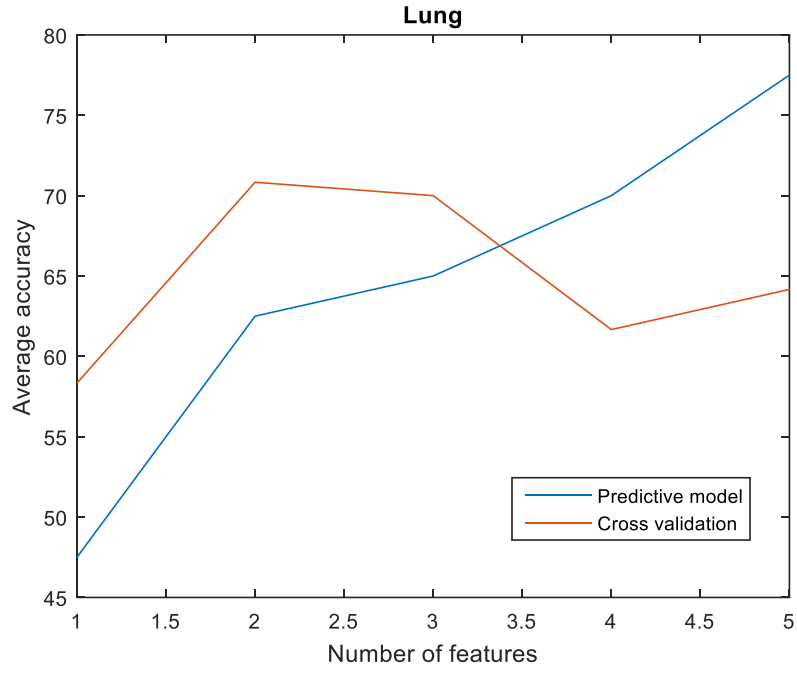
The rMI-SVM algorithm has selected twenty features for lymphoma data set; the average accuracy of the predictive model shows a strictly increasing graph when the number of features increases. The average accuracy of the cross-validation was lower than the average accuracy of the predictive model as the number of features increase. Therefore, the 20 features selected by rMI-SVM algorithm in the predictive model did not show an overfitting issue. The rMI-SVM algorithm had selected five features for lung cancer data set; the average accuracy of the predictive model shows a strictly increasing graph when the number of features increases. The average accuracy of the cross-validation shows there is not much difference from the average accuracy of the predictive model. Therefore, the five features selected by rMI-SVM algorithm in the predictive model did not show an overfitting issue. The rMI-SVM algorithm had chosen fifty features for handwriting data set, the average accuracy of the predictive model shows a strictly increasing graph when the number of features increases. The average accuracy of the cross-validation shows there is not much difference from the average accuracy of the predictive model. Therefore, the 50 features selected by rMI-SVM algorithm in the predictive model did not show an overfitting issue.



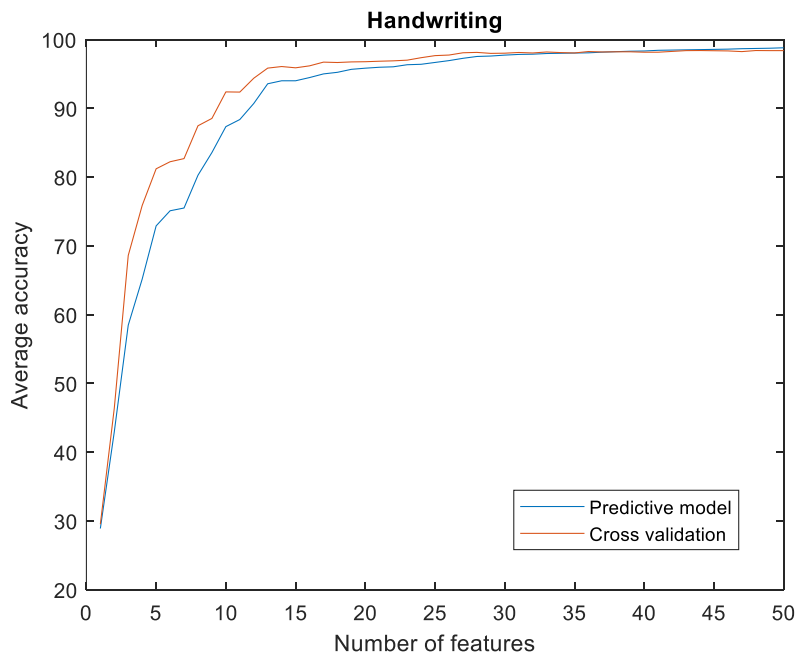
(a) skin cancer data set



(b) lymphoma data set



(c) lung data set



(d) handwriting data set

Figure 7.12: Average accuracy of the predictive model and cross-validation versus the selected features by the rMI-SVM algorithm for the multiclass data set

Overall, the Figures 7.12 (a)-(d) show that the features selected by rMI-SVM algorithm did not show an overfitting issue in the predictive model for skin cancer data set, lymphoma data set, lung cancer data set and handwriting data set.

7.4 Comparison of Performance

7.4.1 Comparison of Performance for Binary Data Set

The colon cancer data set contains 2000 variables with 62 samples, where 40 of the samples belong to tumour group, and 22 samples belong to the healthy group. The optimal baseline from the proposed algorithm in Chapter 3 is 87.78% with 202 ranked features. Table 7.11 shows the average accuracy with the number of features obtained from several feature selection methods for colon cancer data set. The rMI-SVM algorithm can achieve the same average accuracy with the baseline by using only seven features. With the same number of features, the regression method achieved the average accuracy of 82.22% while the mRMR method achieved the average accuracy of 81.11%. Additionally, several studies were conducted on the same data set using different feature selection methods. Yang and Moody (1999) proposed a feature selection using the joint mutual information (JMI). JMI achieved the highest average accuracy of 85.4% using 11 features.

Similarly, Zhu et al. (2007) proposed a Markov blanket-embedded genetic algorithm (MBEGA) for feature selection. MBEGA achieved the

highest average accuracy of 85.66% using 24 features. Joint mutual information maximisation (JMIM) was proposed by Mohamed Bennasar, Yulia Hicks and Rossitza Setchi (2015). JMIM achieved the highest average accuracy of 85.4% using 12 features. Furthermore, Vanitha et al. (2015) proposed a feature selection using support vector machine and mutual information (SVM-MI), SVM-MI selected three features and achieved the average accuracy of 80%.

Jun Wang et al. (2017) proposed a feature selection method using the criterion of maximum relevance and maximum independence (MRI). On an average of 50 groups of feature sets, MRI achieved the average accuracy of 86.84%. A dynamic change of selected feature with a class (DCSF) was proposed by Wanfu Gao, Liang Hu and Ping Zhang (2018). On an average of 30 groups of feature sets, DCSF achieved the average accuracy of 80.48%. Among the feature selection method discussed, the rMI-SVM algorithm is able to achieve the highest average accuracy using only seven features. Although SVM-MI proposed to select only three features, the average accuracy was only 80% which was the lowest among the feature selection method as listed in Table 7.11.

Table 7.11: Average accuracy with the number of features obtained from several feature selection methods for colon cancer data set

	No. of features	Average accuracy
Optimal baseline	202	87.78%
rMI-SVM	7	87.78%
Regression	7	82.22%
mRMR	7	81.11%
JMI	11	85.4%
MBEGA	24	85.66%
JMIM	12	85.4%
SVM-MI	3	80%
MRI	In Average	86.84%
DCSF	In Average	80.48%

The leukaemia data set contains 7128 variables with 72 samples, where 47 of the samples belong to acute lymphocytic leukaemia (ALL), and 25 belong to acute myeloid leukaemia (AML). The optimal baseline from the proposed algorithm in Chapter 3 is 99.05% with 38 ranked features. Table 7.12 shows the average accuracy with the number of features obtained from several feature selection methods for the leukaemia data set. The rMI-SVM algorithm can achieve the 100% average accuracy with only five features. With the same number of features, the regression method achieved the average accuracy of 98.1%, while the mRMR method achieved the average accuracy of 94.28%. There were several studies on the same data set using different feature selection methods. Howard Hua Yang and John Moody (1999) proposed a feature selection using the joint mutual information (JMI). JMI achieved the highest average accuracy of 99% using five features.

Zexuan Zhu, Yew-Soon Ong and Manoranjan Dash (2007) proposed a Markov blanket-embedded genetic algorithm (MBEGA) for feature selection. MBEGA achieved the highest average accuracy of 95.89% using 12 features. A hybrid feature selection using the filters method and wrappers method was also proposed by Hsu et al. (2011). The proposed method achieved the highest average accuracy of 98.6% using 70 features. Joint mutual information maximisation (JMIM) was proposed by Mohamed Bannasar, Yulia Hicks and Rossitza Setchi (2015). JMIM achieved the highest average accuracy of 97.25% using four features. Among the feature selection method that was discussed, the rMI-SVM algorithm is able to achieve the highest average accuracy using only five features.

Table 7.12: Average accuracy with the number of features obtained from several feature selection methods for leukaemia data set

	No. of features	Average accuracy
Optimal baseline	38	99.05%
rMI-SVM	5	100%
Regression	5	98.1%
mRMR	5	94.28%
JMI	5	99%
MBEGA	12	95.89%
Hybrid	70	98.6%
JMIM	4	97.25%

The prostate data set contains 2135 variables with 102 samples, of which 52 samples belong to the tumour group, and 50 samples belong to the healthy group. The optimal baseline from the proposed algorithm in Chapter 3

is 94% with 330 ranked features. Table 7.13 shows the average accuracy with the number of features obtained from several feature selection methods for prostate data set. The rMI-SVM algorithm can achieve the 93.33% average accuracy with only three features. With the same number of features, the regression method produced the same average accuracy of 93.33% while the mRMR method achieved the average accuracy of 88%. Zexuan Zhu, Yew-Soon Ong and Manoranjan Dash (2007) proposed a Markov blanket-embedded genetic algorithm (MBEGA) for feature selection. MBEGA achieved the highest average accuracy of 95.89% using 12 features. Therefore, the rMI-SVM algorithm is able to achieve the highest average accuracy using only three features.

Table 7.13: Average accuracy with the number of features obtained from several feature selection methods for prostate data set

	No. of features	Average accuracy
Optimal baseline	330	94%
rMI-SVM	3	93.33%
Regression	3	93.33%
mRMR	3	88%

The lung cancer data set contains 1626 variables with 181 samples, of which 31 samples belong to malignant pleural mesothelioma (MPM), and 150 samples belong to adenocarcinoma (AD). The optimal baseline from the proposed algorithm in Chapter 3 is 100% with nine ranked features. Table 7.14 shows the average accuracy with the number of features obtained from several feature selection methods for lung cancer data set. The rMI-SVM

algorithm can achieve the 100% average accuracy with only four features. With the same number of features, the regression method and mRMR achieved the average accuracy of 99.26%. Zexuan Zhu, Yew-Soon Ong and Manoranjan Dash (2007) proposed a Markov blanket-embedded genetic algorithm (MBEGA) for feature selection. MBEGA achieved the highest average accuracy of 98% using 24 features. Among the feature selection methods that were discussed, the rMI-SVM algorithm is able to achieve the highest average accuracy by using only four features.

Table 7.14: Average accuracy with the number of features obtained from several feature selection methods for lung cancer data set

	No. of features	Average accuracy
Optimal baseline	9	100%
rMI-SVM	4	100%
Regression	4	99.26%
mRMR	4	99.26%
MBEGA	24	98%

The Parkinson data set contains 22 variables with 195 samples, of which 147 samples belong to the healthy_group, and 48 samples belong to Parkinson disease group. The optimal baseline from the proposed algorithm in Chapter 3 is 88.97% with 22 features. Table 7.15 shows the average accuracy with the number of features obtained from several feature selection methods for Parkinson data set. The rMI-SVM algorithm can achieve the 90.69% average accuracy with seven features. With the same number of features, the regression method achieved the average accuracy of 84.83% while the mRMR method achieved the average accuracy of 86.21%.

There were several studies on the same data set using different features selection method. Howard Hua Yang and John Moody (1999) proposed a feature selection using the joint mutual information (JMI). JMI achieved the highest average accuracy of 89.5% using three features. Joint mutual information maximisation (JMIM) was proposed by Mohamed Bennisar, Yulia Hicks and Rossitza Setchi (2015). JMIM achieved the highest average accuracy of 91% using eight features. Among the feature selection method that was discussed, the rMI-SVM algorithm is able to achieve the average accuracy by using only seven features. Although the highest average accuracy achieved by JMIM, JMIM used eight features to achieve this average accuracy while rMI-SVM algorithm used only seven features to achieve similar average accuracy.

Table 7.15: Average accuracy with the number of features obtained from several feature selection methods for Parkinson data set

	No. of features	Average accuracy
Optimal baseline	22	88.97%
rMI-SVM	7	90.69%
Regression	7	84.83%
mRMR	7	86.21%
JMI	3	89.5%
JMIM	8	91%

The breast data set contains 30 variables with 569 samples, of which 357 samples belong to the benign_group, and 212 samples belong to the malignant group. The optimal baseline from the proposed algorithm in Chapter 3 was 95.06% with 13 ranked features. Table 7.16 shows the average

accuracy with the number of features obtained from several feature selection methods for breast cancer data set. The rMI-SVM algorithm can achieve the highest average accuracy of 96.59% by using eleven features. With the same number of features, the regression method achieved the average accuracy of 92.82% while the mRMR method achieved the average accuracy of 93.29%. There were several studies on the same data set using different features selection method.

Howard Hua Yang and John Moody (1999) proposed a feature selection using the joint mutual information (JMI). JMI achieved the highest average accuracy of 95.8% using 20 features. Joint mutual information maximisation (JMIM) was proposed by Mohamed Bannasar, Yulia Hicks and Rossitza Setchi (2015). JMIM achieved the highest average accuracy of 96.2% using five features. Among the feature selection methods that were discussed, the rMI-SVM algorithm is able to achieve the highest average accuracy by using eleven features.

Table 7.16: Average accuracy with the number of features obtained from several feature selection methods for breast cancer data set

	No. of features	Average accuracy
Optimal baseline	13	95.06%
rMI-SVM	11	96.59%
Regression	11	92.82%
mRMR	11	93.29%
JMI	20	95.8%
JMIM	5	96.2%

7.4.2 Comparison of Performance for Multiclass Data Set

The skin data set contains 22215 variables with 15 samples, of which six samples belong to the healthy group, four belong to actinic keratosis group, and five samples belong to squamous cell carcinoma group. The optimal baseline from the proposed algorithm in Chapter 3 is 86.67% with two ranked features. Table 7.17 shows the average accuracy with the number of features obtained from several feature selection methods for skin cancer data set. The rMI-SVM algorithm can achieve 100% average accuracy by using only four features. With the same number of features, the regression method achieved the average accuracy of 86.67%. Among the feature selection method that was discussed, the rMI-SVM algorithm is able to achieve the highest average accuracy by using only four features.

Table 7.17: Average accuracy with the number of features obtained from several feature selection methods for skin cancer data set

	No. of features	Average accuracy
Optimal baseline	2	86.67%
rMI-SVM	4	100%
Regression	4	86.67%

The lymphoma data set contains 4026 variables with 96 samples. There were nine classes in this data set. The optimal baseline from the proposed algorithm in Chapter 3 is 99.17% with 460 ranked features. Table 7.18 shows the average accuracy with the number of features obtained from several feature selection methods for the lymphoma data set. The rMI-SVM

algorithm can achieve the same average accuracy with the baseline by using only 22 features. With the same number of features, the regression method achieved the average accuracy of 55%. There were several studies on the same data set using different features selection method. Howard Hua Yang and John Moody (1999) proposed a feature selection using the joint mutual information (JMI). JMI achieved the highest average accuracy of 91% using 55 features.

Furthermore, Zexuan Zhu, Yew-Soon Ong and Manoranjan Dash (2007) proposed a Markov blanket-embedded genetic algorithm (MBEGA) for feature selection. MBEGA achieved the highest average accuracy of 97.68% using 34 features. Joint mutual information maximisation (JMIM) was proposed by Mohamed Bannasar, Yulia Hicks and Rossitza Setchi (2015). JMIM achieved the highest average accuracy of 91% using 59 features. In the same year, Devi A.V.C, Devaraj D, and Venkatesulu M proposed a feature selection using support vector machine and mutual information (SVM-MI). SVM-MI selected four features and achieved the average accuracy of 48.33%. Among the feature selection methods that were discussed, the rMI-SVM algorithm is able to achieve the average accuracy which is similar with the baseline by using only 22 features.

Table 7.18: Average accuracy with the number of features obtained from several feature selection methods for lymphoma data set

	No. of features	Average accuracy
Optimal baseline	460	99.17%
rMI-SVM	22	99.17%
Regression	22	55%
JMI	55	91%
MBEGA	34	97.68%
JMIM	59	91%
SVM-MI	4	48.33%

The lung cancer data set contains 1626 variables with 181 samples, of which 31 of the samples belong to malignant pleural mesothelioma (MPM), and 150 samples belong to adenocarcinoma (AD). The optimal baseline from the proposed algorithm in Chapter 3 is 70% with 51 ranked features. Table 7.19 shows the average accuracy with the number of features obtained from several feature selection methods for lung cancer data set. The rMI-SVM algorithm can achieve the 77.5% average accuracy by using only five features. With the same number of features, the regression method achieved the average accuracy of 67.5%. The rMI-SVM algorithm can achieve the highest average accuracy by using only five features.

Table 7.19: Average accuracy with the number of features obtained from several feature selection methods for lung cancer data set

	No. of features	Average accuracy
Optimal baseline	51	70%
rMI-SVM	5	77.5%
Regression	5	67.5%

The handwriting data set contains 649 variables with 2000 samples. This data set consists of the number of handwritten features from zero to nine, and there were ten classes in this data set. The optimal baseline from the proposed algorithm in Chapter 3 is 98.3% with 434 ranked features. Table 7.20 shows the average accuracy with the number of features obtained from several feature selection methods for handwriting data set. The rMI-SVM algorithm is able to achieve the highest average accuracy of 98.8% by using only 50 features. With the same number of features, the regression method achieved the average accuracy of 96.5%. Several studies were conducted on the same data set using different features selection method. Additionally, Howard Hua Yang and John Moody (1999) proposed a feature selection method using the joint mutual information (JMI). JMI achieved the highest average accuracy of 97% by using 33 features. A joint mutual information maximisation (JMIM) was proposed by Mohamed Bennisar, Yulia Hicks and Rossitza Setchi (2015). JMIM achieved the highest average accuracy of 97.5% by using 39 features.

Among the feature selection methods that were discussed, the rMI-SVM algorithm was able to achieve the highest average accuracy using only 50 features. Although JMI produced the average accuracy of 97% by using only 33 features while using the same number of features, rMI-SVM algorithm was able to achieve the average accuracy of 97.97%. JMIM achieved the average accuracy of 97.5% by using only 39 features while using the same number of features; whereas the rMI-SVM algorithm can achieve the average accuracy of 98.3%.

Table 7.20: Average accuracy with the number of features obtained from several feature selection methods for handwriting data set

	No. of features	Average accuracy
Optimal baseline	434	98.3%
rMI-SVM	50	98.8%
Regression	50	96.5%
JMI	33	97%
JMIM	39	97.5%

A total of ten data sets had been tested on the performance of the prediction model using the proposed method, rMI-SVM algorithm. Among the ten data sets, there were six binary classification data set and four multiclass classification data set. The rMI-SVM algorithm had shown an excellent performance using the features selected compared to using all features or using the same number of features in the regression method and mRMR method. Besides that, the performance of the ten data sets were also compared with other feature selection methods such as JMI, MBEGA, JMIM, SVM-MI, MRI, DCFS. The features selected using mutual information were able to retain the maximum information in the predictive model. The sensitivity of the classifier using the selected features test used the ROC curve, and all the ten data sets gave a high recall value. At the same time, the classifier using the selected features also provide high value in the area under the curve. Overall, the rMI-SVM algorithm is able to choose a compact subset of features that yields better performance in the predictive model.

7.5 Biological Meaning of the Selected Features

The proposed algorithm rMI-SVM is able to select the most informative and less redundant features that maximise the performance of the predictive model by ensuring that the next candidate feature chosen will provide new information to the predictive model. Therefore, it is necessary to do further biological analysis on the selected features to the relevant diseases. In this sub-chapter, the features chosen by the proposed algorithm on colon cancer data will further investigate on its biological meaning. There are seven features selected by the proposed algorithm, and the biological description of each feature was shown in Table 7.21.

Table 7.21: Features selected by rMI-SVM for colon cancer data set

No.	Description
1	Human 20-kDa myosin light chain mRNA
2	Homo sapiens cysteine-rich protein gene
3	Homo sapiens desmin gene
4	Thioredoxin (Human)
5	Homo sapiens mRNA for GCAP-II/uroguanylin precursor
6	Alpha Enolase (Human)
7	Major Histocompatibility Complex Enhancer-Binding Protein MAD3

The first feature selected by the proposed algorithm for the colon data set is the most informative because this feature has the highest mutual information score. Human 20-kDa myosin light chain mRNA has been related to colon cancer in several studies (Hadas et al., 1986; Kumar et al., 1989). The

second selected feature is homo sapiens cysteine-rich protein gene, and it is a peptide that exists in the colon (O'Dell, 1992). The next feature is the homo sapiens desmin gene. It was shown in immunohistochemical studies that as the tumour cell increases, the desmin was also present (Hinton and Halliday, 1984). Thioredoxin presents a high level in the cancer cells (Kontou et al., 2004). Experimental studies have shown that thioredoxin contributes to the growth and transformation in the cancer cells. Therefore, by reducing the thioredoxin level, it can regulate the growth of cancer cells (Gallegos et al., 1997).

Homo sapiens mRNA for GCAP-II/uroguanylin precursor activates the guanylate cyclase receptor of the colon cells (Hess et al., 1995). Northern hybridisation showed that the expression of homo sapiens mRNA for GCAP-II/uroguanylin precursor is the highest in the colon (Mägert et al., 1998). It is also shown that this expression plays a vital role in mediated functions in the colon (Hill et al., 1995). The alpha-enolase shows high activities in colon cancer compared to the normal cell (Durany et al., 1997). Major Histocompatibility Complex Enhancer-Binding Protein MAD3 is an antigen-presenting immune cell. A study has shown that it enhances the addition of tumour necrosis factor (Rahat et al., 2001). The seven selected features by rMI-SVM in colon cancer are biologically relevant to colon cancer.

7.6 New Findings

Some new findings were developed through the study of the proposed methods, using mutual information. The new benchmark of the average accuracy is determined by plotting the graph of average accuracy versus the cumulative ranked features. From the proposed algorithm in Chapter 3, the features were ranked according to their relevancy measured by the mutual information score. The higher the score indicates the more information content of the feature in respect to the label class. Therefore, the cumulative ranked features give extra information on two critical information: (1) a new benchmark on the average accuracy of the data set, (2) the number of features required to achieve this new benchmark. The benchmark of the data set provides essential measurement of how good is the classification model and serves as the baseline before feature selection. Researchers can then compare the classification result using various feature selection method with this new benchmark. At the same time, the number of features required to achieve this new benchmark also play an essential role. A classification model that uses more than this number of features to obtain this new benchmark will not be considered as an efficient model.

In past research, the researcher always uses the full features to achieve the benchmark of the average accuracy of the predictive model. However, when all the features are included, the irrelevance and noise are also added to the predictive model. Besides this, there is no guideline on how many features are needed in a classification model or guideline on the number of features

that can be selected. The past researches always show a range of several features versus the average accuracy and leave a problem on how many features should be taken to build a predictive model.

From the second proposed algorithm discussed in Chapter 4, the fast dimension reduction algorithm - rMI-SVM algorithm allows the selection of a compact of features that are needed to produce the same or better performance compared to the new benchmark in a few iterations. The rMI-SVM algorithm can reduce more than 70% of the data in the first step that will make the feature selection process complete in a few step. The rMI-SVM algorithm is different from the iteration algorithm, such as the wrapper method. The wrapper method searches all the features in each iteration to select the next feature. In contrast, the rMI-SVM algorithm only selects the remaining features from the previous step after filtering out the redundant features and noise. By using the rMI-SVM algorithm, the algorithm will filter out those features not giving new information and select only those features that will provide some improvement to the average accuracy of the predictive model.

The selected features using the rMI-SVM algorithm will not face many problems on overfitting, as the next chosen new candidate feature will provide further information to the predictive model and increase the performance of the predictive model, which means that a strictly growing curve for the average accuracy of the predictive model versus the number of features. Therefore, when plotting the average accuracy of the predictive model and

cross-validation versus the number of features, there was no such point where the average accuracy of cross-validation is increasing while the average accuracy of the predictive model starts to decrease, the point that overfitting occurs.

7.7 Contribution of Feature Selection in Malaysia's Medical Research

In clinical data, the feature selection was one of the ways to identify the biomarker. Biomarkers are great significance for the research and development of life science, as well as medical diagnosis, clinical treatment and new drug development. It helps researchers to be more effective in diagnosis or treatment, especially in the prevention and control of complex and chronic diseases such as cancer, cardiovascular disease, diabetes, etc. However, biomarker needs to go through the clinical test or validation test in the lab after the theoretical analysis. Therefore, a smaller number of features will be preferable as running a clinical test for one feature may need plenty of time. They do not have time to run the clinical examination for full features and the combination of features or trial and error. Clinical validity must be established before a biomarker is used in the clinic. The proposed algorithm in Chapter 3 helps the researchers to get a better baseline using the ranked features by mutual information score. This new proposed baseline can give a rough idea on how best will the classification be, as the performance of the predictive model depends on the information contained in the data set.

The rMI-SVM algorithm can reduce the high dimensionality of the data set and obtain a compact subset of features that yield a better performance or lower the classification rate. In most of the proposed feature selection method, the first selected feature can achieve accuracy up to 70% on average. Still, in the clinical test for a biomarker, it is safer to have a combination of a group of features (typically 2-5 proteins), instead of depending on only one protein. For example, a new combination of four proteins (APOE, ITIH3, APOA1 and APOL1) form an accurate biomarker diagnosis of pancreatic cancer as early diagnosis and early treatment may improve prognosis and increase the survival chance.

In Malaysia, most of the cancer diagnosis was detected at a late stage. Therefore, there was an urgent need in Malaysia on developing the new biomarker for cancer diagnosis and other medical diagnoses such as stroke, hypertension, kidney disease, etc. In the latest 2019 Budget of Malaysia, the government will continue programmes such as free mammograms to detect breast cancer, HPV vaccinations, and pap smear tests at government hospitals and clinics. The plans, with an allocation of RM20 million, is expected to benefit as many as 70,000 women. Malaysia's badminton player Dato Lee Chong Wei was detected early for nasal cancer in July 2018. After two month's treatment in Taiwan, now Dato Lee Chong Wei has recovered. Evidently, this shows that early detection of cancer can increase the chance of survival.

Most of the past researchers did not study in depth about the minimum number of features needed in the predictive model, as most of the time the proposed method only shows the accuracy of the predictive model for a range of k . Strictly speaking, the past researcher did not explain how to choose the minimum number of k . The selection of k was based on the researchers, and usually, the range of k was from 1 up to 50 features. The proposed algorithm in Chapter 3 is able to solve this problem by getting the number of features to obtain the optimal baseline.

CHAPTER 8

CONCLUSION

8.1 Summary

The mutual information has been used in this study. There are advantages in using mutual information in feature selection, as mutual information can evaluate both linear and nonlinear dependencies among the features. Besides this, the mutual information score can be easily calculated regardless of the linearity of the features and even the data have some missing values. Besides this, no normal assumption was applied here, unlike some statistical method such as t-test or ANOVA, where the data set must fulfil the normal assumption. The proposed algorithm in Chapter 3 gives an optimal baseline by using the ranked cumulative features with mutual information score. The features were ranked according to the mutual information, and the higher the mutual information score, the more information are contained in that feature. The result shows that the optimal baseline is better than the current benchmark that involved all the features because when all the features are included, the irrelevant, redundant features and noise will also be included as well. At the same time, the number of features to achieve this optimal baseline can be obtained. The number of features will serve as a guideline on how many features are needed in a predictive model.

Secondly, the rMI-SVM algorithm reduces the high dimensional data and is able to select a compact subset of features. The rMI-SVM algorithm filters out the redundant features and noise according to the newly added information in the predictive model. When a new candidate feature is added to the predictive model, the feature will then be selected if this feature provides additional information to the predictive model. Therefore, a strictly increasing performance graph will be obtained with minimal features. The predictive model will face less overfitting as the next added candidate feature must provide some information to the predictive model. The mutual information can help in feature selection to retain the maximum information in the predictive model. In the rMI-SVM algorithm, the redundancy of the features will be minimised, and at the same time, the dynamic change of the already selected features with the newly added candidate feature to the label class will be considered.

The Z-score measured the robustness of the proposed rMI-SVM algorithm and indicated that the features selected by the rMI-SVM algorithm were not chosen by chance. Later, the sensitivity test of the predictive model has been measured using the ROC curve and the area under curve was used to measure the effectiveness of the built model. The performance of the predictive model using the same number of selected features has also been compared with other feature selection method such as the regression method, mRMR method, JMI, JMIM, DCSF, MRI etc.

The biological meaning of the features selected in the colon data set has been studied and confirmed that the chosen features are related to colon cancer. These selected features can further help in finding the potential features for biomarkers (Miyazaki et al., 2016). Biomarkers are important in medical diagnosis, clinical treatment and new drug development. With the help of the biomarker, cancer diseases and other medical diseases can easily be detected and help in clinical treatment to treat the patient more effectively, improve the healing effects, survival rate and reduce the chance of recurrence. It is a big step forward in the medical field. In this way, people can be healthier, the country will be more prosperous, and the world will become better.

8.2 Limitation

In this study, when comes to a tie condition, where the mutual information score is similar, then the selected features will be based on the indexed order, and the feature with the lowest index will then be chosen. Similarly, when the additional improvement on the average accuracy was the same for two features, then the feature with higher mutual information score will then be selected. The sample size of the microarray data was always small, and due to this issue, not much information can be obtained from the small sample data set. Boost strapping method failed to help to increase the amount of information contained in a small sample data set as the information only comes from the original raw data. Besides this, the imbalanced class of the data set was not considered in this study. Therefore, for a multiclass data set

with small sample size, the predictive model might be biased to the particular label class.

8.3 Future Study

Further study on the relationship between the selected features can be further investigated by building a network model using mutual information. Furthermore, ordering of the features has always been an issue in feature selection. Therefore, further study on how to select the feature when the features selection comes to a tie condition is very important as the contribution of these features in respect to the label class might differ. Also, to increase the sample size of the data set and to increase the amount of information contained in the data set, merging the two data sets will be the next challenging task. Therefore, this is an important issue and there is an urgent need in feature selection with small sample size. Other than that, the current algorithm can be improved by searching more relevant features by combining with other selection methods. Last but not least, the imbalanced class issue also plays a vital role in feature selection to avoid the bias of the predictive model to a particular class.

REFERENCES

- Aguilar-Ruiz J.S., Azuaje F. and Riquelme, J.C., 2004. Data mining approaches to diffuse large B-Cell lymphoma gene expression data interpretation. *Lecture Notes in Computer Science*, 3181, pp. 279-288.
- Aksakalli V. and Malekipirbazari M., 2016. Feature selection via binary simultaneous perturbation stochastic approximation. *Pattern Recognition Letters*, 75, pp. 41–47.
- Alon U., Barkai N., Notterman D. A., Gish K., Ybarra S., Mack D. and Levine A.J., 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America*, 96, 6745–6750.
- Ang J.C., Mirzal A., Haron H. and Hamed H.N.A., 2016. Supervised, unsupervised, and semi-supervised feature selection: A Review on Gene Selection. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 13, pp. 971-989.
- Aranda E., Aparicio J., Alonso V., Garcia-Albeniz X., Garcia-Alfonso P., Salazar R., Valladares M., Vera R., Vieitez J. M., and Garcia-Carbonero R., 2015. SEOM clinical guidelines for diagnosis and treatment of metastatic colorectal cancer 2015. *Clinical & translational oncology: official publication of the Federation of Spanish Oncology Societies and of the National Cancer Institute of Mexico*, 17(12), pp. 972–981.
- Avery O.T., MacLeod C.M. and McCarty M., 1944. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *J Exp Med*, 79, pp. 137–158.
- Battiti R., 1994. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5, pp. 537–550.
- Bir-Jmel A., Douiri S. M. and Elbernoussi S., 2019. Gene selection via a new hybrid ant colony optimization algorithm for cancer classification in high-dimensional data. *Computational and Mathematical Methods in Medicine*, 2019: 7828590.
- Bennasar M., Hicks Y. and Setchi R., 2015. Feature selection using joint mutual information maximisation. *Expert Systems with Applications*, 42, pp. 8520-8532.
- Bins J. and Draper B. A., 2001. Feature selection from huge feature sets. *Proceedings Eighth IEEE International Conference on Computer Vision*, 2001, pp. 159-165.

Bolón-Canedo V., Sánchez-Marño N. and Alonso-Betanzos A., 2013. A review of feature selection methods on synthetic data. *Knowledge and Information Systems*, 34, pp. 483–519.

Cai Z., Xu D., Zhang Q., Zhang J., Ngai S-M., and Shao J. C., 2015. Classification of lung cancer using ensemble-based feature selection and machine learning methods. *Molecular BioSystems*, 11, pp. 791–800.

Chou P.Y. and Fasman G.D., 1974. Conformational parameters for amino acids in helical, β -sheet, and random coil regions calculated from proteins. *Biochemistry*, 13 (2), pp. 211-222.

Cover T. and Thomas J., 2006. *Elements of information theory*. New York: John Wiley & Sons.

Dayhoff M.O. and Ledley R.S., 1962. Comproteins: a computer program to aid primary protein structure determination. *Proceedings of Fall Joint Computer Conference*, 4-6 December 1962 New York, NY: ACM, pp. 262–274.

Dayhoff M.O. (1965). *Atlas of Protein Sequence and Structure*, Vol. 1. Silver Spring, MD: National Biomedical Research Foundation.

Dayhoff M.O., Schwartz R.M. and Orcutt B.C., 1978. Chapter 22: a model of evolutionary change in proteins. *In: Atlas of Protein Sequence and Structure*. Washington. DC: National Biomedical Research Foundation.

Dessì N., Pascariello E. and Pes B. 2013. A comparative analysis of biomarker selection techniques. *BioMed research international*, 2013, pp. 1-10.

Durany N., Joseph J., Campo E., Molina R., and Carreras J., 1997. Phosphoglycerate mutase, 2,3-bisphosphoglycerate phosphatase and enolase activity and isoenzymes in lung, colon and liver carcinomas. *Br J Cancer*, 75(7), pp. 969-977.

Elyasigomari V., Lee D.A., Screen R.C. and Shaheed M.H., 2017. Development of a two-stage gene selection method that incorporates a novel approach using the cuckoo optimization algorithm and harmony search for cancer classification. *Journal of Biomedical Informatics*, 67, pp. 11-20.

Estévez P. A., Tesmer M., Perez A. and Zurada J.M., 2009. Normalized mutual information feature selection. *IEEE Transactions on Neural Networks*, 20, pp. 189–201.

Feng D.F. and Doolittle R.F., 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol*, 25, pp. 351–360.

Fleischmann R.D., Adams M. D., White O., Clayton R. A., Kirkness E. F., Kerlavage A. R., Bult C. J., Tomb J. F., Dougherty B. A., Merrick J. M., McKenney K., Sutton G., FitzHugh W., Fields C., Gocyne J. D., Scott J.,

Shirley R., Liu L. I., Glodek A., Kelley J. M., Weidman J. F., Phillips C. A., Spriggs T., Hedblom E., Cotton M. D., Utterback T. R., Hanna M. C., Nguyen D. T., Saudek D. M., Brandon R. C., Fine L. D., Fritchman J. L., Fuhrmann J. L., Geoghagen N. S. M., Gnehm C. L., McDonald L. A., Small K. V., Fraser C. M., Smith H. O., and Venter J. C., 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, 269, pp. 496–512.

Gallegos A., Berggren M., Gasdaska J.R., and Powis G., 1997. Mechanisms of the regulation of thioredoxin reductase activity in cancer cells by the chemopreventive agent selenium. *Cancer Res*, 57(21), pp. 4965-4970.

Gao W., Hu L. and Zhang P., 2018. Class- specific mutual information variation for feature selection. *Pattern Recognition*, 79, pp. 328-339.

Gauthier, J., Vincent, A.T., Charette, S.J., and Derome, N., 2018. A brief history of bioinformatics. *Briefings in Bioinformatics*, pp. 1-16.

Golub T. R., Slonim D.K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J. P., Coller H., Loh M. L., Downing J. R., Caligiuri M. A., Bloomfield C. D., and Lander E. S., 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286, pp. 531-537.

Guyon I., Weston J., Barnhill S. and Vapnik V., 2002. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46, pp. 389–422.

Hadas E., Fink A., Gembom E., Harpaz, N., Shani, A., Bentwich, Z., and Eshhar, Z., 1986. Organ specific neoantigens reactive in the leukocyte adherence inhibition assay: affinity purification of human colon carcinoma antigen and its cross-reactive protein using monoclonal antibodies. *Cancer Res*, 46(10), pp. 5201-5205.

Hershey A.D. and Chase M., 1952. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J Gen Physiol*, 36, pp. 39–56.

Hess R., Kuhn M., Schulz-Knappe P., Raida M., Fuchs M, Klodt J., Adermann K., Kaefer V., Cetin Y., and Forssmann W., 1995. GCAP-II: isolation and characterization of the circulating form of human uroguanylin. *FEBS Lett.*, 374(1), pp. 34-38.

Hill O., Cetin Y., Cieslak A., Mägert H.J., and Forssmann W.G., 1995. A new human guanylate cyclase-activating peptide (GCAP-II, uroguanylin): precursor cDNA and colonic expression. *Biochim Biophys Acta.*, 1253(2), pp. 146-149.

Hinton D.R. and Halliday W.C., 1984. Primary rhabdomyosarcoma of the cerebellum--a light, electron microscopic, and immunohistochemical study. *J Neuropathol Exp Neurol*, 43(4), pp. 439-449.

Hong Z. and Yang J., 1991. Optimal discriminant plane for a small number of samples and design method of classifier on the plane. *Pattern Recognition*, 24, pp. 317-324.

Hoque N., Bhattacharyya D.K. and Kalita J.K., 2014. MIFS-ND: a mutual information based feature selection method. *Expert Systems with Applications*, 41, pp. 6371–6385.

John M., 2005. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Current opinion in structural biology*, pp. 285–289.

Kontou M., Will R.D., Adelfalk C., Wittig R., Poustka A., Hirsch-Kauffmann M. and Schweiger M, 2004. Thioredoxin, a regulator of gene expression. *Oncogene*, 23(12), pp. 2146-2152.

Kwak N. and Choi C., 2002. Input feature selection for classification problems. *IEEE Transactions on Neural Networks*,13, pp. 143–159.

Kukreja S.J., Lofberg J. and Brenner M.J., 2006. A least absolute shrinkage and selection operator (LASSO) for nonlinear system identification. *IFAC Proceedings Volumes*, 39, pp. 814-819.

Kumar C.C., Mohan S.R., Zavodny P.J., Narula S.K., and Leibowitz P.J., 1989. Characterization and differential expression of human vascular smooth muscle myosin light chain 2 isoform in nonmuscle cells. *Biochemistry*, 28(9), pp. 4027-4035.

Lewis, D. D., 1992. Feature selection and feature extraction for text categorization. *Proceedings of the workshop on speech and natural language*, pp. 212–217.

Li Z., Xie W. and Liu T., 2018. Efficient feature selection and classification for microarray data. *PLoS ONE*, 13(8):e0202167.

Liu X., Zheng W., Wang W., Shen H., Liu L., Lou W., Wang X. and Yang P., 2017. A new panel of pancreatic cancer biomarkers discovered using a mass spectrometry-based pipeline. *British Journal of Cancer*, 117, pp.1846–1854.

Mägert H.J., Reinecke M., David I., Raab H.R., Adermann K., Zucht H.D., Hill O., Hess R. and Forssmann W.G., 1998. Uroguanylin: gene structure, expression, processing as a peptide hormone, and co-storage with somatostatin in gastrointestinal D-cells. *Regul Pept.*, 73(3), pp. 165-176.

Maxam A.M. and Gilbert W., 1977. A new method for sequencing DNA. *Proc Natl Acad Sci USA*,74, pp. :560–564.

Michael L and Arieh W., 1975. Computer simulation of protein folding. *Nature*, 253, pp. 694–698.

Murata M., Richardson J.S. and Sussman J.L., 1985. Simultaneous comparison of three protein sequences. *Proc Natl Acad Sci USA*, 82, pp. 3073–3077.

Naseriparsa M., Bidgoli A. and Varae T., 2013. Improving Performance of a Group of Classification Algorithms Using Resampling and Feature Selection. *World of Computer Science and Information Technology Journal*, 3, pp. 70-76.

National Cancer Registry Department. Summary of Malaysian National Cancer registry report 2007-2011.

Needleman S.B. and Wunsch C.D., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48, pp. 443–53.

Nestor M. S. and Zarraga, M. B., 2012. The incidence of nonmelanoma skin cancers and actinic keratoses in South Florida. *The Journal of clinical and aesthetic dermatology*, 5, pp. 20–24.

O'Dell B.L., 1992. Cysteine-rich intestinal protein (CRIP): a new intestinal zinc transport protein. *Nutr Rev*, 50(8), pp. 232-233.

Papageorgis P., Ozturk S., Lambert A. W., Neophytou C. M., Tzatsos A., Wong C. K., Thiagalingam S. and Constantinou A. I., 2015. Targeting IL13Ralpha2 activates STAT6-TP63 pathway to suppress breast cancer lung metastasis. *Breast Cancer Research*, 17, pp. 1-15.

Pauling L. and Zuckerkandl E., 1963. Chemical paleogenetics: molecular “restoration studies” of extinct forms of life. *Acta Chem Scand*, 17, pp. 9–16.

Pedrycz, W. and Chen, S., 2020. *Deep Learning: Algorithms And Applications*. Springer, pp.118-121.

Peng H., Long F. and Ding C., 2005. Feature selection based on mutual information: criteria of max- dependency, max- relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27, pp. 1226–1238.

Podolsky M. D., Barchuk A.A., Kuznetsov V. I., Gusarova N. F., Gaidukov V. S. and Tarakanov S. A., 2016. Evaluation of machine learning algorithm utilization for lung cancer classification based on gene expression levels. *Asian Pac J Cancer Prev*, 17, pp. 835-838.

Rahat M.A., Chernichovski I., and Lahat N., 2001. Increased binding of IFN regulating factor 1 mediates the synergistic induction of CIITA by IFN-gamma and tumor necrosis factor-alpha in human thyroid carcinoma cells. *Int Immunol*, 13(11), pp. 1423-1432.

Ramani R. G. and Sivagami G., 2011. Parkinson Disease Classification using Data Mining Algorithms. *International Journal of Computer Applications*, 32, pp. 17-22.

Ranjan, P., Kumari, A., and Chakrawarty, A., 2015. How can doctors improve their communication skills?. *Journal of clinical and diagnostic research : JCDR*, 9(3), JE01–JE4.

Salama G.I., Abdelhalim M.B., and Zeid M.A., 2012. Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers. *International Journal of Computer and Information Technology*, 1, pp. 36-43.

Sanger F. and Thompson E.O.P., 1953. The amino-acid sequence in the glycol chain of insulin. I. The identification of lower peptides from partial hydrolysates. *Biochem J* ,53, pp. 353-366.

Sanger F. and Thompson E.O.P., 1953. The amino-acid sequence in the glycol chain of insulin. II. The investigation of peptides from enzymic hydrolysates. *Biochem J* ,53, pp. 366–374.

Sanger F., Nicklen S. and Coulson A.R., 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA*, 74, pp. 5463–5467.

Shannon C.E., 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27, pp. 379–423, 623–656.

Sievers F. and Higgins D.G., 2014. Clustal Omega, accurate alignment of very large numbers of sequences. *Methods Mol Biol*, 1079, pp. 105–116.

Sogawa K., Takano S., Lida F., Satoh M., Tsuchida S., Kawashima Y., Yoshitomi H., Sanda A., Koderia Y., Takizawa H., Mikata R., Ohtsuka M., Shimizu H., Miyazaki M., Yokosuka O. and Nomura F., 2016. Identification of a novel serum biomarker for pancreatic cancer, C4b-binding protein α -chain (C4BPA) by quantitative proteomic analysis using tandem mass tags. *British Journal of Cancer*, 115, pp. 949–956.

Staden R., 1979. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res*, 6, pp. 2601–2610.

Fitch W.M., 1970. Distinguishing homologous from analogous proteins. *Syst Zool*, 19, pp. 9–113.

University of Sheffield, 2016. Over 750 biomarkers identified as potentials for early cancer screening test. [Online]. Available at: www.sciencedaily.com/releases/2016/08/160801093000.htm [Accessed: 1 August 2016]

Uttley L., Whiteman B.L., Woods H.B., Harnan S. and Philips S. T., 2016. Building the evidence base of blood-based biomarkers for early detection of cancer: A Rapid Systematic Mapping Review. *EBioMedicine*, 10, pp. 164-173.

Vergara J. and Estévez P., 2014. A review of feature selection methods based on mutual information. *Neural Computing and Applications*, 24, pp. 175–186.

Vinh N. X. and Bailey J., 2013. Comments on supervised feature selection by clustering using conditional mutual information-based distances. *Pattern Recognition*, 46(4), pp.1220-1225.

Wang J., Wei J., Yang Z. and Wang S., 2017. Feature Selection by Maximizing Independent Classification Information. *IEEE Transactions on Knowledge and Data Engineering*, 29, pp. 828-841.

Watson J.D. and Crick F.H.C., 1953. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 171, pp. 737–738.

Yang H. and Moody J., 1999. Feature selection based on joint mutual information. In *Proceedings of international ICSC symposium on advances in intelligent data analysis*, pp.22–25.