Towards Breast Cancer Survivability Prediction Models through Multiple Hospital Information Systems

Tasos Mavrikeos

This thesis is presented in fulfillment of the requirements for the degree of Master of Philosophy

School of Engineering and Science
Faculty of Health, Engineering and Science
Victoria University

Victoria, Australia

2015

Towards Breast Cancer Survivability Prediction Models in Thai Hospital Information Systems

Jaree Thongkam

This thesis is presented in fulfilment of
the requirements for the degree of
Doctor of Philosophy

School of Engineering and Science
Faculty of Health, Engineering and Science
Victoria University

Victoria, Australia

2009

Thesis Supervisor
Professor Yanchun Zhang
Thesis Co-supervisor
Dr. Fuchun Huang

Author
Jaree Thongkam

Towards Breast Cancer Survivability Prediction Models in Thai Hospital Information Systems

Abstract

Finding suitable ways to develop models for predicting unknown data classes is a challenging task in data mining and machine learning. The improvement of the quality of data sets and combining AdaBoost with a weak learner is an important contribution to the development of these prediction models.

The objectives of this thesis are to build accurate, stable and effective breast cancer survivability prediction models using breast cancer data obtained from the Srinagarind Hospital in Thailand. To achieve these objectives, five approaches were proposed including: 1) *k*-means and RELIEF to improve accuracy and stability of prediction models generated from AdaBoost algorithms; 2) C-Support Vector Classification Filtering (C-SVCF) to identify and eliminate outliers; 3) a combination of C-SVCF and oversampling approaches to handle both outliers and imbalanced data problems; 4) a hybrid AdaBoost and Random Forests to build stronger prediction models; and 5) C4.5 to form breast cancer survivability decision trees and rules. To illustrate capability, performance and effectiveness of these approaches, extensive experimental studies have been conducted using WEKA version 3.5.6, AdaBoost MATLAB Toolbox, LIBSVM and C4.5 program.

Empirical results in this study show that *k*-means and RELIEF algorithms improve the accuracy and stability of prediction models built from AdaBoost algorithms. Although the application of these algorithms is unable to achieve a significant improvement in accuracy, this study is useful for investigating the limitation of AdaBoost. On the other hand, C-SVCF is not only effective in identifying and eliminating outliers yet also can improve the accuracy and Area Under the receiver operating characteristic Curve (AUC) of the prediction models up to 24.12% and 29.69%, respectively. Moreover, a combination of C-SVCF and over-sampling approaches, which is able to handle an imbalanced problem of data, provides the improvement of accuracy, sensitivity, specificity, AUC score and *F*-measure of the models up to 29.83%, 29.83%, 47.34%, 38.59% and 33.38%, respectively. Furthermore, a hybrid AdaBoost and Random Forests provides an accuracy of prediction models up to 97.55% which is better than AdaBoost and Random Forests. In addition, C4.5 and C4.5rules are used to provide decision trees and decision rules which are easily understood by health practitioners.

This thesis has systematically investigated the survivability analysis of breast cancer via data mining and provided suitable approaches for developing accurate and reliable breast cancer survivability prediction models. Furthermore, this thesis also provides accurate decision trees and rules for assisting medical practitioners in their decision-making processes for breast cancer patients in Thailand.

Declaration

I, Jaree Thongkam, declare that the PhD thesis entitled "Towards Breast Cancer Survivability Prediction Models in Thai Hospital Information Systems" is no more than 100,000 words in length including quotes and exclusive of tables, figures, appendices, bibliography, references and footnotes. This thesis contains no material that has been submitted previously, in whole or in part, for the award of any other academic degree or diploma. Except where otherwise indicated, this thesis is my own work.

Signature

26/12/09
Date

Table of Contents

Abstract i

Declaration iii

Table of Contents iv

List of Figures xi

List of Tables xiii

List of Algorithms xvi

Preface xvii

Acknowledgements xviii

Dedication xix

Chapter 1 Introduction 1

1.1 Background 1

1.2 Research issues investigated in this thesis 3

1.3 Contributions of the thesis 6

1.4 Outline of the thesis 7

Chapter 2 Literature Review 9

2.1 Data mining 9

2.1.1 Data mining processes 11

2.1.2 Data mining tasks 13

2.2 Classification 13

2.2.1 Classification problems 15

2.2.2 Data selection 16
2.2.2.1 10-fold cross-validation 16
2.2.2.2 Stratified 10-fold cross-validation 17
2.2.3 Evaluation methods 18
2.2.3.1 Accuracy, sensitivity and specificity 19
2.2.3.2 Receiver operating characteristic curve 20
2.2.3.3 Area under the receiver operating characteristic curve 21
2.2.3.4 *F*-measure 22
2.2.3.5 Kappa statistics 23
2.3 Basic classification techniques 24
2.3.1 C4.5 25
2.3.2 Classification and regression tree 31
2.3.3 Naïve Bayesian 34
2.3.4 *K*-nearest neighbour 35
2.3.5 Support vector machine 36
2.3.6 Rule-based classifier 38
2.4 Chapter summary 39
Chapter 3 Breast Cancer Survivability 41
3.1 Breast cancer nature and treatments 41
3.2 Breast cancer research 43
3.3 Survival analysis 44
3.4 Traditional survival analysis tools 46
3.5 Data mining in breast cancer 47
3.6 Problems of breast cancer data in data mining 50

3.6.1 Missing data 50
3.6.2 Outliers 51
3.6.3 Imbalanced data 52
3.7 Data understanding and preparation 54
3.7.1 Breast cancer attributes 55
3.7.2 Examining breast cancer databases 56
3.7.3 5-year breast cancer survival analysis 57
3.8 Chapter summary 60
Chapter 4 Data Pre-processing via AdaBoost 61
4.1 Motivation 62
4.2 Algorithms and research framework 63
4.2.1 *K*-means algorithm 63
4.2.2 RELIEF attribute selection 64
4.2.3 AdaBoost algorithms 65
4.2.4 Pre-processing research framework 68
4.3 Methodology 69
4.3.1 Data preparation 70
4.3.2 Applying pre-processing 71
4.4 Approach validation 72
4.5 Experimental evaluations 73
4.5.1 Accuracy comparisons 73
4.5.2 Sensitivity comparisons 74
4.5.3 Specificity comparisons 75
4.5.4 Comparisons of well-known classifiers 76

4.5.5 Discussion 77
4.6 Chapter summary 79
Chapter 5 Identifying and Eliminating Outliers via C-Support Vector Classification Filtering 80
5.1 Problems of outliers 81
5.2 Outlier filtering approaches and framework 83
5.2.1 C-support vector classification 83
5.2.2 C-support vector classification filtering approach 85
5.2.3 Outlier filtering research framework 86
5.3 Data sets 87
5.4 Validations 88
5.5 Performance and effectiveness of classifiers 89
5.5.1 Default baseline learners 89
5.5.2 Outliers identification 90
5.5.3 Accuracy of classifiers 91
5.5.4 AUC of classifiers 93
5.5.5 Discussion of classifier results 96
5.6 Chapter summary 97
Chapter 6 Improving Data Space via Combining Outlier Filtering and Over-Sampling 98
6.1 Overview and approaches 99
6.1.1 Outlier filtering 100
6.1.2 Over-sampling 101
6.1.3 Combined approaches 102

6.2 Breast cancer survivability data sets 104
6.3 Evaluation approaches 106
6.4 Experimental results 106
6.4.1 Default algorithm setting 107
6.4.2 Imbalance data 107
6.4.3 Accuracy, sensitivity and specificity results 108
6.4.4 AUC results 111
6.4.5 *F*-measure results 113
6.4.6 Discussion of experimental results 115
6.5 Chapter summary 116
Chapter 7 Breast Cancer Survivability Prediction Models 118
7.1 Overview and motivation 119
7.2 Related and a hybrid method 120
7.2.1 Basic AdaBoost 120
7.2.2 Random Forests 122
7.2.3 Hybrid AdaBoost and Random Forests 124
7.3 Part-I: Prediction models of 5-year breast cancer survivability on the outlier-filtered data set 125
7.3.1 Data of 5-year breast cancer survivability 125
7.3.2 Methods for evaluating classifiers 126
7.3.3 Results of classifiers 127
7.3.3.1 Default parameters 127
7.3.3.2 Performance of classifiers 128
7.3.3.3 Performance of AdaBoost with weak learners 129

7.3.3.4 Statistical analysis of multiple classifiers 131
7.3.3.5 Discussion of a 5-year breast cancer survivability prediction model 133
7.4 Part-II: Breast cancer survivability prediction models on outliers-filtered and balanced data sets 134
7.4.1 Data sets of 3-, 5-, 8- and 10-year breast cancer survivability 134
7.4.2 Performance evaluation methods 136
7.4.3 Experimental evaluation results 136
7.4.3.1 Parameters setting 137
7.4.3.2 Accuracy classifications 139
7.4.3.3 Area under the receiver operating characteristic curve classifications 141
7.4.3.4 *F*-measure classifications 142
7.4.3.5 Kappa statistics classifications 143
7.4.3.6 Discussion of classification results 144
7.5 Chapter summary 146
Chapter 8 Breast Cancer Survivability Outcomes 148
8.1 Overview and techniques 148
8.1.1 C4.5 decision tree 150
8.1.2 C4.5rules decision rule 150
8.2 Breast cancer survivability data sets 151
8.3 Breast cancer survivability decision tree models 152
8.3.1 Decision tree for predicting 3-year breast cancer survivability 152
8.3.2 Decision tree for predicting 5-year breast cancer survivability 157
8.3.3 Decision tree for predicting 8-year breast cancer survivability 159

8.3.4 Decision tree for predicting 10-year breast cancer survivability 162
8.4 Breast cancer survivability decision rules 163
8.4.1 Decision rules to predict 3-year breast cancer survivability data 164
8.4.2 Decision rules to predict 5-year breast cancer survivability data 166
8.4.3 Decision rules to predict 8-year breast cancer survivability data 167
8.4.4 Decision rules to predict 10-year breast cancer survivability data 169
8.5 Discussion of decision trees and rules 170
8.6 Chapter summary 171
Chapter 9 Conclusions and Future Work 172
9.1 Summary of results 172
9.2 Limitations of the current study 174
9.3 Future research 175
References 176
Appendices 195

List of Figures

Figure 2.1: Simple classification 15

Figure 2.2: A confusion matrix 18

Figure 2.3: Receiver operating characteristic curve 21

Figure 2.4: The area under the ROC curve 22

Figure 2.5: Final decision tree model 30

Figure 2.6: A binary linearly separable classification problem 36

Figure 3.1: Breast cancer in females: Percentages of a 5-year relative survival and age of patients at the first diagnosis from 1985 to 2004 57

Figure 3.2: Breast cancer in females: Percentages of a 5-year relative survival and stage at diagnosis from 1985-2004 58

Figure 3.3: Relative survival proportions 59

Figure 4.1: Pre-processing framework 69

Figure 4.2: RELIEF scores 71

Figure 4.3: Accuracy before pre-processing 73

Figure 4.4: Accuracy after pre-processing 73

Figure 4.5: Sensitivity before pre-processing 75

Figure 4.6: Sensitivity after pre-processing 75

Figure 4.7: Specificity before pre-processing 76

Figure 4.8: Specificity after pre-processing 76

Figure 5.1: Outlier filtering research framework 87

Figure 6.1: Combined outlier filtering and over-sampling framework..... 103

Figure 7.1: Accuracy comparisons 130

Figure 7.2: Sensitivity comparisons..... 130

Figure 7.3: Specificity comparisons 130

Figure 8.1: Decision tree model for predicting 3-year breast cancer survivability 155

Figure 8.2: Decision tree model for predicting 5-year breast cancer survivability 158

Figure 8.3: Decision tree model for predicting 8-year breast cancer survivability 160

Figure 8.4: Decision tree model for predicting 10-year breast cancer survivability 162

List of Tables

Table 2.1: Example of 5-year breast cancer survivability 28

Table 2.2: Kernel functions 37

Table 3.1: Breast cancer attributes 55

Table 3.2: Descriptive statistics 56

Table 4.1: Input attributes before applying pre-processing 70

Table 4.2: Input attributes after applying pre-processing 72

Table 4.3: Accuracy, sensitivity and specificity of classifiers 77

Table 5.1: Input attributes of 5-year breast cancer survivability data 88

Table 5.2: Number of outliers in the data set 90

Table 5.3: Accuracy of classifiers using C-Support Classification Filtering 91

Table 5.4: Accuracy of classifiers using AdaBoost Filtering 91

Table 5.5: Accuracy of classifiers using Bagging Filtering 92

Table 5.6: Accuracy of classifiers using AdaBoost with SVM Filtering 92

Table 5.7: Accuracy of classifiers using Bagging with SVM Filtering 92

Table 5.8: AUC scores of classifiers using C-Support Vector Classification Filtering.. 94

Table 5.9: AUC scores of classifiers using AdaBoost Filtering 94

Table 5.10: AUC scores of classifiers using Bagging Filtering 94

Table 5.11: AUC scores of classifiers using AdaBoost with SVM Filtering 95

Table 5.12: AUC scores of classifiers using Bagging with SVM Filtering 95

Table 6.1: Input attributes 104

Table 6.2: The number of instances in original data sets 105

Table 6.3: The number of instances using outlier filtering , over-sampling and OOS approaches 107

Table 6.4: Accuracy, sensitivity and specificity of AdaBoost 109

Table 6.5: Accuracy, sensitivity and specificity of Bagging 109

Table 6.6: Accuracy, sensitivity and specificity of C4.5 109

Table 6.7: Accuracy, sensitivity and specificity of SVM 110

Table 6.8: AUC scores of AdaBoost 111

Table 6.9: AUC scores of Bagging 112

Table 6.10: AUC scores of C4.5 112

Table 6.11: AUC scores of SVM 112

Table 6.12: *F*-measure of AdaBoost 113

Table 6.13: *F*-measure of Bagging 114

Table 6.14: *F*-measure of C4.5 114

Table 6.15: *F*-measure of SVM 114

Table 7.1: Input attributes of breast cancer data 126

Table 7.2: Performance of single classifier on the training and test sets 129

Table 7.3: Statistics of accuracy of ensemble classifiers on test sets 132

Table 7.4: Statistics of sensitivity of ensemble classifiers on test sets 132

Table 7.5: Statistics of specificity of ensemble classifiers on test sets 132

Table 7.6: The list of attributes 134

Table 7.7: The number of instances in data sets 135

Table 7.8: Accuracy of classifiers 140

Table 7.9: AUC scores of classifiers 141

Table 7.10: *F*-measure of classifiers 142

Table 7.11: Kappa statistics of classifiers 143

Table 8.1: Input attributes 151

Table 8.2: Instances within data sets 151

Table 8.3: Rules for predicting 3-year breast cancer survivability 164

Table 8.4: Rules for predicting 5-year breast cancer survivability 166

Table 8.5: Rules for predicting 8-year breast cancer survivability 168

Table 8.6: Rules for predicting 10-year breast cancer survivability 169

List of Algorithms

Algorithm 2.1: A basic decision tree 26

Algorithm 2.2: Linear support vector machine 37

Algorithm 4.1: *K*-means algorithm 64

Algorithm 4.2: Basic AdaBoost 66

Algorithm 5.1: C-support vector classification 84

Algorithm 5.2: C-support vector classification filter 85

Algorithm 7.1: Gentle AdaBoost 121

Algorithm 7.2: Random Forests 123

Algorithm 7.3: The hybrid AdaBoost and Random Forests 124

Preface

The following list of publications arises from this thesis.

1) J. Thongkam, G. Xu and Y. Zhang, An analysis of data selection methods on classifier accuracy measures, The Journal of KKU Engineering, Khon Kaen University, 35, Pages 1-10, Jan.-Feb., Thailand, 2008.
2) J. Thongkam, G. Xu, Y. Zhang and F Huang, Breast cancer survivability via AdaBoost algorithms, in Proceedings of the Second Australasian Workshop on Health Data and Knowledge Management (HDKM 2008), pp. 1-10, Jan. 22-25, Wollongong, Australia, 2008.
3) J. Thongkam, G. Xu, Y. Zhang and F Huang, Support vector machines for outlier detection in cancer survivability prediction, in Proceedings of the International Workshop on Health Data Management (IWHDM'08), pp. 99-109, April 28, Shenyang, China, 2008.
4) J. Thongkam, G. Xu and Y. Zhang, AdaBoost algorithm with random forests for predicting breast cancer survivability, in Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN2008), pp. 1-8, Jun. 1-6, Hong Kong, 2008.
5) J. Thongkam, G. Xu, Y. Zhang and F Huang, Toward breast cancer survivability prediction models through improving training space, Expert Systems with Applications (accepted), 2009.

Acknowledgements

First and foremost, I would like to present my most sincere appreciation and deepest gratitude to my supervisor, Professor Yanchun Zhang, for his wonderful support. Professor Zhang is one of the most reliable and amazing people I have ever met. His guidance and advice have been the major contributors toward completing this work. He has been very supportive of my research and offered me a continuous stream of ideas.

I would like to thank Dr. Fuchun Huang and Dr. Guandong Xu for their support and feedback on my research studies. I am grateful to Associate Professor Vatinee Sukmak who guided me during the early stages of my PhD. They have been very kind in supplying me with relevant research papers. Special thanks go to Dr. Petre Santry and Mrs. Angela Rojter for proofreading and extensive advice on English as well as Mrs. Elizabeth Smith for serving on my PhD confirmation committee.

I would like to take this opportunity to thank Professor Somporn Potinam and Mahasarakham University for offering me a scholarship over two years, and also Victoria University for waiving the tuition fee for three semesters. I would like to offer thanks to the Information Technology and Cancer Department staff at Srinagarind Hospital for providing data.

Many thanks go to my parents, brothers, sisters, brother-in-law, son and husband. This great family has been a continuous source of whole-hearted encouragement.

Dedicated to my parents, husband
and sons

Chapter 1

Introduction

This thesis aims to develop accurate, stable and effective breast cancer survivability prediction models using data mining processes to predict breast cancer survivability at Srinagarind Hospital in Thailand. Building accurate and reliable prediction models from historical data is expected to provide valuable information for assisting medical practitioners in their decision-making processes, thus enhancing provision of care, treatment monitoring, and quality assurance for patients [1]. Currently, no relevant studies have been conducted using breast cancer data at Srinagarind Hospital in Thailand, making the present study significantly useful to applications, not only in Thailand, but also in the less developed South East Asian region. Furthermore, this study's development of effective new approaches to building accurate and reliable prediction models in data mining, adds to its significance.

1.1 Background

Breast cancer is the second most common cause of cancer death among women in Thailand [2]. It has been increasing in the past several years, with more than 5,000 new cases reported every year. Several research studies have contributed to investigating factors in diseases such as lifestyle changes, dietary patterns, and genetic issues [3]. Also, much research has analysed the causes and outcomes of the disease

which can assist patients in understanding how to make decisions about their quality of life in accordance with their finances [4] [5]. Traditional tools for analysing patients' survival including Kaplan-Meier and the Cox proportional hazard model, are commonly used to estimate the survival rate of a particular patient suffering from a disease over a particular time period [6]. However, these tools are unable to provide an accurate prediction for an individual patient's outcomes, due to the fact that they only use singular variate and linear analysis techniques to predict future trends [7].

Currently, data mining is widely used in medical domains, including early diagnosis of diseases [8] and patients' risk factors [9] [10], and prescription of suitable drugs and treatments [11] [12] [13]. In data mining, classification is one of the most commonly utilised techniques in building a model for predicting the unseen data [14] [15] [16] [17]. Besides, classification techniques have been proven to be more accurate than the traditional tools mentioned above [7] [18]. In relation to techniques of analysing breast cancer survivability, Ohno-Machado [18] found Neural Networks to be better than Cox Proportional Hazards (traditional tool), whereas Delen, Walker and Kadam [4] showed that the accuracy of a decision tree (C5) outperforms Neural Networks and logistic regression using a large 5-year breast cancer survivability data set from SEER databases. Similarly, Bellaachia and Guven [19] utilised SEER databases to build a 5-year breast cancer survivability prediction model, confirming that the accuracy of the decision tree (C4.5) is superior to Neural Networks and Naïve Bayes. Despite these findings, Jonsdottir, Hvannberg, Sigurdsson and Sigurdsson [20] argued that it is difficult to find an algorithm that is accurate and consistent in all data sets especially in a small data size. This is due to problems including selecting suitable attributes to build the model and quality of medical data (*e.g.* outliers and imbalanced data), which lead to reduction in

the performance, effectiveness and stability of prediction models [21] [22] [23]. In order to solve these problems in the present study, four major problems are addressed:

1) How to identify the suitable attributes in order to improve the accuracy and stability of prediction models.
2) How to identify and eliminate outliers in order to improve the performance and effectiveness of classifiers.
3) How to improve the performance and effectiveness of classifiers in imbalanced data problems.
4) How to build accurate and effective breast cancer survivability prediction models.

1.2 Research issues investigated in this thesis

In order to build accurate breast cancer survivability prediction models, five approaches are proposed. The first approach uses new AdaBoost algorithms (Real, Gentle and Modest) to build a 5-year breast cancer survivability prediction model. The performance of these models is investigated to gain a better understanding of the relative importance of their features. This is done through employing a *k*-means algorithm and RELIEF attributes selection to improve the performance and stability of AdaBoost classifiers. Unlike Qahwaji, Al-Omari, Colak and Ipson [24] who only compared the performance of Real, Gentle and Modest classifiers using simple random to divide data sets without reducing bias and variance of the results, the present study utilises random stratified 10-fold cross-validation to divide a data set with reducing bias and variance of the results. This approach is chosen due to the belief that employment of a *k*-means algorithm to transform the numeric attribute into a discrete attribute and RELIEF to

select suitable attributes, can enhance performance and stability of the models. Moreover, using random stratified 10-fold cross-validation is expected to increase the reliability of experiment results.

The second approach proposes C-Support Vector Classification filtering (C-SVCS) to improve quality of breast cancer survivability data sets. This approach can significantly improve the accuracy and Area Under the receiver operating characteristic Curve (AUC) scores of prediction models and is computationally inexpensive [25]. Outlier filtering framework has traditionally been built in three main steps including: 1) generation of a data set without outliers; 2) adding a level of outliers to the outlier free data set; and 3) using learning algorithms to evaluate performance of approaches. However, this thesis employs a different outlier filtering framework to achieve better prediction performance, following three main steps: 1) obtain original data; 2) apply C-SVCF to identify and eliminate outliers; and 3) evaluate the capability of C-SVCF throughout the performance of prediction results using well-known learning algorithms. In the traditional framework, the capability of the filtering approach has been evaluated using only a small number of learning algorithms. For example, Verbaeten and Assche [26] simply utilised precision, tree size and accuracy of a model generated from a decision tree learning algorithm to evaluate the performance of their filtering approach. However, this thesis applies the accuracy and AUC score of models generated from seven well-known learning algorithms (C4.5, Conjunctive Rule, Naïve Bayes, Nearest Neighbour, Random Committee, Random Forests and Radial Basis Function Network) to evaluate the capability and effectiveness of the C-SVCF approach. Even though C-SVC is commonly used to build prediction models [27] [28] [29], in this thesis it is employed to identify outliers, thus providing an appropriate ap-

proach to identifying outliers while significantly improving the performance and effectiveness of the classifiers.

The third approach proposes the combination of Outlier filtering and Over-Sampling (OOS) to resolve problems related to outliers and imbalanced data. A recent study by Padmaja, Dhulipalla, Bapi and Krishna [30] utilised a three step framework to improve data quality: 1) employ *k*-Nearest Neighbours (*k*-NN) to eliminate outliers in a minority class; 2) apply over-sampling to increase the size of this minority class; and 3) exploit under-sampling to reduce the size of the majority class. However, their framework only filtered outliers from the minority class, allowing outliers in the majority class to remain. Therefore, in order to tackle this problem, the present study adopts four main steps to improve data quality: 1) remove outliers from data sets using C-Support Vector Classification filtering approach; 2) divide the data set into majority and minority classes; 3) resize the minority class to the same size as the majority class using an over-sampling approach; and 4) combine majority and minority classes into one data set. This framework is preferred to eliminate outliers in both minority and majority classes. Furthermore, this study employs five evaluation methods including accuracy, sensitivity, specificity, Area Under the receiver operating characteristic Curve (AUC) and *F*-measure of models, to evaluate the capability and effectiveness of outlier filtering approaches, and to increase the significance and reliability of experimental results.

The fourth approach uses AdaBoost which is an attractive ensemble technique in machine learning since it is used to improve the performance of classifiers by combining with a weak learner [31] [32] [33]. Several studies have successfully combined AdaBoost with weak learners to solve classification problems. For example, Li, Wang

and Sung [34] combined AdaBoost with Support Vector Machine (SVM) to demonstrate that this combination has a better generalisation performance than only SVM. On the other hand, Leshem and Ritov [33] combined AdaBoost with Random Forests to build a motor traffic flow prediction model, showing that this combination has low error rates and is better than basic AdaBoost (AdaBoost with Decision Stump). Nevertheless, few research studies have employed this combination in the medical field. Therefore, a combination of AdaBoost and Random Forests is used to build the 3-, 5-, 8- and 10-year breast cancer survivability prediction models in this study. Moreover, several evaluation methods (accuracy, sensitivity, specificity, Area Under the receiver operating characteristic Curve (AUC), *F*-measure and Kappa statistics) are employed to measure the performance and effectiveness of prediction models.

The final approach utilises C4.5 and C4.5rules to build 3-, 5-, 8- and 10-year breast cancer survivability decision trees and rules models. These models are easily understood by medical practitioners and combine with previous practitioner knowledge to enhance decision making systems.

1.3 Contributions of the thesis

The main contributions of this thesis are to develop new medical data mining approaches to build accurate and reliable breast cancer survivability prediction models as follows:

1) Utilise AdaBoost algorithms to construct breast cancer survivability prediction models, followed by *k*-means and RELIEF to improve the accuracy and stability;
2) Apply C-Support Vector Classification Filtering to identify and eliminate outliers from breast cancer survivability data sets to improve data quality;

3) Combine C-Support Vector Classification Filtering and Over-sampling approaches to further improve data quality in both outliers and imbalanced data problems;
4) Integrate AdaBoost with Random Forests to build accurate and reliable breast cancer survivability prediction models, which opens up an avenue to extending the medical outcomes applications; and
5) Employ C4.5 and C4.5rules to build 3-, 5-, 8- and 10-year breast cancer survivability prediction decision trees and rules.

1.4 Outline of the thesis

In presenting the development of accurate and effective breast cancer survivability prediction models, this thesis comprises nine chapters. *Chapter 1* contains general knowledge about the aims and outline of this thesis.

Chapter 2 discusses data mining and knowledge discovery. Issues about data mining classification together with the well-known techniques including C4.5, Classification And Regression Tree (CART), Naïve Bayesian (NB), *k*-Nearest Neighbour (*k*-NN), Support Vector Machines (SVM) and rule-based, are discussed.

Chapter 3 presents the background of breast cancer and its treatments. Survival analysis and traditional tools for analysing patients' survivability are described. Furthermore, in order to comprehend data mining tools used in the medical field, an analysis of data mining problems in the medical field is presented.

Chapter 4 proposes a *k*-means clustering and RELIEF data selection algorithm to transform the numerical attribute into groups, and to choose relevant attributes to build pre-

diction models. The capability and effectiveness of the proposed approach are evaluated using accuracy, sensitivity and specificity of prediction models.

Chapter 5 presents a C-Support Vector Classification Filter (C-SVCF) to identify and remove outliers to improve the quality of 5-year breast cancer survivability data. The capability and effectiveness of this approach are measured using the accuracy and Area Under the receiver operating characteristic Curve (AUC) of prediction models.

Chapter 6 proposes a combination of Outlier filtering and Over-Sampling (called OOS) to improve data quality in relation to outliers and imbalanced data. In order to assess the capability and effectiveness of the OOS approach, several measurement methods including accuracy, sensitivity, specificity, Area Under the receiver operating characteristic Curve (AUC) and *F*-measure of models are utilised.

Chapter 7 presents a combination of AdaBoost and Random Forests algorithms to develop breast cancer survivability prediction models. Performance and effectiveness of prediction models are evaluated using accuracy, sensitivity, specificity, AUC, *F*-measure and Kappa statistics.

Chapter 8 utilises C4.5 and C4.5rules to build 3-, 5-, 8- and 10-year breast cancer survivability decision trees and decision rules, as they provide an actual tree and rule which are easily understood by medical practitioners in accessing the knowledge-base for decision-making processes.

Lastly, *Chapter 9* concludes the dissertation and presents future directions for the research into data mining for breast cancer survivability.

Chapter 2

Literature Review

This chapter reviews the background of data mining, discussing its processes and presenting its tasks. As data mining commonly utilises a classification for building prediction models, this classification, supervised and unsupervised, is analysed, and classification problems are discussed in order to understand these problems. To investigate the performance and effectiveness of prediction models, data selections including 10-fold and stratified 10-fold cross-validation and evaluation methods including accuracy, sensitivity, specificity, Receiver Operating Characteristic (ROC) curve, Area Under the receiver operating characteristic Curve (AUC), *F*-measure and Kappa statistics, are also discussed. Finally, basic classification techniques including C4.5, Classification And Regression Tree (CART), Naïve Bayesian (NB), *k*-Nearest Neighbour (*k*-NN), Support Vector Machines (SVM) and Rule-Based are reviewed in order to understand the strength and limitation of these basic techniques.

2.1 Data mining

Data mining refers to processes used to extract useful information and knowledge from large databases [35]. Dale and Bench-Capon [36] pointed out that data mining is a new term for knowledge discovery in a large data set to extract relationships from a mass

of data. Wong, Lam, Leung, Ngan and Cheng [37] strongly argued that data mining can be considered as one step in knowledge discovery for databases (KDD). Similarly, Ramirez, Cook, Peterson and Peterson [38] referred to data mining as a particular step in the process of knowledge discovery, while KDD is referred to as the overall process. In fact, the two terms are sometimes drawn upon interchangeably.

Recently, data mining has been widely used, not only for analysing medical data but also in Web, text and images data. In relation to medical mining, many research studies have utilised data mining to analyse medical data. For example, Yi and Fuyong [39] applied a data mining process using C-Support Vector Machine (C-SVM) to analyse breast cancer data sets from the University of Wisconsin Hospitals. Their results indicated that this technique significantly improved the accuracy of the classifiers in unseen test sets, however, in their study this technique was limited to small data sets. Furthermore, Chang [40] employed data mining processes, decision tree and association rules to classify delay levels according to physical illness, language, motor and social emotional developmental delays. He concluded that his results can assist healthcare professionals in understanding the development of young children during the process of evaluation and diagnosis.

In relation to Web mining, several research studies have utilised data mining to analyse on-line user behaviour and user on-line traversal patterns (internet sites) [41] [42] [43]. For example, Madria, Raymond, Bhowmick and Mohania [44] employed Web data mining, *e.g.* association rule, to define relationships between nodes and links of a Web data set. Their results showed that data mining provided a further step in revealing the patterns and linked properties of a document in Web data mining. In addition, Yan, Shen, Peng and Pan [45] used a parallel Web mining algorithm to build a link prediction

model in the environment of Web cluster servers. Their results demonstrated that their model reduced the time complexity and cost of analysis.

On the other hand, text mining uses data mining processes to categorise class labels for new documents based on previous documents [46] [47] [48]. For instance, Zhang and Yang [49] presented that Ridge Regression is better than Support Vector Machine (SVM) in the case of imbalanced data. Conversely, Li and Staunton [48] designated Association Rule-based Classifier By Categories (ARC-BC) in order to provide high accuracy for building associative text classifiers in small data sets.

Similar to text mining, many research studies have made use of data mining to recognise images. For example, Olukunle and Ehikioya [50] suggested that an association rule technique is suitable for extracting hidden information from medical image data. However, Melgani and Bruzzone [27] recommended that a Support Vector Machine (SVM) classifier is superior to both Radial Basis Function Neural Networks and *k*-Nearest Neighbour classifiers when classifying hyper-spectral remote sensing images.

In this study, the term data mining is used and medical mining is of interest because analysing medical data can assist medical practitioners to enhance the provision of care and quality assurance. In order to understand data mining, its processes and tasks are reviewed.

2.1.1 Data mining processes

In relation to business areas, data mining processes consist of six steps [51] including business requirements analysis, data requirements analysis, data mining opportunity identification, data mining project implementation, business application and business results analysis. Likewise, Shearer [52] introduced the Cross-Industry Standard Process for Data Mining (CRISP-DM) with six steps containing business understanding, data

understanding, data preparation, modeling, evaluation and deployment for developing prediction models. However, Han and Kamber [53] presented seven steps including data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge presentation. In short, according to Han and Kamber [53], data mining processes can be summarised into four main steps as follows.

1) Pre-processing includes data integration, data selection, data transformation and data cleaning. *Data integration* combines multiple data sources. *Data selection* identifies and extracts the data from domains. *Data transformation* transforms or consolidates data into an appropriate form for data mining. *Data cleaning* removes outlier/noise in data.
2) Mining involves the process of selecting and/or developing the technique of data mining to extract description and prediction patterns.
3) Evaluating includes the process of measuring performance of the prediction models.
4) Knowledge presentation refers to the step of presenting the knowledge from the mining stage in a suitable form for end-users.

In order to build an accurate and reliable prediction model, pre-processing and mining processes are the most challenging steps in the data mining process [53] [54] [5] [20]. This may be due to the fact that both processes significantly improve the performance and effectiveness of models [54] [5] [20] [55] [30]. Therefore, in this thesis, pre-processing is also discussed, as well as mining processes.

2.1.2 Data mining tasks

Data mining tasks incorporate two primary tasks including prediction and description [53]. *Prediction* involves using known class values in a data set to predict unknown class values. The main prediction tasks in data mining contain classification and regression. C*lassification* is utilised to assign an unlabelled class value to predefine categorical classes. On the other hand, *regression* is exploited to map a data record to a numerical class attribute.

Unlike prediction, *description* embraces describing the data and its behaviour in order to be interpretable to users. The main descriptive tasks in data mining include clustering, summarisation and dependency modelling. *Clustering* is employed to define a finite set of categories or clusters describing data. *Summarisation* is used to illustrate a subset of the data in a compact way. *Dependency modelling* is exploited to express the significant dependencies between data attributes using a particular model.

In this thesis, classification for predicting the unlabelled class attribute is the main concern due to the fact that it provides techniques to develop effective models used in decision-making systems.

2.2 Classification

Classification is a learning process which involves learning knowledge from a labelled class attribute and applying learned knowledge to an unlabelled class attribute [53]. It consists of supervised and unsupervised classifications. *Supervised classification* refers to tasks for building a model from a labelled class attribute in training data [53]. In contrast to supervised classification, *unsupervised classification* refers to tasks for formu-

lating a class attribute from an unlabelled class attribute by grouping high dimension data into a similar class [53] [56] [57].

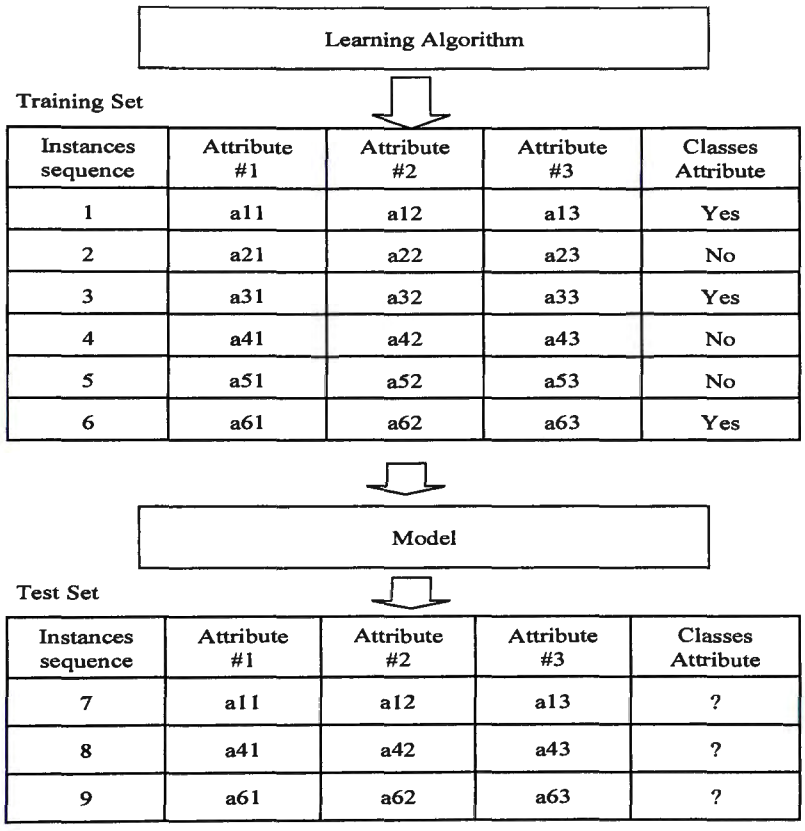
Many research studies have employed supervised classification to form prediction models. For example, Buciu, Kotropoulos and Pitas [58] utilised Support Vector Machine (SVM) for building a prediction model to detect human face data. Their results indicated that SVM was stable in terms of bias and variance. However, Okun and Priisalu [59] employed a Random Forests algorithm to build an ensemble decision tree for cancer classification based on gene expression. They claimed that this decision tree provided an important relevance gene.

On the other hand, several research studies have utilised unsupervised classification to cluster image data to improve the quality of images. An example of this is a study carried out by Shalvi and DeClaris [11] in which a self-organisation map was used for grouping two dimensional images of morphology, and then data mining and data visualisation techniques were applied to illustrate the morphology images. However, Pham [56] introduced Edge-Adaptive Fuzzy *C*-Means (EAFCM) clustering techniques to group two dimensional images of tissue samples. His results showed that EAFCM outperformed fuzzy *C*-means clustering.

This thesis concentrates on supervised classification to develop accurate and reliable prediction models. In order to reduce confusion about the term, the word *classification* will be used to indicate supervised classification. Moreover, to understand the concepts and terms of classification the following are discussed: types of classification, a simple example of the classification problem, data selection for classification procedures and evaluation methods.

2.2.1 Classification problems

Classification problems refer to problems of separating a data set into a smaller class, and determining whether particular data in the data set is in a particular class or not [60]. Many classification problems have been addressed in data mining, machine learning, pattern recognition and statistics [61] [14] [62] [24] [63]. In order to understand classification problems, a simple binary classification problem is given in Figure 2.1 [64].



Instances sequence	Attribute #1	Attribute #2	Attribute #3	Classes Attribute
1	a11	a12	a13	Yes
2	a21	a22	a23	No
3	a31	a32	a33	Yes
4	a41	a42	a43	No
5	a51	a52	a53	No
6	a61	a62	a63	Yes

Instances sequence	Attribute #1	Attribute #2	Attribute #3	Classes Attribute
7	a11	a12	a13	?
8	a41	a42	a43	?
9	a61	a62	a63	?

Figure 2.1: Simple classification

Figure 2.1 shows a simple classification problem. Data in this problem consist of instances and attributes. Instances refer to data in each row, and these are also called examples, vectors and tuples. On the other hand, attributes refer to values in each column, and these are also called features and variables. For the rest of this thesis, the term an *instance* will refer to data in a row and an *attribute* will refer to data in a column. There

are two types of the instance including numerical and categorical. Numerical instances refer to continuous values, while categorical instances refer to a finite set of categories. Attribute also has two types including predictor attributes and dependent attributes (called class attribute). In order to build a model and evaluate the model, a data set is commonly divided into training and test sets using a data selection method. The training set is composed of instances with labelled classes. This training set is used to build a prediction model or classifier. On the other hand, a test set contains instances with unlabelled classes. It is used to measure the performance of the models.

2.2.2 Data selection

Data selection methods (cross-validation, random sampling and bootstrap) are used to divide an original data set into training and test sets. Different methods have their own strategies to partition the original data set and combine the results of multiple partitions. The most commonly used data selection method is cross-validation due to the fact that it provides less bias and variance of classification results than random sampling and bootstrap [53] [4] [65]. Two cross-validation methods, including 10-fold and stratified 10-fold, are discussed.

2.2.2.1 10-fold cross-validation

The 10-fold cross-validation method is widely used in data mining, machine learning and patterns recognition due to the fact that it spends less time dividing the data set into training and test sets [66]. However, the distribution of the class attribute in training and test sets using a data selection method is different from the original class attribute. This leads to instability in predicting results [67]. This method divides a data set (D) into 10 subsets $\{d_1, d_2, \ldots, d_{10}\}$. The first experiment uses partition $\{d_1\}$ for a test set

(T_2), and the remaining partition $\{d_2\text{-}d_{10}\}$ is used for a training set (T_1). The second experiment uses partition $\{d_2\}$ for a test set, and the remaining partitions $\{d_1, d_3, \ldots, d_{10}\}$ are used for a training set. These training and test sets are performed 10 times. Several research studies have investigated the effectiveness of data selection methods. For example, de Lacerda, de Carvalho and Ludermir [66] pointed out that 10-fold cross-validation outperforms booststep and holdout (random sampling) methods using a genetic classifier.

2.2.2.2 Stratified 10-fold cross-validation

Stratified 10-fold cross-validation is widely employed in medical research [68] [65], as this kind of method reduces the bias and variance of classification results in the evaluation process [53] [64]. Moreover, it leads to an increase in the performance of models, which results in greater reliability. There are four main processes of stratified 10-fold cross-validation including [53]:

1) division of the data set into a set of subclasses;
2) assignment of a new sequence number to each set of subclasses;
3) random selection of subclasses into 10 subsets; and
4) combination of each fold of each subclass into a single fold.

As a result, the size of each single fold is approximately equal to the original data set. In this way, many research studies have utilised this method. For instance, Flores and Gonzalez [68] intensively utilised the stratified 10-fold cross-validation to evaluate Neural Networks and decision tree algorithms in mammograms of breast cancer data. Also, Kohavi [67] demonstrated that stratified 10-fold cross-validation performs better than other methods in terms of the bias and variance of estimated accuracies.

Therefore, choosing the most suitable data selection method leads to generating a model which provides a better performance of the classification result [67] [66]. For this reason, stratified 10-fold cross-validation has been chosen to reduce the bias and variance as well as to improve the classification results in the present study.

2.2.3 Evaluation methods

Evaluation methods are used to illustrate and evaluate the performance and effectiveness of models. The way of evaluating the performance of the model includes generalisation and reconstitution errors [53]. *Generalisation errors* refer to evaluation of a model using a test set, while *reconstitution errors* refer to evaluation of a model using a training set. The effectiveness of such a model is not only predicted correctly in an unseen test set, but also provides meaningful and understandable classifier behaviours. In this section, accuracy, sensitivity, specificity, Receiver Operating Characteristic (ROC) curve, Area Under the receiver operating characteristic Curve (AUC), *F*-measure and Kappa statistics, based on the confusion matrix, are discussed.

Confusion matrix refers to the relational table of actual classes and predicted classes used for calculating the performance of classifiers [53]. It is also used to measure the level of effectiveness of the classification model by presenting the number of correct and incorrect classifications of each possible class value being classified [69]. The confusion matrix is shown in Figure 2.2.

		Predicted Classes	
		Positive	Negative
Actual Classes	Positive	*TruePost*	*FalseNeg*
	Negative	*FalsePost*	*TrueNeg*

Figure 2.2: A confusion matrix

Figure 2.2 illustrates a confusion matrix of classes 'Positive' and 'Negative'. The confusion matrix is used to calculate the performance of the classifier in which:

- *TruePost* is the true positive value which is the number of correct predictions in a positive class;
- *FalsePost* is the false positive value which is the number of incorrect predictions in a positive class;
- *TrueNeg* is the true negative value which is the number of correct predictions in a negative class; and
- *FalseNeg* is the false negative value which is the number of incorrect predictions in a negative class.

2.2.3.1 Accuracy, sensitivity and specificity

Accuracy, sensitivity and specificity are basic performance measurements in classification problems [53]. *Accuracy* reflects the possible discrepancies between a predicted class and an actual class. It also refers to the percentage of correctness of positive and negative classes among the test set defined in Equation 2.1. *Sensitivity* refers to the true positive rate in the test set defined in Equation 2.2 while *specificity* refers to the true negative rate in the test set defined in Equation 2.3.

$$accuracy = \frac{TruePost + TrueNeg}{TruePost + FalsePost + TrueNeg + FalseNeg} \tag{2.1}$$

$$sensitivity = \frac{TruePost}{TruePost + FalsePost} \tag{2.2}$$

$$specificity = \frac{TrueNeg}{TrueNeg + FalseNeg} \tag{2.3}$$

When the classifier can predict all cases correctly, the value of a perfect test will be 100 percent of both sensitivity and specificity. Much research has measured the performance of classifiers using accuracy, sensitivity and specificity including Delen, Walker and Kadam [4], Delen and Patil [65], Bellaachia and Guven [19], and Kazmierska and Malicki [70].

2.2.3.2 Receiver operating characteristic curve

Receiver Operating Characteristic (ROC) curve is widely used in evaluating medical images, because it provides the analysis of diagnostic tasks such as disease prevalence and cost-benefit relations in decision-making systems [71] [72]. Moreover, the ROC curve provides more robust evaluation than traditional comparisons such as error rates [73] [72]. As a result, it has often been employed both as an evaluation criterion for the predictive performance of classification in data mining and as an alternative single-number measure for evaluating the performance of learning algorithms [74].

The ROC curve utilises two-dimensional graphs to show the true positive rate (*TPR*) defined in Equation 2.4 and the false positive rate (*FPR*) defined in Equation 2.5 [75] [76] [72]. *TPR* is plotted on the *Y* axis, while *FPR* is plotted on the *X* axis.

$$TPR = \frac{TruePost}{TruePost + FalseNeg} \tag{2.4}$$

$$FPR = \frac{FalsePost}{TrueNeg + FalsePost} \tag{2.5}$$

Two important points in the ROC curve include the lower left point (0: 0) and the upper right point (1: 1) (see Figure 2.3). The lower left point (0: 0) represents the strategy of never issuing a positive classification, as these classifiers commit no false positive rates and gain on true positives rates. The upper right point (1: 1) represents the opposite

strategy of unconditionally issuing positive classifications. The ROC curve is displayed in Figure 2.3.

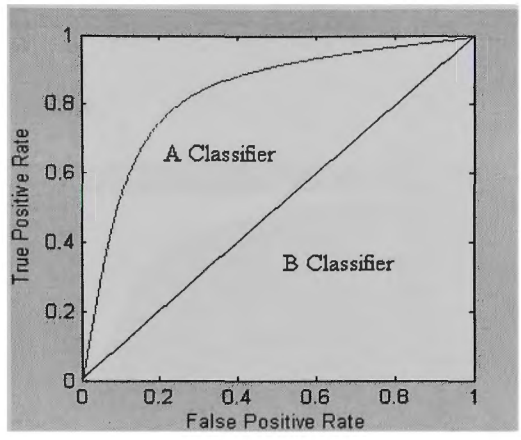


Figure 2.3: Receiver operating characteristic curve

Figure 2.3 shows that classifier A is better than classifier B. Several research studies have utilised the ROC curve to evaluate classifiers. For instance, Biesheuvel, Vergouwe, Steyerberg, Grobbee and Moons [77] employed the ROC curve to interpret the results of two regression models including Polytomous and Dichotomous logistic regression for diagnosis of cancer. Their results showed that the ROC curve provided a simple illustration for demonstrating the performance of the classifiers. Likewise, Liao, Nolte and Collins [78] made use of the ROC curve to present the performance of multisensor decision fusion algorithms that combine the local decisions of existing detection algorithms for different sensors. However, Webb and Ting [79] pointed out that ROC is unsuited to predicting models under varying class distributions, and presents difficulties in distinguishing the performance of classifiers in some cases.

2.2.3.3 Area under the receiver operating characteristic curve

Area Under the receiver operating characteristic Curve (AUC) is commonly used for evaluating medical diagnosis systems [75] [76] [80]. Recently, this method has been

proposed as an alternative measurement criterion for evaluating the predictive ability of learning algorithms by randomly selecting the instance of one class which has a smaller estimated probability among other classes [75] [76] [80]. The AUC of *A* and *B* classifiers is exhibited in Figure 2.4 below.

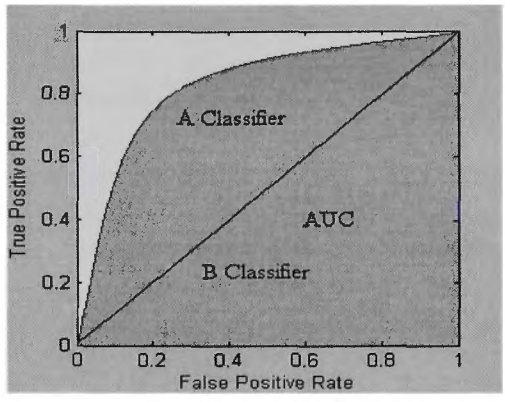


Figure 2.4: The area under the ROC curve

Figure 2.4 shows that the AUC of *A* classifier is larger than the *B* classifier, meaning that *A* classifier is better than *B* classifier. Besides, it can be interpreted into a numeric range which has scores between 0 and 1. As a result, many research studies have utilised AUC scores for comparing classifiers' performance. For example, Jiang [81] successfully employed average AUC scores to analyse the optimal linear in Artificial Neural Networks (ANN) output. Furthermore, Huang and Ling [82] found that AUC is a more accurate measurement method than the ROC curve.

2.2.3.4 *F*-measure

F-measure incorporates the evaluation of effectiveness expressed in terms of hits, misses, false alarms and correct rejections used in text recognition and information retrieval systems [83]. It is computed using both precision (P) defined in Equation 2.6 and recall (R) defined in Equation 2.7. Therefore, *F*-measure is defined in Equation 2.8.

$$P = \frac{TruePost}{TruePost + FalsePost} \tag{2.6}$$

$$R = \frac{TruePost}{TruePost + FalseNeg} \tag{2.7}$$

$$F - measure = \frac{2PR}{P + R} \tag{2.8}$$

In order to achieve a high *F*-measure, both recall and precision have to be balanced. That means achieving a high value for either one of them does not lead to a high *F*-measure value. Several research studies have utilised *F*-measure to measure performance and effectiveness of classification models. By way of illustration, Li and Park [84] employed *F*-measure to measure the categorised effectiveness of Artificial Neural Networks in text categorisation. Their results showed that *F*-measure was superior for evaluating the effectiveness of text models. Likewise, Musicant, Kumar and Ozgur [85] measured the minimisation number of misclassified points using *F*-measure. Their results indicated that *F*-measure was able to present the effectiveness of Support Vector Machine classifiers.

2.2.3.5 Kappa statistics

Kappa statistics are one of the most extensively used methods for nominal scales, intra-class correlation coefficients [86] and measurement of the agreement normalised for chance agreement [87] [88]. It has been used in evaluation functions for assessing the value of individual members of the population [87]. Kappa statistics are used for evaluating a classifier based on both the probability of actual agreement *P*(*A*) defined in Equation 2.9 and the probability of chance agreement *P*(*E*) defined in Equation 2.10. Therefore, Kappa statistics are defined in Equation 2.11.

$$P(A) = \frac{TruePost + TrueNeg}{TrueNeg + FalseNeg} \tag{2.9}$$

$$P(E) = \frac{(TruePost + FalseNeg)(TruePost + FalsePost)}{(TrueNeg + FalseNeg)^2} + \frac{(FalseNeg + TrueNeg)(FalsePost + TrueNeg)}{(TrueNeg + FalseNeg)^2} \tag{2.10}$$

$$Kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{2.11}$$

In this way, Kappa statistics are an alternative method to evaluate classifiers in classification problems. Although Thomsen, Olsen and Nielsen [89] pointed out that the Kappa statistics method is unable to handle the bias between observers, several researchers have successfully exploited Kappa statistics as a criteria measurement for model selection. For instance, Gao, Warren and Warren-Forward [90] utilised Kappa statistics to evaluate the reliability of inter-rater and intra-rater of visual mammography density. Their results showed that a higher Kappa statistics value increased the reliability of visual mammography density.

In this dissertation, the accuracy, sensitivity, specificity, AUC, *F*-measure and Kappa statistics of classifiers are employed to evaluate the performance and effectiveness of classification models generated by classification techniques. In order to understand the strength and weakness of these techniques, basic classification techniques are discussed.

2.3 Basic classification techniques

Classification is one of the most widely used techniques in data mining, knowledge discovery, artificial intelligence, pattern recognition and machine learning. It enables researchers to extract models describing important data classes or to predict future data trends [53] [91] [92] [93] [1] [94]. These techniques have been rapidly developed and

diffused across many disciplines including finance and investment, manufacturing, business and marketing and medical health care [95]. They also allow users to describe and predict information from large and small sets of data. Therefore, in this section, basic classification techniques including C4.5, Classification And Regression Tree (CART), Naïve Bayesian (NB), Neural Networks (NN), *k*-Nearest Neighbour (*k*-NN), Support Vector Machine (SVM) and rule-based classifier are discussed in order to understand basic concepts for developing prediction models (classifiers).

2.3.1 C4.5

C4.5 [96] is a classic decision tree algorithm in machine learning. It is used to build a tree structure for classifying a data set related to a class attribute consisting of nodes and leaves [53] [97] [98]. In C4.5, nodes represent rules which categorise data according to attributes, and leaves represent the condition in each rule. A basic decision tree algorithm is a straightforward algorithm, as shown in Algorithm 2.1 [53].

Algorithm 2.1: A basic decision tree

Input:
D: Data set;
attribute_list: the set of candidate attributes;
Attribute_selection_method: a procedure to determine the splitting criterion;
Output:
N: Decision tree classifier;

```
(1)   Create a node N;
(2)   If instances in D are all of the same class (C) then
(3)     return N as a leaf node labelled with the class C;
(4)   End if
(5)   If attribute_list is empty then
(6)     Return N as a leaf node labelled with the majority class in D: //majority voting
(7)     Apply Attribute_selection_method (D,attribute_list) ;
(8)     Label node N with splitting_criterion;
(9)   End if
(10)  If splitting_attribute is discrete-valued and
        Multiway splits allowed then // not restricted to binary trees then
(11)      attribute_list ← attribute_list - Splitting_attribute;
(12)    For j=1 of splitting_criterion do
            //partition the tuples and grow subtree from each partition
(13)        let Dj be the set of data tuples in D satisfying outcome j; // a partition
(14)        if Dj is empty then
(15)          attach a leaf labelled with the majority class in D to node N;
(16)        Else
(17)          Attach the node and return it by
                  Generate_decision_tree(Dj,attribute_list) to node N;
(18)      End if
(19)    End for
(20)  End if
(21)  Return N.
```

Algorithm 2.1 displays a decision tree algorithm to generate a decision tree model from a data set. *D* refers to a complete data set with its class labels. *Attribute_list* refers to a list of attributes described in the data set and Attribute_selection_method refers to a heuristic procedure used to select the best attribute from a given data set.

C4.5 starts with partitioning the instances into smaller subsets by using Gain Ratio for selecting the best attribute with unequal class labels from a large number of instances. Following this, the recursive strategy applies the top-down greedy algorithm to build a tree. The Gain Ratio is defined in Equation 2.12 as follows:

$$GainRatio(A) = \frac{Gain(A)}{SplitInfor(A)}. \tag{2.12}$$

In Equation 2.12, Gain Ratio depends on *Gain* (called Gain Information) and *SplitInfor* (called Split Information). *Gain Information* is computed using Equations 2.13. It is based on information theory defined in Equations 2.14 and 2.15, respectively.

$$Gain(A) = Infor(D) - Infor_A(D). \tag{2.13}$$

$$Infor(D) = -\sum_{j=1}^{v} p_j \log_2(p_j). \tag{2.14}$$

$$Infor_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} x \sum_{j=1}^{v} p_j \log_2(p_j). \tag{2.15}$$

On the other hand, *Split Information* represents the potential information generated by splitting a data set (*D*) into partitions (*v*) which corresponds to outcomes of attribute (*A*) in a data set. Split Information is computed from Equation 2.16.

$$Splitfor(A) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} x \log_2(\frac{D_j}{D}). \tag{2.16}$$

The attribute with the maximum Gain Ratio is selected as the splitting attribute. However, when the split information is 0, then the Gain Ratio becomes unstable. Therefore, a constraint on the Information Gain is added to avoid this situation. An example of computing Gain Ratio is given in Example 2.1.

Example 2.1: In order to understand the process of building the decision tree using C4.5 technique, the weather data set introduced by Quinlan [96] is customised and presented as a 5-year breast cancer survivability data set. This is due to the fact that this simple data set is easy to understand for building the decision tree [53] [96]. The data set is shown in Table 2.1.

Table 2.1: Example of 5-year breast cancer survivability

Record No.	age	stage	received-chemo	received-surgery	survivability
1	youth	IV	no	no	'Alive'
2	youth	IV	no	yes	'Alive'
3	middle-aged	IV	no	no	'Dead'
4	senior	III	no	no	'Dead'
5	senior	II	yes	no	'Dead'
6	senior	II	yes	yes	'Alive'
7	middle-aged	II	yes	yes	'Dead'
8	youth	III	no	no	'Alive'
9	youth	II	yes	no	'Dead'
10	senior	III	yes	no	'Dead'
11	youth	III	yes	yes	'Dead'
12	middle-aged	III	no	yes	'Dead'
13	middle-aged	IV	yes	no	'Dead'
14	senior	III	no	yes	'Alive'

Table 2.1 shows a simple data set which consists of five attributes and 14 instances. These attributes include 'age', 'stage', 'received-chemo', 'received-surgery' and 'survivability' (class attribute). This class attribute has two distinct values ('Dead' and 'Alive'). The 'Dead' class refers to patients who die within five years after the first diagnosis while the 'Alive' class refers to patients who are still alive for five years or more after the first diagnosis. In order to build the decision tree from the above instances, Gain Ratio (Equations 2.12), Gain Information (Equations 2.13) and Split Information (Equation 2.16) are computed below.

In order to calculate Gain Information, both Information and Information of each attribute are needed. *Information* is computed using Equation 2.14.

$$Infor(D) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2(\frac{5}{14}) = 0.940 \text{ bits.}$$

Then *Information of each attribute* is computed using Equation 2.15.

$$Infor_{age}(D) = \frac{5}{14} * (-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}) + \frac{4}{14} * (-\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4})$$

$$+\frac{5}{14}*(-\frac{3}{5}\log_2\frac{3}{5}-\frac{2}{5}\log_2\frac{2}{5})$$

$$= 0.694 \text{ bits.}$$

$$Infor_{stage}(D) = \frac{4}{14}*(-\frac{1}{4}\log_2\frac{1}{4}-\frac{3}{4}\log_2\frac{3}{4})$$

$$+\frac{6}{14}*(-\frac{2}{6}\log_2\frac{2}{6}-\frac{4}{6}\log_2\frac{4}{6})$$

$$+\frac{4}{14}*(-\frac{2}{4}\log_2\frac{2}{4}-\frac{2}{4}\log_2\frac{2}{4})$$

$$= 0.911 \text{ bits.}$$

$$Infor_{received\_chemo}(D) = \frac{7}{14}*(-\frac{4}{7}\log_2\frac{4}{7}-\frac{3}{7}\log_2\frac{3}{7})$$

$$+\frac{7}{14}*(-\frac{1}{7}\log_2\frac{1}{7}-\frac{6}{7}\log_2\frac{6}{7})$$

$$= 0.788 \text{ bits.}$$

$$Infor_{received\_surgery}(D) = \frac{8}{14}*(-\frac{2}{8}\log_2\frac{2}{8}-\frac{6}{8}\log_2\frac{6}{8})$$

$$+\frac{6}{14}*(-\frac{3}{6}\log_2\frac{3}{6}-\frac{3}{6}\log_2\frac{3}{6})$$

$$= 0.892 \text{ bits.}$$

Gain Information of each attribute is generated below.

$Gain(age) = Infor(D) - Infor_{age}(D)$ = 0.94-0.694 = 0.246 bits.

$Gain(stage) = Infor(D) - Infor_{stage}(D)$ = 0.94-0.911 = 0.029 bits.

$Gain(chemo) = Infor(D) - Infor_{chemo}(D)$ = 0.94-0.788 = 0.151 bits

$Gain(surgery) = Infor(D) - Infor_{surgery}(D)$ = 0.94-0.892 = 0.048 bits

Splitting Information of each attributes is computed below.

$$SplitInfo_{age}(D) = -\frac{5}{14}*\log_2(\frac{5}{14})-\frac{4}{14}*\log_2(\frac{4}{14})-\frac{5}{14}*\log_2(\frac{5}{14}) = 1.577 \text{ bits}$$

$$SplitInfo_{stage}(D) = -\frac{4}{14}*\log_2(\frac{4}{14})-\frac{6}{14}*\log_2(\frac{6}{14})-\frac{4}{14}*\log_2(\frac{4}{14}) = 1.557 \text{ bits}$$

$$SplitInfo_{chemo}(D) = -\frac{7}{14}*\log_2(\frac{7}{14})-\frac{7}{14}*\log_2(\frac{7}{14}) = 1.000 \text{ bits}$$

$$SplitInfo_{surgery}(D) = -\frac{8}{14} * \log_2(\frac{8}{14}) - \frac{6}{14} * \log_2(\frac{6}{14}) = 0.985 \text{ bits}$$

Lastly, Gain Ratio is computed to select the best attribute to become a root of a decision tree.

$$GainRatio(age) = \frac{0.246}{1.577} = 0.156$$

$$GainRatio(stage) = \frac{0.029}{1.557} = 0.019$$

$$GainRatio(chemo) = \frac{0.151}{1.000} = 0.151$$

$$GainRatio(surgery) = \frac{0.048}{0.985} = 0.048$$

It appears that of 'age' attribute is a root of the decision tree due to the fact that it has the highest Gain Ratio among other attributes. The 'age' values ('youth', 'middle_age' and 'senior') become the branches which are growing in each attribute value. The final decision tree of this example is shown in Figure 2.5.

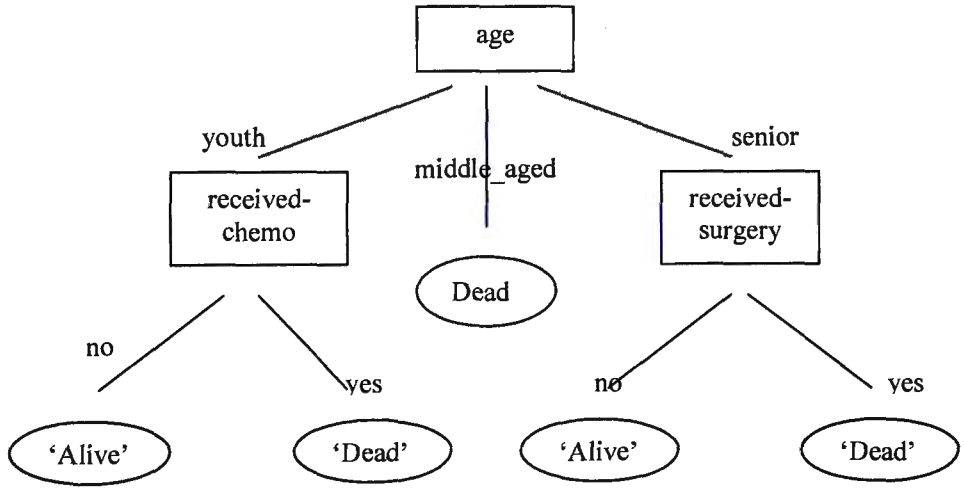


Figure 2.5: Final decision tree model

Figure 2.5 displays an example of a final 5-year breast cancer survivability decision tree model. It indicates that if a patient has breast cancer at a senior level and receives surgery then this patient is predicted to die within five years after the first diagnosis.

In this way, the C4.5 decision tree model is easy to interpret from a tree structure [98] [99] and provides a short computation time for building a model [98] [99] [100]. However, it is limited in robustness and overfitting and is time consuming in memory usage [99] [100]. As a result, much research has provided several approaches to enhance the performance of the C4.5 classifier. For example, Ruggieri [97] adopted a binary search of the threshold in a training set to improve the generating time used in the main memory. His results indicated that the computation time in the main memory increased up to five times, while producing the same decision tree. On the other hand, Yao, Liu, Lei and Yin [99] utilised attribution selection to improve the effectiveness and robustness of a C4.5 classifier.

2.3.2 Classification and regression tree

Classification And Regression Tree (CART) [101] is a widely used decision tree technique. It uses a rule-based approach to generate a binary tree through a binary recursive partitioning process that splits nodes based on the 'yes' and 'no' answer of the predictors. These rules, generated at each step, maximise the class purity within the two resulting subsets ('yes' and 'no'). Each subset is then split further, based on the independent rules. CART uses a Gini Index purity criterion of a single attribute to split a node, based on a rule. The Gini Index purity criterion of classification and regression tree is shown in Equation 2.17.

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2. \tag{2.17}$$

In this equation, D refers to a data set; and p_i refers to the probability that an instance in D belongs to class C_I. The value of p_i is estimated by $|C_{i,D}|/|D|$ and the sum is computed over m classes. Therefore, Gini Index provides a binary split for each attribute.

When considering a binary split, a weighted sum of the impurity of each resulting partition is defined as in Equation 2.18 below.

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2). \quad (2.18)$$

In this way, each of the possible binary splits of each attribute is considered. For a discrete attribute, the subset that gives the minimum Gini Index for that attribute is selected as its splitting subset. However, some attributes may be used many times while others may not be used at all. In order to understand the computing of the Gini Index, an example of the Gini Index is given in Example 2.2 below.

Example 2.2: This example uses data from Table 2.1 to build a decision tree using the Gini Index as follows. Firstly, Equation 2.17 is used to compute the Gini Index impurity of a data set (D).

$$Gini(D) = 1 - (\frac{9}{14})^2 - (\frac{5}{14})^2 = 0.459.$$

Secondly, the Gini Index of each attribute is computed as show below (*e.g.* the conditions of 'age' attribute are 'youth' and 'middle_age' ($Gini_{age \in \{youth, middle\_age\}}(D)$).

$$\begin{aligned} Gini_{age \in \{youth, middle\_age\}}(D) &= \frac{9}{14} Gini(D_1) + \frac{5}{14} Gini(D_2) \\ &= \frac{9}{14}\left(1 - (\frac{6}{9})^2 - (\frac{3}{9})^2\right) + \frac{5}{14}\left(1 - (\frac{3}{5})^2 - (\frac{2}{5})^2\right) \\ &= 0.457 \end{aligned}$$

$$\begin{aligned} Gini_{age \in \{youth, senior\}}(D) &= \frac{9}{14} Gini(D_1) + \frac{5}{14} Gini(D_2) \\ &= \frac{10}{14}\left(1 - (\frac{5}{10})^2 - (\frac{5}{10})^2\right) + \frac{4}{14}\left(1 - (\frac{4}{4})^2 - (\frac{0}{4})^2\right) \\ &= 0.357. \end{aligned}$$

$$\begin{aligned} Gini_{age \in \{middle\_age, senior\}}(D) &= \frac{9}{14} Gini(D_1) + \frac{5}{14} Gini(D_2) \\ &= \frac{9}{14}\left(1 - (\frac{7}{9})^2 - (\frac{2}{9})^2\right) + \frac{5}{4}\left(1 - (\frac{2}{5})^2 - (\frac{3}{5})^2\right) \\ &= 0.394 \end{aligned}$$

In order to split the 'stage' attribute into a binary split, the Gini Index of 'stage' attribute is computed.

$$Gini_{stage \in \{II,III\}}(D) = \frac{10}{14}Gini(D_1) + \frac{4}{14}Gini(D_2) = 0.443$$

$$Gini_{stage \in \{II,IIII\}}(D) = \frac{8}{14}Gini(D_1) + \frac{6}{14}Gini(D_2) = 0.458$$

$$Gini_{stage \in \{III,IIII\}}(D) = \frac{10}{14}Gini(D_1) + \frac{4}{14}Gini(D_2) = 0.450$$

Therefore, Gini $_{\text{stage }\{\text{'II' and 'III'}\}}$ has been obtained as the best split with a Gini Index of 0.443. Moreover, Gini $_{\text{stage }\{\text{' II,IIII'}\}}$ and Gini $_{\text{stage }\{\text{' III,IIII'}\}}$ have Gini Index values of 0.458 and 0.450, respectively.

Finally, the Gini Index of each attribute is compared and selected in order to find the best binary split using the minimum Gini Index. As a result, the 'age' attribute and subset of $Gini_{age \in \{youth,senior\}}$ gives the minimum Gini Index overall, with a reduction in impurity to 0.459-0.357 = 0.102. As a result, 'age' attribute is selected as the root node as C4.5.

In real world data, models generated from CART often encounter an overfitting problem [53] [102]. Moreover, CART often generates classifiers with many nodes from a few predictors that make the trees extremely complex and difficult to interpret [53] [102] [103]. Several research studies have developed an alternative CART algorithm to address these problems. For example, Chipman, George and McCulloch [104] proposed a Bayesian CART using stochastic search methods to avoid greediness and instability problems. On the other hand, Tibshirani and Knight [105] used a bootstrap-based method for searching through a data set to improve the performance of classifiers.

2.3.3 Naïve Bayesian

Naïve Bayesian (NB) [106] is commonly used to build a model in machine learning and data mining. Because it is both time and space efficient when it builds the frequency table, it does not need to store the training data set in memory [53] [70]. However, it has an overfitting problem [70] [107]. The Naïve Bayesian classifier, Naïve Bayesian is defined in Equation 2.19.

$$P(X \mid C_i) = P(x_1 \mid C_i)xP(x_2 \mid C_i)x...xP(x_n \mid C_i)$$
$$P(X \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i) \tag{2.19}$$

where, P refers to the prior probability of each class, X refers to the vector of data *e.g.* $(x_1,x_2, \ldots,x_n)$, n refers to the number of instances and C_i refers to a class i in a data set. This technique can handle both categorical and numerical attributes. In the case of an attribute value as a category, $P(x_k|C_i)$ is the number of instances in a class C_i having the value x_k for the attribute, divided by $|C_{(i,D)}|$, the number of instances in class C_i. On the other hand, if an attribute value is a number, the mean (μ) and standard deviation (σ) is used to compute $P(x_k \mid C_i)$. The $P(x_k \mid C_i)$ is defined as in Equation 2.20 below.

$$P(x_k \mid C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i}) \tag{2.20}$$

where:

$$g(x_k, \mu_{C_i}, \sigma_{C_i}) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{2.21}$$

To predict the class label of X, $P(X\backslash C_i)P(C_i)$ is the evaluation for each C_i. The classifier predicts that the class label of instance X is the class C_i.

$$P(X\backslash C_i)P(C_i) \text{ for } 1 \le j \le m, j \ne i. \tag{2.22}$$

In this way, several research studies have employed Naïve Bayesian to build optimising models from their data. For instance, Kazmierska and Malicki [70] employed Naïve Bayesian to build a model to assess the individual risk of cancer progression after re-

ceiving radiotherapy treatment for brain tumour patients. Their results showed that this model provided high accuracy (84%), specificity (87%) and sensitivity (80%).

2.3.4 *K*-nearest neighbour

The *k*-Nearest Neighbour (*k*-NN) [108] is a commonly used technique to build a model for classifying data in pattern recognition and machine learning [53] [108] [109]. Although it handles both numeric and categorical attributes, it suffers from noisy data which leads to an overfitting problem and is time consuming in training a model [109]. In order to build a model containing numeric attributes, Euclidean Distance is used to measure the closest point between two points of instances (X_1 and X_2), as repesented in Equation 2.23.

$$dist(X_1, X_2) = \sqrt{\sum_{i-1}^{n} (x_{1i} - x_{2i})^2} \ . \tag{2.23}$$

where, X_1 refers to (x_{11},x_{12},...x_{1n}) and X_2 refers to (x_{21},x_{22},...,x_{2n}). On the other hand, in order to build a model with category attributes, a distance-based method is used to compare them intrinsically and assign equal weight to each attribute.

In this way, many research studies have utilised *k*-NN to classify text documents. For example, Li and Staunton [48] performed the *k*-NN algorithm to classify the effect of multi-texture segmentation. Their results indicated that the number of *k* equal to five was suitable for classifying their multi-texture segmentation. Likewise, Colvo, Larrañaga and Lozano [110] employed *k*-NN to build models using binary unbalanced data. Their results demonstrated that *k*-NN classifier increased the classification performance in independent classes. Hu, Yu and Xie [109] employed *k*-NN for selecting the suitable attributes in their data. Their results showed that *k*-NN was suitable for selecting attributes which leads to the improvement of the accuracy of the classifiers.

2.3.5 Support vector machine

Support Vector Machine (SVM) [111] is a novel classification technique based on Neural Networks technology. It uses a statistical learning theory to classify a binary data set by finding a linear optimal hyper-plane to maximise the margin for separating both positive and negative classes [112] [111]. It provides a flexible and low error rate for classification tasks [64]. Nevertheless, it has problems with high dimensionality and complexity of models that is hard to interpret [113]. In order to clearly understand the linear support vector machine, a binary linearly separable line is exhibited in Figure 2.6.

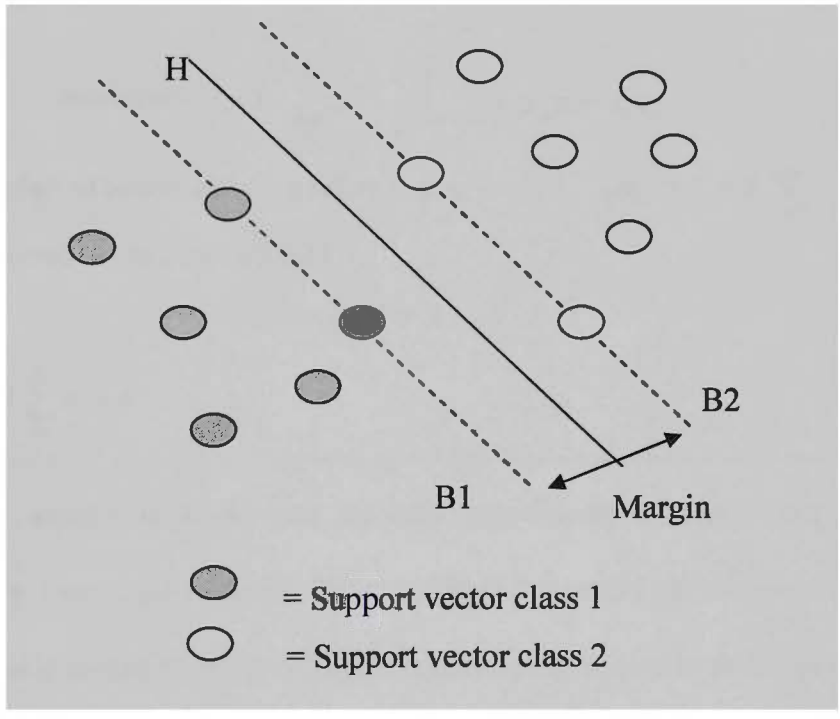


Figure 2.6: A binary linearly separable classification problem

Figure 2.6 displays the separation line for a binary linearly separable classification problem. The boundaries of two classes are separated with boundaries B1 and B2. The data points on a boundary (called support vector) are used to find the maximum margin in the hyper-plane (called H). As a result, the liner SVM algorithm is illustrated in Algorithm 2.2 below.

Algorithm 2.2: Linear support vector machine

Input:

S: a training set (x_i,y_i), i=1,2,…,n; $x_i \in R^n$; $y_i \in \{+1,-1\}$;

n: the number of instances;

Output:

$f(x)$: output space;

(1) Find an optimal separating hyper-plane using Equation 2.24

$$\text{minimise (in } w, b) = \frac{1}{2} \cdot |w^2| \quad \text{with } i = 1,2,..,n \tag{2.24}$$

$$\text{subject to } y_i(w \cdot x_i + b) \geq 1 \tag{2.25}$$

In the Equation 2.24, w refers to weight vector; b refers to a bias for all elements of the training set;

(2) Use characteristics of the language multipliers (α) to solve optimisation problem using the Equation 2.26

$$\text{maximise (in } \alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i-1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i x_j \tag{2.26}$$

where x_i refers to data in class i, x_j refers to data in class j, $(\alpha_i) \geq 0$ and $\sum_i \alpha_i y_i = 0$; and

(3) $f(x)$ is showed in the Equation 2.27

$$f(x) = sign(w \cdot x + b) \tag{2.27}$$

where $w = \sum_{i=1}^{N} \alpha_i y_i x_i$. (2.28)

Unfortunately, most real world data are often non-linear; therefore, four kernel functions including linear, polynomial, radial basis and sigmoid are commonly utilised to solve a quadratic optimisation problem in a data set to optimise the hyper-plane. The kernel functions are shown in Table 2.2.

Table 2.2: Kernel functions

Kernel functions	Mathematical forms
Linear kernel	$K(x_i, x_j) = (x_i \cdot x_j)$
Polynomial kernel of degree d	$K(x_i, x_j) = (\gamma x_i \cdot x_j + r)^d$
Radial basis function	$K(x_i, x_j) = \exp\{-\gamma \|x_i - x_j\|^2\}$
Sigmoid kernel with $r \in N$	$K(x_i, x_j) = \tanh(\gamma x_i \cdot x_j + r)$

Note: γ refers to gamma and r refers to the coefficient of data.

Many research studies have employed SVM to build a prediction model. For instance, Yi and Fuyong [39] applied SVM to discover breast cancer diagnosis patterns from the University of Wisconsin Hospitals. Their results showed that SVM was suitable for diagnosing breast cancer patterns. Similarly, Coussenment and Poel [113] made use of SVM to build churn prediction models from their customer relationship management (CRM) system. Their results revealed that SVM outperforms logistic regression in a balanced data set.

2.3.6 Rule-based classifier

Rule-based classifier refers to a set of 'If-Then' rules for representing information and knowledge in databases [64] [53]. Moreover, it can be made more comprehensible by reducing the number of conditions in classification rules [114]. Although a common rule is usually derived from human experts as linguistic knowledge [115], a rule generated from data mining algorithms can be easy to understand [44] [50]. In addition, it can be used to combine with previous knowledge of the domain in the knowledge-bases, and decision-making as well as guideline systems. Most rules are generated either directly from a data set (called association rule) or decision tree (called decision rule) [53].

Association rule provides a useful measure of the relationship between data by defining the association between an attribute and an instance within a data set [53] [44]. It commonly uses an Apriori algorithm to find frequent item sets. Although it is easy to understand the outcomes, it is time consuming in producing the rules [44] [50] [116]. Much research has employed association rules to discover relationship patterns in medical data and Web mining. For example, Ordonez, Omiecinski, Braal, Santana, Ezquerra, Taboada, Cooke, Krawczynska and Garcia [117] introduced association rules

to predict patient heart disease. Their results demonstrated that some rules can confirm the high risks of heart disease. Similarly, Madria, Raymond, Bhowmick and Mohania [44] made use of association rules to explore the warehouse of Web data. Their results presented that the Apriori algorithm is better than the search content and structure method. Conversely, Olukunle and Ehikioya [50] successfully employed the FP-growth algorithm to enhance the medical images extracting time.

On the other hand, *decision rule* is a popular technique and provides several rules generated from decision tree with high performance. The 'If-Then' rules are extracted from a root to a leaf node and split along a path using 'And' to form the rule antecedent ('If') until class prediction then forms the rule consequent ('Then'). Although it is easy to understand the outcomes and it has high prediction results, it often generates the duplication rules [64] [53]. Several research studies have utilised the decision rule technique to explore the relationship of attributes in a data set. For instance, Bensaid, Bouhouch, Bouhouch, Fellat and Amri [118] used a fuzzy rule-based technique induced from an ID3 model for classifying an electrocardiogram database. Alternatively, Tsakonas, Dounias, Jantzen, Axer, Bjerregaard and von Keyserlingk [119] employed a crisp rule-based technique to induce the rules from a genetic algorithm for classifying medical databases.

2.4 Chapter summary

In this chapter, data mining and Knowledge Discovery in Databases (KDD) have been analysed. Also, data mining processes, and classification and examples of its problems, have been discussed. Subsequently, data selection used in evaluation procedures and measurement methods including accuracy, sensitivity, specificity, Receiver Operating

Characteristic (ROC) curve, Area Under the receiver operating characteristic Curve (AUC), *F*-measure and Kappa statistics were discussed. In order to understand the strangeness and weakness of these well-known learning algorithms, C4.5, Classification And Regression Tree (CART), Naïve Bayesian, *k*-Nearest Neighbour, Support Vector Machines and rule-based classifier have been discussed. In the next chapter, breast cancer survivability will be considered in order to understand data and prepare data sets for building the prediction models.

Chapter 3

Breast Cancer Survivability

In the previous chapter, the background of data mining and its processes were reviewed. In order to investigate and evaluate prediction models, classification learning, evaluation procedures, measurement methods and basic classification techniques were discussed.

As the main purpose of this thesis is to build accurate and reliable breast cancer survivability prediction models using data mining methods, this chapter reviews breast cancer and its treatments to investigate its causes and outcomes. In order to understand breast cancer survivability, breast cancer research studies, survival analysis and traditional tools for analysing the patient's survival are discussed. In order to investigate data mining used in medical fields, the current data mining tools used in the medical field are enumerated. The chapter concludes with an understanding of the breast cancer survivability data at Srinagarind Hospital in Thailand, in order to prepare data sets suited to data mining tools.

3.1 Breast cancer nature and treatments

Cancer refers to abnormal, out of control cell growth in the body [120]. Breast cancer is a major cause of concern not only in the United States of America and Austra-

lia but also in Thailand. In the United States, it is the second highest cause of death among women who are diagnosed at a rate of almost one-in-three [121]. Likewise, the National Breast Cancer Centre (NBCC) reported that breast cancer is the most common invasive cancer diagnosed in Australia [121]. Similarly in Thailand, breast cancer is the second most frequent cause of cancer incidence among women [2] and it has been increasing every year [122].

In clinical practice, the developmental stages of breast cancer start from stage 0, and progress to stage IV [120]. Stage 0 is used to describe a non-invasive breast cancer. Stage I describes invasive breast cancer in which the tumour measures up to two centimetres with no lymph nodes involved. Stage II describes both invasive breast cancer in which the tumour measures from two to five centimetres, and cancer that has spread to the lymph nodes under the arm on the same side as the breast cancer. Stage III is divided into subcategories of IIIA and IIIB. In stage IIIA, the tumour measures more than five centimetres, or there is significant involvement of lymph nodes. In stage IIIB, the tumour has spread to the breast-skin, chest-wall or internal mammary lymph nodes. Stage IV describes an invasive breast cancer where the tumour has spread beyond the breast to under the arm and internal mammary lymph nodes. At this stage, a tumour may have spread to the supraclavicular lymph nodes, lungs, liver, bone or brain. Clearly, in improving the ways that medical practitioners can access information to assist them in diagnosing these stages and prescribing the most suitable treatment, their patients can expect improved outcomes [123].

Treatment of breast cancer provides a typical example of inconsistency in clinical practice because the cancer varies depending on types and stages, as well as overall condition [124]. Recently, medical treatments have included biological therapies, chemo-

therapy, complementary medicine and surgery. However, cancer treatments may vary depending on whether the goals of treatments are to cure, to prevent spreading, or to relieve symptoms. Thus, research in the field of breast cancer detection and treatment helps patients to have an idea of the prognosis of the likely course and outcome of their disease as well as the latest treatments. For instance, Andreetta and Smith [125] reported that treatment using adjuvant aromatase inhibitor in the first stage can reduce cancer progressions. Nevertheless, one or more treatment modalities are usually used to obtain the most effective treatment to increase the patient survival period [1].

3.2 Breast cancer research

In the breast cancer research context, numerous research studies have investigated the risk factors of breast cancer. For example, Parker and Folsom [126] reported that sudden weight loss in the past can increase the risk of breast cancer. Similarly, Wasserman, Flatt, Natarajan, Laughlin, Matusalem, Faerber, Rock, Barrett-Connor and Pierce [127] found that certain types of dietary intake can be a risk factor for women in relation to breast cancer. Breast cancer research is commonly studied in the context of laboratory, observation and clinical trials [120] [123]. *Laboratory* studies prove a hypothesis under controlled conditions to yield detailed results. However, these results are generally only a small part of the sample data used. *Observational* studies inspect the characteristics of total populations to illustrate the factors of breast cancer related to specific outcomes; however, these studies are often unable to present the causes and effects of the outcomes. *Clinical trials* involve medical studies of humans and are able to show a cause and effect in the relationships between attributes and outcomes of breast cancer. These

have been used extensively in the development of drugs and procedures in the treatment of breast cancer.

This thesis focuses on the use of clinical data to develop prediction models and to discover the relationships between available attributes and patients' survivability in the context of breast cancer. These models are designed to assist medical research studies in identifying certain patterns of breast cancer progression in the service of health maintenance, and in providing information about a patient's condition.

3.3 Survival analysis

In the breast cancer context, "survival" is the length of time lived after the initial diagnosis of cancer [121]. Similarly, Delen, Walker and Kadam [4] denoted "survival" as a patient remaining alive for a specified period of time following the diagnosis of cancer. In relation to medical prognoses, survival analysis has generally been divided into either a single point in time or several points in time [18].

Single point in time refers to the specific time period of patient survival during their treatment [18]. This kind of analysis is extensively used in oncology studies, and aims to produce an estimate of the probability of occurrence of the event of interest before a certain time. Although this method is able to present a useful estimate of survival isolated as a single point in time, it is unable to provide predictions of how quickly breast cancer could develop in a given patient nor able to illustrate the temporal patterns of breast cancer development. However, a new period analysis has been introduced as a superior method for analysing medical data in relation to data mining processes. Period analysis deals with building survival analysis models in a medical prognosis field [4]. This period analysis refers to the actual survival period of patients from their diagnosis

until time of death. It is used to monitor survival rates and provide up-to-date information [18] [4]. Using period analysis to predict patients' survival is important in the field of breast cancer because it can determine the most suitable type of therapy, matches patients for clinical trials, and provides more accurate patient information [128].

In contrast to a single point in time, *Several points in time* refers to the probability of disease development during the life of an individual [18]. Outcomes of these types of models can be either discrete (death) or continuous (survival-length of stay in the intensive care unit). This kind of analysis gives a prolonged period of time, allowing a meaningful survival curve to be generated. It also provides better survival curves than using single point, due to the fact that it is usually based on the assumption of outcome dependency to produce the hazards function in order to analyse the survival data.

In analysing patients' survival rate in order to study their long term survival, 10 years or longer has traditionally been used to calculate breast cancer outcomes [6]. However, more recently, a 5-year survival analysis has been introduced for developing models in relation to the improvement of early detection and treatment related to breast cancer [121] [129]. As a result, several recent research studies have built a 5-year breast cancer survivability prediction model using a new single point in time. For instance, Burke, Rosen and Goodman [128] utilised a 5-year survival period to predict patient survivability. Similarly, Delen, Walker and Kadam [4] showed that the 5-year breast cancer survivability analysis provides helpful information for medical practitioners. Likewise, Clinical Best Practice in Australia [121] has changed and now provides a 5-year relative survival report rather than a 10-year survival report. Therefore, in the present study a new single point in time, a 5-year survival period, is mainly used for building breast cancer survivability prediction models.

3.4 Traditional survival analysis tools

The prediction models are valuable tools used to assist in determining a prognosis and in deciding whether to apply an appropriate treatment [130]. The most commonly used tools for building the prediction models include Kaplan-Meier survivor curves, Cox Proportional Hazards and Logistic Regression.

Kaplan-Meier survivor curves are simple, non-parametric models used to summarise survival data for an individual at time period (t), which is the condition of his or her survival at a previous time period (t-1) [18]. These provide actuarial life tables and produce-limit estimator of survivor functions for illustrating survival curves of a group of individuals who share a common particular variable. For instance, Boorjian, Crispen, Lohse, Leibovich and Blute [131] employed Kaplan-Meier to estimate the survival rate of their patients who were suffering from synchronous and metachronous renal cell carcinomas. Their results indicated that metachronous tumours had a greater degree of pathological concordance than synchronous lesions. As a result, two different variables can be compared but the number of patients in each resulting group must not fall below a minimum point, beyond which the reliable survival cure cannot be generated [18].

Cox Proportional Hazards are multivariate semi-parametric regression models that allow continuous covariates and involve the assumptions of a simplifying transformation in an initial data and the hazards for the different groups of proportions of the survival periods. Thus, these models are multiple-point models used to estimate the survival of particular patients by calculating the related conditions of the patient with the baseline hazard. Although D'Ambrosio [132] applied this method to build models using descriptive attributes in Biochemical Failure (BF) after definitive radiotherapy of prostate cancer data, it is unable to provide generalisation errors in an unseen test set [7] [18].

Logistic regression is a generalisation of linear regression used for predicting the binary or several class attributes by using a single point in time [80] [133]. Although it used to build the model to predict odds of occurrence, it is unable to build a model with discrete attributes. For instance, Mertens, Flisher, Satre and Weisner [134] employed logistic regression models to examine substance-abuse related medical condition, integrated medical and chemical dependency, and on-going primary care predicting remission of chemical dependency problems at five years. The data set consisted of 598 chemical dependency patients. Their results illustrated mean, median, minimum and maximum numbers of substance-abuse related medical conditions (SAMCs) per participant, as 1.7, 2.0, 0 and 8, respectively. Similarly, Heidema and Nagelkerke [135] demonstrated that the accuracy of the Logistic Regression models is better than the accuracy of Classification and Regression Tree (CART) models using a validation set from breast cancer patients data.

3.5 Data mining in breast cancer

In relation to analysing breast cancer data, data mining is one of the most promising and challenging tools for model generation [53] [136] [1] [19]. It has been applied in breast cancer research including diagnosis of diseases [137] [138], prediction of the effectiveness of treatments [1], prognostic and predictive factors [139] [140] and especially breast cancer survivability [19] [4]. This is because data mining requires less domain experts to propose a hypothesis, has a high performance in terms of results and is capable of mining large data sets with high dimension attributes [4] [53] [65]. As a result, various techniques have been applied for building reliable and accurate prediction models including Neural Networks, decision tree, rule-based and support vector machine.

Neural Networks is an supervised learning classification which uses Multi-Layer Perceptron (MLP) with back-progration by utilising a set of weights to connect between input and output units [53]. They provide a robust approximator function to solve classification problems [84]. Although they have the ability to build highly complex models for non-linear functions, generating a model is time consuming and the model is hard to be interpreted [81]. A large body of research has employed Neural Networks (NN) algorithms to build predictive and descriptive models. For example, Lundina, Lundina, Burked, Toikkanenb, Pylkkänenc and Joensuua [141] utilised a Neural Networks algorithm to build 5-, 10- and 15-year breast cancer survival prediction models from 951 breast cancer patients. They found that a Neural Networks model outperforms logistic regression models for 5-, 10- and 15-year survival periods using AUC in unseen test sets.

Decision Tree is a tree structure consisting of nodes and leaves. Nodes represent rules which categorise data according to attributes while leaves represent the condition in each rule [53]. This technique provides the most promising results, easy interpretation of the tree structure, and the ability to convert to rule-based classifiers. For this reason, decision tree techniques have been used to build the prediction model in breast cancer data. For example, Delen, Walker and Kadam [4] demonstrated that decision tree model (C5) provides better accuracy than Artificial Neural Networks (ANN) and Logistic Regression using 5-year breast cancer survivability data sets from SEER databases. This data set consists of 16 attributes including race, marital status, primary site code, histology, behaviour, grade, extension of disease, lymph node involvement, radiation, stage of cancer, site specific surgery code, age, tumour size, number of positive nodes, number of nodes and number of primaries.

Rule-based refers to a set of 'If-Then' rules which can be generated from the decision tree or directly from the training set. Although rule-based models sometimes have a lower performance than other techniques, they are easily understood by medical practitioners [53] [8]. Several research studies have employed rule-based algorithms to build prediction rules using breast cancer data sets. For example, Kohli, Krishnamurti and Jedidi [138] utilised a Conjunctive Rule algorithm to build prediction rules using the University of Wisconsin's data set. They found that these rules can be readily used in a clinical setting because these rules are simple and have the same structure as the rules currently used in clinical diagnosis.

Support Vector Machine (SVM) [112] is a novel classification technique. It uses Neural Networks to find a linear optimal hyper-plane to maximise the margin for separating a binary class attribute in classification problems [112] [111]. Although models generated from SVM are difficult to interpret, they are accurate, flexible and significantly resistant to overfitting problems [64] [39]. Much research has employed Support Vector Machine (SVM) to generate the prediction models. For example, Yi and Fuyong [39] exploited C-Support Vector Machines (C-SVM) to discover breast cancer diagnosis patterns at the University of Wisconsin Hospitals. Their results showed that SVM was suitable for diagnosing breast cancer patterns. Likewise, Wang, Wu, Liang and Guo [142] showed that Least Square Support Vector Machine (LS-SVM) based on an Independent Components Analysis (ICA) provides a good diagnosis of breast cancer tumours.

3.6 Problems of breast cancer data in data mining

One of the problems in mining cancer data is the uncertain format of breast cancer data [143]. Most researchers manually retrieve the cancer data from the database into the text file before correcting and transforming them into the data mining format. Unlike cancer data, financial data are commonly based on codes which can retrieved directly into the data mining format. However, when applying data mining, problems in breast cancer data are still occur such as missing data, outliers and imbalanced data frequently occurs [13] [23] [21]. These problems directly affect the performance and effectiveness of the prediction models [23] [21].

3.6.1 Missing data

Missing data refers to unknown and null values in data sets [53] [21]. Although this kind of data decreases the ability of algorithms to learn from observations and accurate prediction, most learning algorithms can handle missing data well [144]. In relation to breast cancer data in the medical field, there are three causes of missing data including occasional effects, medical decisions and progress in laboratory examinations [21]. Firstly, *occasional effects* refer to an unclear hypothesis of the attributes based on data collection. These sources of data are often observed in business databases. Secondly, *medical decisions* refer to physical and laboratory examinations data that medical experts have neglected to record in the system for any reason. Finally, *progress in laboratory examinations* refers to data that are often removed by medial experts in order to gain accurate diagnosis and treatment prior to recording it in the system.

Several techniques in data mining can handle missing data well, such as *k*-Nearest Neighbour, C4.5 and Naive Bayesian. For instance, Liu, Lei and Wu [145] investigated

the performance of classifiers in different levels of missing data. Their results indicated Naïve Bayesian as superior to *k*-NN and C4.5 in handling data with missing values. On the other hand, Mussa and Tshilidzi [144] combined Neural Networks and genetic algorithms to handle missing data in their data sets. Their results showed that the number of missing data affects the accuracy of prediction results.

3.6.2 Outliers

Outliers refer to instances which do not follow the common rules, whereas data mining discards these instances as noise or exceptions [146] [147] [148]. This kind of data commonly results in the poor performance of the learned model in unseen data (called overfitting problems) [26]. In relation to methods for handling outliers, there are three commonly used approaches including, outlier filtering, outlier correction and robust algorithms [149].

Outlier filtering approaches employ the learning algorithm to identify and eliminate outliers from mislabelled instances in the data set [25]. Although models from outlier-free may lead to misinterpretation, these prediction models provide better performance and less model building time [150]. Several research studies have employed outlier filtering approaches for identifying and eliminating outliers from misclassified instances in data sets. For example, Verbaeten and Assche [26] utilised Inductive Logic Programming (ILP) to remove outliers. Their results showed that the accuracy of the decision tree increased rapidly after removing outliers. In contrast, Zhou and Jiang [151] made use of a Neural Networks ensemble to identify and eliminate misclassified instances from data sets. Their results indicated that using Neural Networks ensembles can remove outliers resulting in improving the accuracy of the decision tree (C4.5).

Outlier correction approaches are built upon the assumption that each attribute in the data is correlated with others [150]. These approaches are often used in attributes tracking of video images to correct and retain such images using a simple iterative scheme [152]. Although they are computationally expensive and sometimes introduce undesirable features into a data set during the process of correction, these methods provide the benefit of preserving the maximum information available in the data set to increase performance [25] [150]. Several research studies have employed outlier correction approaches to improve video images. For example, Huynh, Hartley and Heyden [152] used statistics to identify and correct outlier instances in an image sequence. Their results showed that statistics provided less re-projection errors in the image sequence data set. Likewise, Broersen [153] employed a statistic linear interpolation to identify and correct outliers in Turbulence data. His experimental results showed that after correcting outliers, the accuracy of the ARMAsel model increased.

Robust algorithms are used to build a complex control mechanism to overcome overfitting problems and improve the generalisation of a learned model [154]. Several algorithms such as Inductive Learning by Logic Minimisation (ILLM), C4.5 and *k*-Nearest Neighbour (*k*-NN) provide this mechanism. ILLM uses a saturation filter [155], while C4.5 utilises a pruning mechanism [96]. On the other hand, the *k*-NN algorithm exploits the appropriate choice of number (*k*) of nearest neighbours [108].

3.6.3 Imbalanced data

Imbalanced data are the number of instances in which one class in the data set outnumbers instances in the other classes [22] [156]. This is a common problem in not only breast cancer data but also credit card fraud detection and software defect prediction [157] [158]. Usually, the classification algorithm exhibits poor performance while deal-

ing with imbalanced data, biasing the results towards the majority class [30]. Re-sampling approaches including under-sampling and over-sampling in pre-processing are commonly used for re-balancing a data set [156] [159] [160].

Under-sampling is used to decrease the size of the majority class to the same size as the minority class. Although the models from under-sampling may lose some useful information and lead to misinterpretation, they can be used to improve the performance of classification in an imbalanced data set [22] [158]. Many research studies have utilised an under-sampling approach to re-balance imbalanced data. For example, Alejo, Garcia, Sotoca, Mollineda and Sánchez [158] successfully exploited an under-sample approach using the Nearest Neighbour rule to reduce the majority class. Their results demonstrated that this approach was suitable for enhancing the accuracy of the Neural Networks classifier. In contrast, Barandelaa, Sánchez, Garcíaa and Rangel [22] demonstrated that the under-sampling approach was ineffective in improving the performance of classifiers in four data sets including Phoneme, Satimage, Glass and Vehicle from the UCI Databases Repository.

Unlike under-sampling, *over-sampling* is used to increase the size of the minority class to the same size as the majority class in order to improve the distribution differential between classes in an imbalanced data set. Although it is difficult to find the best distribution of minority and majority classes in the training set, it can lead to increasing the performance of classifiers after re-balancing the data set [157] [161]. Much research has utilised an over-sampling approach to re-balance data sets. For example, He, Han and Wang [157] employed this over-sampling to increase the size of a minority class to the same size as the majority class in different level distributions to find the best performance. Their results specified that using 500 over-sampling rates improved the per-

formance of C4.5 by at least 6% or more. Likewise, Pelayo and Dick [161] employed the Synthetic Minority Over-sampling TEchnique (SMOTE) to increase the size of the minority class. Their results showed that SMOTE improved the average accuracy by up to 23% on four benchmark data sets from a NASA project.

3.7 Data understanding and preparation

Understanding and preparation of data is an important and time consuming step towards building the prediction models [93] [53]. In this thesis, breast cancer data were obtained at Srinagarind Hospital which is the only medical school associated hospital in Northeast Thailand. It was established in 1972 as a part of the faculty of medicine at Khon Kaen University. Here, patients have increased by 20,000 annually with the current number of patients totalling 100,000. In particular, Patient Statistics Records showed that in 1996, 1997 and 1998 the total cases of breast cancer in the Northeast were 1,163, 1,210 and 1,218, respectively [2]. Since the medical data used in this study is related to human beings, ethical, legal and social issues play an important role. Thus, approval of official permission according to Victoria University ethics procedures was required. In addition, privacy, security and confidentiality are of great concern in the medical data in data mining. As a result, all data in this study is de-identified following the title 45 Code of Federal Regulations part 46 (called 45 CFR 46) for protection of human rights.

These data involved patient information and the choice of treatment for patients diagnosed with breast cancer between January 1985 and December 2006 in Northeast Thailand. The raw data comprise 4,312 patients, and include fields for each record in the database. In order to understand and prepare the input data to suit the data mining for-

mat, three aspects are given: 1) understanding the meaning of attributes; 2) examining the data in databases; and 3) analysing 5-year patients' survival related to their age and stage at first diagnosis from 1985 to 2004.

3.7.1 Breast cancer attributes

In order to understand the meaning of each attribute in the database, breast cancer attributes referred to a field that is fixed-width within the database are shown in Table 3.1.

Table 3.1: Breast cancer attributes

No.	Attributes	Description
1	Patient ID	Patient's unique identification number
2	Sex	Patient's sex
3	Age	Patient's age at the first diagnosis
4	Marital status	Patient's marital status (single, married, monk, unknown)
5	Occupation	Patient's occupation at diagnosis
6	Race	Patient's race
7	Region	Patient's region of residence
8	Diagnosis date	Patient's diagnosis date
9	Basis of diagnosis	Patient's basis of diagnosis
10	Topography	Topography diagnosis of a patient with breast cancer
11	Morphology	Morphology diagnosis of a patient with breast cancer
12	Extent	Extent of breast cancer at diagnosis
13	Stage	Stage of breast cancer at diagnosis
14	Received surgery	Choice of received surgery (yes/no/unknown)
15	Received radiation	Choice of received radiation (yes/no/unknown)
16	Received chemotherapy	Choice of received chemotherapy (yes/no/unknown)
17	Received hormonal therapy	Choice of received hormonal therapy (yes/no/unknown)
18	Received immunotherapy	Choice of received immunotherapy (yes/no/unknown)
19	Received other therapy	Choice of received other treatments (yes/no/unknown)
20	Received supportive therapy	Choice of received treatments for particular symptoms (yes/no/unknown)
21	Last follow up data	Last date of patient's visit to hospital
22	Survival status	Patient's survival status on 17 January 2007
23	Cause of death	Cause of patient's death (cancer/other/unknown)

Table 3.1 displays 23 attributes including 'Patient ID', 'Sex', 'Age', 'Marital status', 'Occupation', 'Race', 'Region', 'Diagnosis date', 'Basis of diagnosis', 'Topography', 'Morphology', 'Extent', 'Stage', 'Received surgery', 'Received radiation', 'Received chemotherapy', 'Received hormonal therapy', 'Received immunotherapy', 'Received other therapy', 'Received supportive therapy', 'Last follow up data', 'Survival status' and 'Cause of death'.

3.7.2 Examining breast cancer databases

In order to successfully build prediction models, it is necessary to have an overall summarisation of data by assigning a case number for each patient and a unique record number for each specific tumour [53]. The basic statistics including the percentage of missing data, mean distribution and the range of maximum and minimum values are shown in Table 3.2.

Table 3.2: Descriptive statistics

No.	Attributes	Missing (%)	Mean	Min	Max
1	Age diagnosis	0	48.60	8	91
2	Marital Status	0.02	1.90	1	9
3	Occupation	22.33	559.87	0	996
4	Race	0.02	1.02	1	9
5	Region	0.02	1	1	9
6	Basis of diagnosis	0	6.15	1	7
7	Topography	0	508.43	500	509
8	Morphology	0	8397.55	8000	9591
9	Extent	0	5.23	1	9
10	Stage	62.73	6.68	0	9
11	Received surgery	25.41	1.27	1	2
12	Received radiation	0	1.70	1	2
13	Received chemotherapy	0	1.35	1	2
14	Received hormonal therapy	46.54	5.23	1	9
15	Received immunotherapy	46.54	5.25	1	9
16	Received other therapy	0	1.95	1	2
17	Received supportive therapy	0	1.91	1	2
18	Survival period (months)	0	32.49	0	260

Table 3.2 displays the statistical results of 18 attributes which exclude patients' ID, sex, last follow up data, survivability status and cause of death. It indicates that some attributes have more than 30% of missing values and some have only one value. This may be due to the fact that some patients were diagnosed in this hospital but received treatments in other hospitals.

3.7.3 5-year breast cancer survival analysis

In order to understand and prepare the survival behaviours of data, 5-year breast cancer survival rates are analysed. The 5-year survival rate is used to generate the following reports: relative survival according to the patient's age; relative survivals recording to the patient's stage of breast cancer at diagnosis; and relative survival proportion, from 1985 to 2004.

Firstly, *relative survival according to the patient's age* is used to show the comparison between the number of breast cancer survivals and the total number of breast cancer patients of the same age range and sex using a 5-year survival rate. Percentages of the 5-year relative survival related to patient's age of diagnosis are illustrated in Figure 3.1.

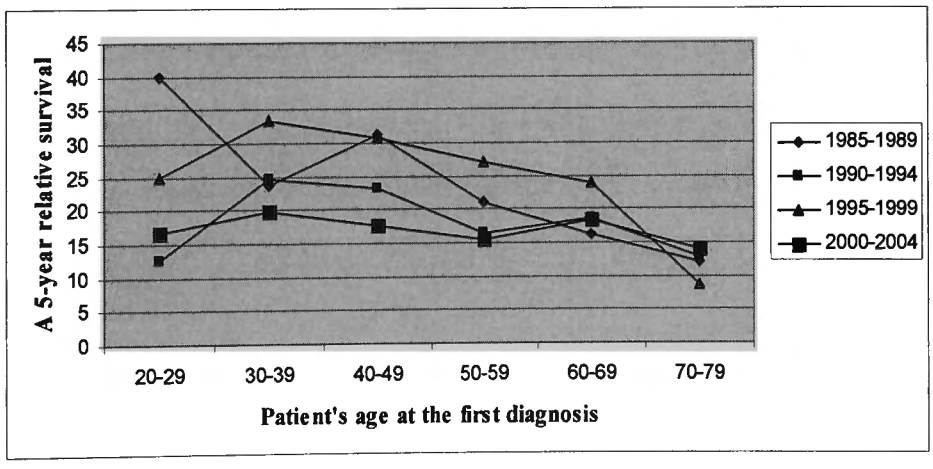


Figure 3.1: Breast cancer in females: Percentages of a 5-year relative survival and age of patients at the first diagnosis from 1985 to 2004

Figure 3.1 displays the percentage of a 5-year breast cancer survival in relation to patients aged between 20 and 79 years, diagnosed from 1985 to 2004. Results indicate that in the years 1985 and 1989, women aged 20-29 years have 40% of 5-year relative survival, then fell to 23.70% for women aged 30-39, 31.33% for women aged 40-49, 21.03% for women aged 50-59, 16.13% for women aged 60-69 and 12% for women aged 70-79, respectively. From 1990-2004, there is a more significant increase in the relative survival of women aged 30-39 years compared to other age groups.

Secondly, *Relative survival among breast cancer stages* is used to present the comparison between survival patients after five years with breast cancer among the number of total patients in the same breast cancer stage and sex at the first diagnosis. The percentage of a 5-year relative survival related to the stage of breast cancer is exhibited in Figure 3.2.

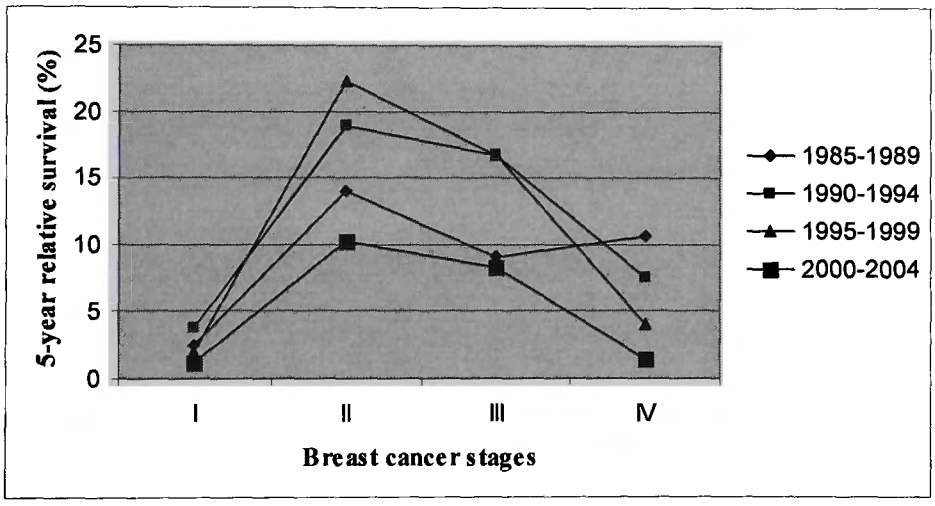


Figure 3.2: Breast cancer in females: Percentages of a 5-year relative survival and stage at diagnosis from 1985-2004

Figure 3.2 presents the percentage of a 5-year breast cancer relative survival related to the stage of breast cancer at the first diagnosis from 1985 to 2004. The results show that the 5-year relative survival for women diagnosed from 1995-1999 is highest at around 22.33% for stage II breast cancer. Furthermore, patients with breast cancer at

stage II have a high probability of surviving for more than five years after the first diagnosis. Patients at stage I have less chance of survival than patients at stages III and IV. This may be due to the fact that these patients are rarely diagnosed at stage I when compared to the total population of that period.

Lastly, *Relative survival proportions* are used to present the percentage of survival period of patients diagnosed with breast cancer each year. Relative survival proportions are displayed in Figure 3.3.

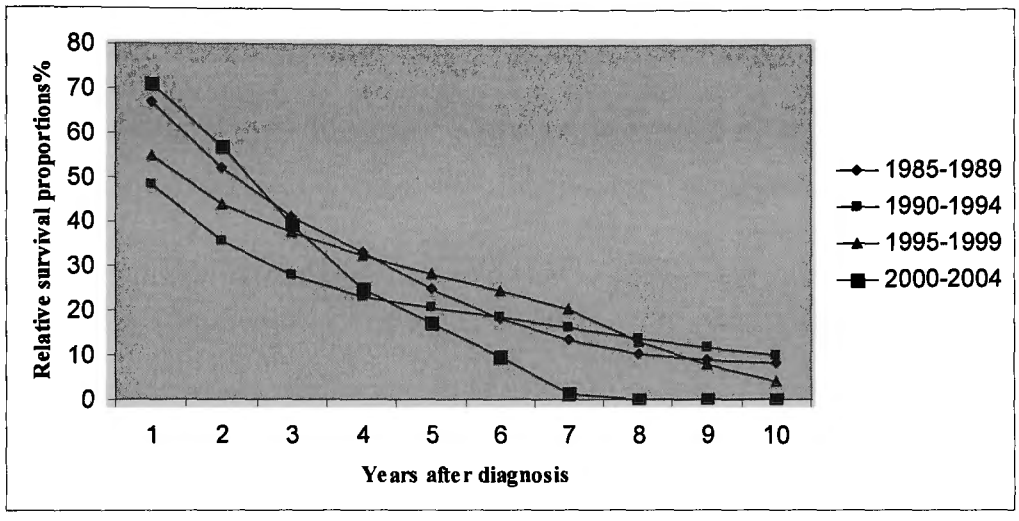


Figure 3.3: Relative survival proportions

Figure 3.3 presents the relative survival proportion after the first diagnosis of breast cancer in females from 1985-2004. The results indicate that 1-year relative survival increased from 66.81% to 71.08% and 5-year relative survival decreased from 24.75% to 17.25%. This may be due to the fact that the diagnosis of breast cancer was already in stages III and IV.

As a result, data from Srinagarind Hospital databases have a number of limitations including missing and imbalanced data. However, outliers are not discussed in this chapter due to the difficulty of finding a suitable method to deal with this problem.

3.8 Chapter summary

In this chapter, breast cancer and its treatments have been reviewed and survival analysis in the field of medical prognosis discussed in order to provide up-to-date information. Following this, breast cancer research, survival analysis, traditional survival analysis tools, data mining tools and the problem of breast cancer data in data mining were discussed. In order to develop accurate and reliable breast cancer survivability prediction models, attributes and missing data were examined and a 5-year breast cancer survival follow-up was analysed. In the next chapter, dimension reduction methods and attribute selection in pre-processing will be investigated in order to improve the performance of the AdaBoost classifiers.

Chapter 4

Data Pre-processing via AdaBoost

In the previous chapter, breast cancer and its treatments were reviewed to better understand the commonly recognised causes and outcomes. Moreover, traditional statistical survival analysis and data mining tools were discussed and the problems of data mining related to the medical fields were presented as well. Subsequently, understandings of the behaviours of these data were also discussed, in order to prepare data sets for building breast cancer survivability prediction models at Srinagarind Hospital in Thailand. In this chapter, the main problem for identifying the suitable attributes that improve the performance and stability of prediction models generated by new AdaBoost is addressed. This work has been published as follows:

- J. Thongkam, G. Xu, Y. Zhang and F Huang, Breast cancer survivability via AdaBoost algorithms, in Proceedings of the Second Australasian Workshop on Health Data and Knowledge Management (HDKM 2008), pp. 1-10, Jan. 22-25, Wollongong, Australia, 2008.

In addressing the above problem, a pre-processing using *k*-means and RELIEF are proposed to improve data quality and select suitable attributes. In order to evaluate the capability and effectiveness of the proposed process, random stratified 10-fold cross-validation, accuracy, sensitivity and specificity are employed. Experimental results are then provided and discussed to evaluate this approach. This chapter concludes with the chapter summary.

4.1 Motivation

AdaBoost (a successor of boosting) is a most popular ensemble method [32]. It is used to combine with a base learner to form a better model (classifier) [31]. Moreover, AdaBoost was reported to be low in error rates in a low outlier data set [32] [31]. Much research has utilised AdaBoost to solve classification problems in object detection including face recognition, video sequences and signal processing systems [162] [163]. For instance, Jinbo, Li and Wenhuang [164] utilised Gentle and Modest AdaBoost algorithms for predicting customer churn data using a top-decile lift criterion to evaluate the prediction models (classifiers). Their results showed that Gentle and Modest AdaBoost prediction models outperformed SVM and C4.5 prediction models. Although AdaBoost algorithms provide prediction models with low error rates, problems in medical data often decrease the performance of prediction models and cause unstable classification results [21]. Therefore, pre-processing in data mining is needed in order to improve the quality of data to generate better prediction models [80] [53]. The most commonly used methods in pre-processing include data transformation and attribute selection.

Data transformation is used to change numerical attributes into discrete attributes, which leads to the reduction of the complexity of data sets and increases the performance of prediction models [165] [166]. Medical practice commonly uses ranking to transform a numerical attribute such as age into groups (*e.g.* 10 years) for analysing data. In relation to data mining, a *k*-means algorithm is commonly used to dynamically partition data into groups, using the nearest distance from the assign point [167] [168]. Several research studies have employed *k*-means to partition their data. For example, Xia, Lyu, Lok and Hyabg [169] employ a *k*-means algorithm to reduce the number of data to reduce the CPU time and memory used in the training process. Their results in-

dicated that the performance of Support Vector Machine (SVM) improved after using the *k*-means algorithm.

On the other hand, *attribute selection* is used to select the relevant attributes to build accurate and stable models [170] [171]. It is also known as variable selection, feature reduction, feature selection, or variable subset selection, however, in this study the term attribute selection is in use. Machine learning has provided several attribute selection methods including RELIEF, Information Gain and Support Vector Machine. However, RELIEF is one of the most successful attribute selection methods, due to its simplicity and effectiveness in forming discrete class attributes [172] [173]. Therefore, in this chapter a *k*-means algorithm and RELIEF attribute selection are proposed to improve the data quality in order to enhance the performance and stability of breast cancer survivability prediction models generated from new Modest AdaBoost algorithms.

4.2 Algorithms and research framework

In this section, the *k*-means algorithm, RELIEF for attribute selection, and AdaBoost algorithms are reviewed. Following this, the pre-processing framework is illustrated.

4.2.1 *K*-means algorithm

K-means is an unsupervised learning algorithm, most well-known in data mining [53] [69]. It is used to separate the nearest distance between instances into the number of cluster (*k*) [73]. Although finding the number of clusters is difficult and related to the performance of a model, it is an effective method of finding the approximate optimal solution [53] [168] [174]. The *k*-means algorithm is presented in Algorithm 4.1.

Algorithm 4.1: *K*-means algorithm

Input :
D: a data set;
K: a number of clusters;
Output:
A: a set of data in each cluster;

(1) Assign the number of clusters (*K*) to initial cluster centers;
(2) **For** $k = 1$ **to** *K*
(3) Assign or re-assign each instance to the cluster to which the instance is the most similar, based on the mean value of the objects in the cluster;
(4) Update the cluster means and calculate the mean value of the instance for each cluster;
(5) **End for**
(6) Return *A*.

Several research studies have utilised *k*-means algorithm in image mining to improve the quality of images. One interesting example is provided by Zhang, Lee and Whangbo [175]. They employed a *k*-means algorithm to cluster the side face attributes and decide the indexes of the side face. Their results showed that this algorithm was able to group the side face data into 5 groups. On the other hand, the *k*-means algorithm was used to identify and eliminate outliers by Tang and Khoshgoftaar [176]. Their results indicated that using *k*-means for identifying and eliminating outliers led to improving the performance of C4.5. As a result, the *k*-means algorithm is a powerful algorithm not only used to reduce instance dimensions, but also to improve the performance of classification models. Therefore, the age attribute is grouped using the *k*-means algorithm to improve the data quality.

4.2.2 RELIEF attribute selection

RELIEF is an instance-based attribute ranking algorithm [173] which utilises the random sampling to locate the nearest neighbour from the same and opposite class. This

method is suited to selecting important attributes to improve the effectiveness of prediction models [172]. Although Sun and Li [177] argued that this algorithm has problems with the original attribute space as it is not in the weighted space and lacks a mechanism to eliminate outliers, it generates a clear score which is easy to understand. RELIEF for attribute selection defined from the weight w_j on input feature j can be computed in Equation 4.1.

$$w_j = P(x_j \neq x_j^d) - P(x_j \neq x_j^s) \tag{4.1}$$

In the Equation 4.1 above, P refers to probability of the nearest instances, x_j refers to the randomly selected training sample, and x_j^d and x_j^s are the two nearest training instances to x_j in the death and alive classes, respectively. Several research studies have notably employed RELIEF to select appropriate attributes in their data. For example, Hall and Holmes [178] demonstrated that C4.5 achieved a higher performance after applying RELIEF for selecting dependent attributes. Therefore, due to the simplicity and effectiveness in machine learning [172], RELIEF was applied as the attribute selection method for selecting relative significant attributes from data sets in this chapter.

4.2.3 AdaBoost algorithms

AdaBoost is the most popular ensemble method in machine learning and it has been shown to significantly enhance the prediction accuracy of the base learner [179] [180]. It is not only used to maintain a distribution or set of weights over the training set, but also for presenting self-rated confidence scores which estimate the reliability of predictions [180]. Although it has disadvantages for classifying noisy data [172] [181] and imbalanced data [182] in binary classification problems, it requires less input parameters, needs little prior knowledge about the base learner, handles the numerical class

well, and has a high flexibility in combining with other methods for finding weak hypotheses [32] [31] [179]. In order to understand the mechanism of AdaBoost, eight steps of the AdaBoost algorithm are reviewed in Figure 4.2.

Algorithm 4.2: Basic AdaBoost

Input:

S: a training set where $S=(x_i, y_i)$ and $(i=1,2,\ldots,n)$;
K: the number of iterations;

Output:

$H_{(x)}$: a set of classifier;

(1) Assign S sample $(x_1,y_1),..,(x_n,y_n)$; $x_i \in X, y_i \in (-1,+1)$;

(2) Initialise the weights of $D_1(i)=1/n$, $i=1,\ldots,n)$;

(3) **For** $k=1$ **to** K **do**

(4) Call WeakLearn, providing it with the distribution D_k;

(5) Get weak hypothesis $h_k: X \rightarrow (-1,+1)$ with its error: $\varepsilon_k = \sum_{i=h_k(x_i)\neq y_i} D_k(i)$;

(6) Update distribution D_k: $D_{k+1}(i) = \frac{D_k(i)\exp(-\alpha_k y_k k_k(x_k))}{z_k}$;

(7) **End for**

(8) Return $H_{(x)} = sign\left(\sum_{k=1}^{K} \alpha_k h_k(x)\right)$.

In the above algorithm 4.2, a training set (S) consists of $(x_1 ,y_1),\ldots(x_n,y_n)$, where each x_i belongs to some domain or instance space X, and each label y_i is in the label set $Y=(-1,+1)$. Although AdaBoost repeatedly assigns a weak learning algorithm in a series of rounds $k=1,\ldots,K$, the weight on the training example i on round t is denoted as $D_k(i)$. The same weight will set at the starting point ($D_1(i)=1/N$, $i=1,\ldots,N$). Then the weight of the misclassified example is increased to concentrate on the hard examples in the training set. Z_k refers to the normalisation constant (chosen so that D_k+1 will be a distribution). In (6) α_k is used to improve the generalising result, and also solve the overfitting and noise sensitivity problem [183]. Consequently, the final hypothesis $H_{(x)}$ refers to a weighted majority vote of the k weak hypotheses where it is the weight assigned to h_k. Although a Decision Stump algorithm is commonly used as base learner in AdaBoost,

in this chapter Classification and Regression Trees (CART) [32] is utilised to generate a tree structure through recursively splitting until the predictor space is completely partitioned into a set of non-overlapping subspaces effective in capturing the local character and complex interaction. In order to show the performance of AdaBoost prediction models, three types of AdaBoost including Real, Gentle and Modest are employed as follows:

1) *Real AdaBoost* [183] is developed by replacing the α_k with

$$\alpha_k = \frac{1}{2}\ln(\frac{W_{+1}}{W_{-1}}). \tag{4.2}$$

$$W_b = \sum_{i=y_k h_k(x_k)=b} D(i). \tag{4.3}$$

In this case D is minimised and W_b refers to the class probability estimate to construct a real value of $\alpha_k h_k(x)$ and become a basic boosting algorithm.

2) *Gentle AdaBoost* [183] is developed from the Real AdaBoost algorithm by updating α_k as Equation 4.4.

$$\alpha_k = \frac{1}{2}\ln(\frac{W_{+1} - W_{-1}}{W_{-1} + W_{-1}}). \tag{4.4}$$

This change improves the generalising error and also solves the overfitting and noise sensitivity problems. Although several research studies have suggested that this algorithm has a similar performance to the Real AdaBoost, Vezhnevets and Vezhnevets [32] found that Gentle often outperforms Real AdaBoost in terms of stability [31] [183].

3) *Modest AdaBoost* [32] modified the formula of α_k as in Equation 4.5.

$$\alpha_k = W_{+1}(1 - \overline{W_{-1}} - W_{-1}(1 - \overline{W_{-1}}) \tag{4.5}$$

This formula is a further change to the formula for calculating the weight of instances to improve the generalising error. Thus, Vezhnevets and Vezhnevets [32] investigated the performance of this Modest AdaBoost in the UCI Machine Learning Repository data sets (including breast cancer, ionosphere, diabetes and hepatitis) using error rates of prediction models. Their results showed that Modest AdaBoost outperformed Gentle AdaBoost in breast cancer, ionosphere and hepatitis.

4.2.4 Pre-processing research framework

Pre-processing is an important and time consuming process [53] [38]. It commonly involves data transformation and attributes selection methods to improve data quality. *Data transformation* is used to transform the low level into the higher level information (*e.g.* numerical values into discrete values). *Attribute selection* is used to reduce dimensions of attributes by removing irrelevant or redundant attributes making the patterns more easily understood. Several research studies have utilised transformation and attribute selection to reduce the number of attributes under consideration and to find invariant representation of the data [38] [184] [185]. In this context, the research framework involved in pre-processing [186] and selecting the best models is illustrated in Figure 4.1.

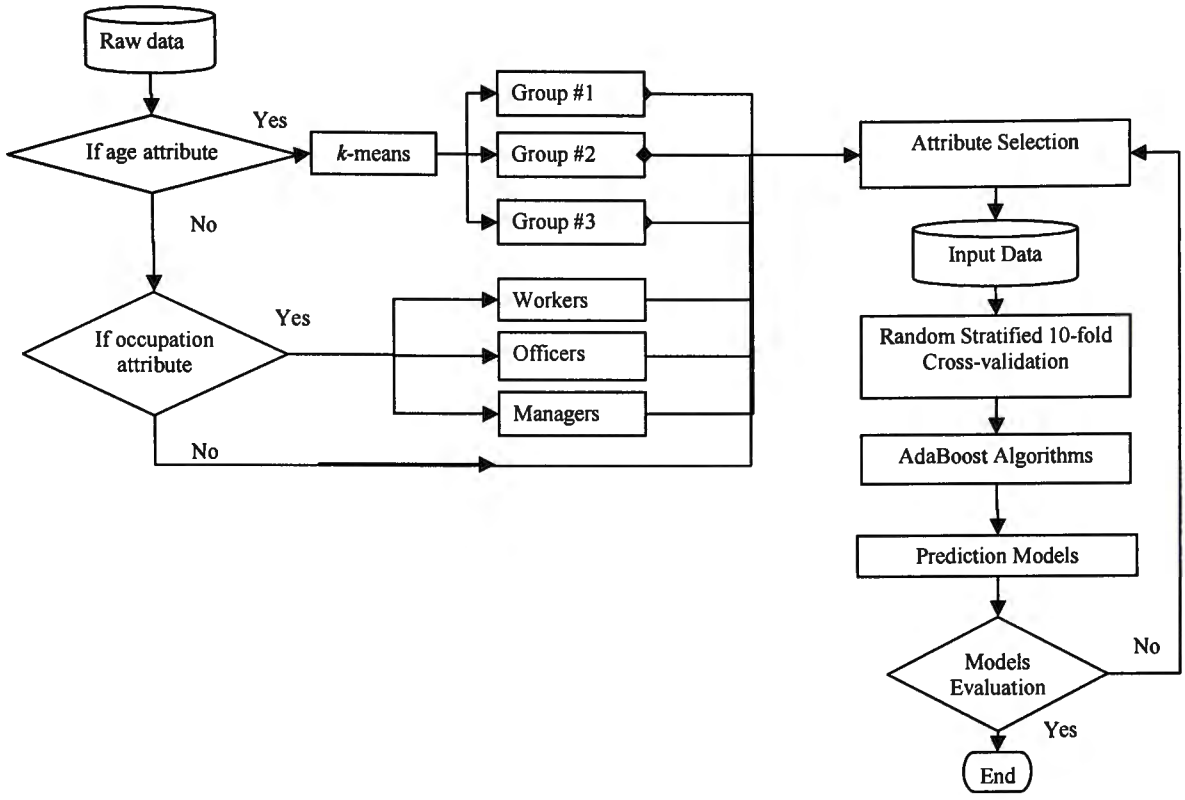


Figure 4.1: Pre-processing framework

Figure 4.1 shows the pre-processing research framework used in this study. It is mainly divided into four steps as follows:

1) Employ *k*-means to transform age attribute into three groups;
2) Transform the occupation attribute into workers, officers and managers;
3) Utilise RELIEF to select the relevant attributes; and
4) Apply AdaBoost algorithms to build the breast cancer prediction models and select the best performance.

4.3 Methodology

In relation to the data mining task, the methodology of the data preparation and steps for applying pre-processing are described.

4.3.1 Data preparation

In this chapter, the data set is generated using data collected from Srinagarind Hospital databases between January 1990 and December 2001. Thus, the initial data comprises 2,462 patients with several attributes. After selecting only identified female patients and removing the incomplete and unknown instances (records), the input data set consists of 736 instances. Twelve attributes including a numeric attribute, 10 discrete attributes, and one class attribute are selected. The numerical attribute consists of an 'Age', while discrete attributes consist of 'Marital status', 'Occupation', 'Basis of diagnosis', 'Topography', 'Morphology', 'Extent', 'Stage', 'Received surgery', 'Received radiation', 'Received chemotherapy' and 'Survivability status'. The list of 12 attributes is presented in Table 4.1 below.

Table 4.1: Input attributes before applying pre-processing

No.	Attributes	Field Name	Attribute Values
1	Age at diagnosis	Age	-
2	Marital status	Married	3
3	Occupation	Occ	27
4	Basis of diagnosis	Basis	6
5	Topography	Top	9
6	Morphology	Mor	14
7	Extent	Extent	4
8	Stage	Stage	4
9	Received surgery	Surg	2
10	Received radiation	Radi	2
11	Received chemotherapy	Chem	2
12	Survivability status (class attribute)	Class	2

Table 4.1 shows the initial attributes used in this chapter. According to binary classification problems related to 5-year breast cancer survivability, the class attribute consists of two classes including 'Dead' and 'Alive'. These classes are generated such that a patient surviving less than 60 months is coded as 0 ('Dead') and 60 months or more is code as 1 ('Alive'). Therefore, the class attribute includes 394 patients ('Dead') and 342 patients ('Alive').

4.3.2 Applying pre-processing

In order to improve the quality of data, the first three steps of the research framework (Figure 4.1) are applied. Firstly, the *k*-means algorithm is applied to group the age attribute into three groups including youth (22-43 years), middle-age (44-56 years) and senior (57-81 years) age ranks. Secondly, based on occupation, the occupation attribute is transformed into three groups including workers, officers and managers. Thirdly, RELIEF is used as the attribute selection method to align the relative attributes with respect to their scores. The results obtained from RELIEF are presented in Figure 4.2.

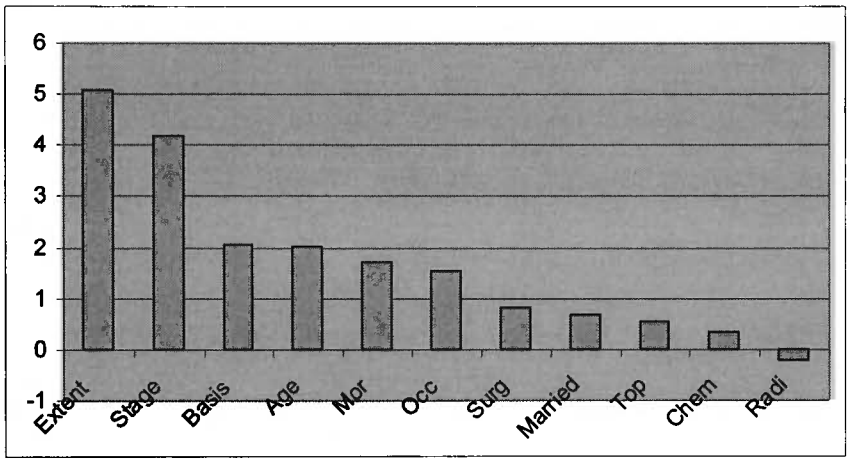


Figure 4.2: RELIEF scores

Figure 4.2 illustrates that extension of breast cancer has the highest score among other attributes in the 5-year breast cancer survivability data set. After running pre-investigation by eliminating the low sore attribute each time, eight attributes including 'Extent', 'Stage', 'Basis', 'Age', 'Morphology', 'Occupation', 'Received surgery' and 'Survivability status' are considered to produce a better performance in building the breast cancer survivability prediction model. The selected attributes are shown in Table 4.2.

Table 4.2: Input attributes after applying pre-processing

No.	Attributes	Attribute Values
1	Extent	4
2	Stage	4
3	Basis	6
4	Age	3
5	Morphology	14
6	Occupation	3
7	Received surgery	2
8	Survivability status (class attribute)	2

Table 4.2 displays new input attributes and the number of values after applying dimension reduction methods, and the attribute selection used for building prediction models. Initiating this process may increase the accuracy and stability of prediction models.

4.4 Approach validation

In order to evaluate the performance and effectiveness of the proposed approach, the accuracy, sensitivity and specificity (see Section 2.2.3.1) of AdaBoost prediction models based on a confusion matrix are employed. *Accuracy* refers to the percentage of correctness of outcomes among the test sets of the prediction results of a classifier. *Sensitivity* refers to the true positive rate of prediction results. In contrast, *specificity* refers to the true negative rate of the prediction results. Moreover, random stratified 10-fold cross-validation is utilised to divide the data set into training and test sets in order to minimise the bias and variance associated with random sampling of training and test sets [67]. Nine folds are used for building the model, and the remaining fold for evaluating the model. Besides, we demonstrate that random stratified 10-fold cross-validation has the highest accuracy using Self-Organizing map and Naïve Bayesian (NB) classifiers in the same data, training and test sets [187]. Therefore, the overall accuracy, sensitivity and specificity are defined using 10 individual experiments as shown in Equation 4.6.

$$CA = \frac{1}{k}\sum_{i=1}^{k} A_i. \tag{4.6}$$

In this Equation 4.6, CA refers to the average of accuracy, sensitivity and specificity, k refers to the number of folds, and A_i is the accuracy, sensitivity and specificity measure of each fold.

4.5 Experimental evaluations

In order to evaluate the fact that applying pre-processing leads to increasing the performance and stability of prediction models, experiments are conducted using MATLAB 7 release 14, GML AdaBoost MATLAB Toolbox [188] and WEKA version 3.5.6. Moreover, parameters of Real, Gentle and Modest algorithms are set to three splitting leaves and levels. Finally, the results of the experiments are presented and discussed.

4.5.1 Accuracy comparisons

In order to evaluate the performance and stability of AdaBoost prediction models before and after applying pre-processing, the accuracy of prediction models is utilised. The results of this experiment are shown in Figures 4.3 and 4.4.

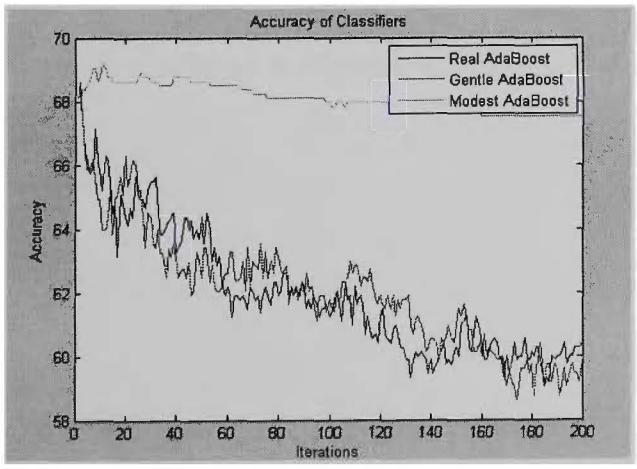


Figure 4.3: Accuracy before pre-processing

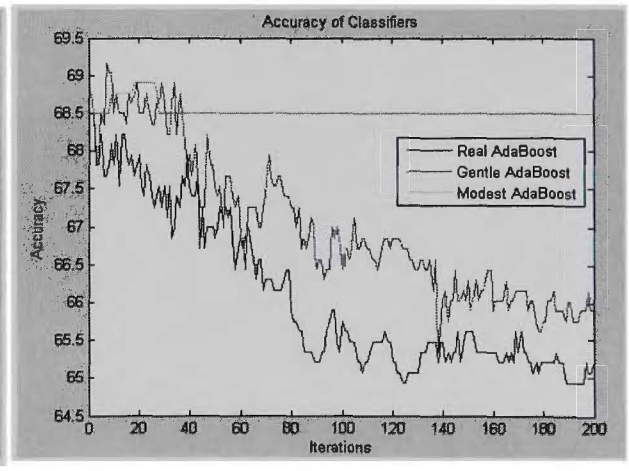


Figure 4.4: Accuracy after pre-processing

Figures 4.3 and 4.4 illustrate the accuracy of the prediction models generated from Real, Gentle and Modest AdaBoost algorithms using 200 iterations before and after the application of pre-processing, respectively. The results of the experiment show that the accuracy of Real and Gentle classifiers decreases rapidly, while the accuracy of Modest seems to increase from 0 up to 15 iterations, before gradually decreasing prior to pre-processing. This may be due to the fact that Real and Gentle classifiers have an overfitting problem which leads to a decrease in the accuracy of these classifiers. However, the stability of Real and Gentle classifiers slightly improves while Modest remains stable following pre-processing. This proves that the Modest classifier provides better prediction accuracy compared to Real and Gentle classifiers in this context. This may be due to the fact that Modest can handle well in small dimensions resulting in decreasing the generalisation error. Moreover, this also proves that applying pre-processing can increase the accuracy and stability of prediction models.

4.5.2 Sensitivity comparisons

In order to investigate the performance and stability of classifiers in the positive class ('Dead' class), sensitivity is used before and after applying pre-processing. The sensitivity results of Real, Gentle and Modest classifiers after running the tests for 200 iterations are displayed in Figures 4.5 and 4.6.

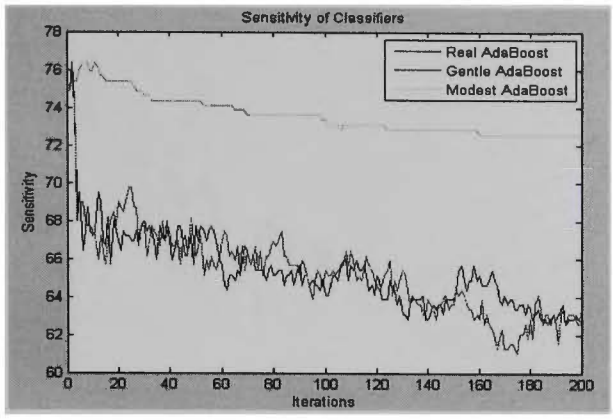


Figure 4.5: Sensitivity before pre-processing

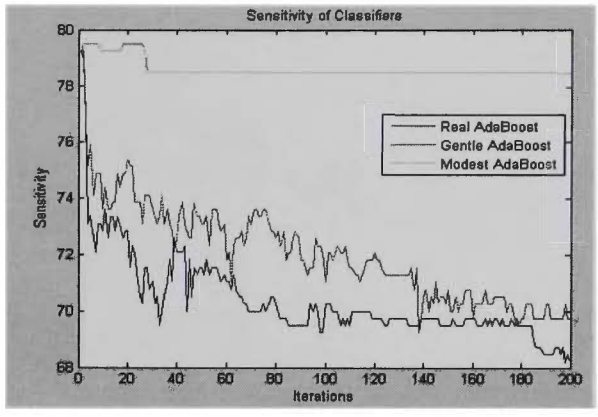


Figure 4.6: Sensitivity after pre-processing

Figures 4.5 and 4.6 show the sensitivity of Real, Gentle and Modest classifiers using 5-year breast cancer survivability data sets. The results of this experiment exhibit that the sensitivity of Real and Gentle classifiers hastily decreases from the 5th iteration until the 200th iteration. On the other hand, the sensitivity of the Modest classifier slightly increases from the start until the 10th iteration and then decreases leisurely until the 200th iteration before the application of pre-processing. However, following pre-processing, the sensitivity of Modest increases significantly. This may be due to the fact that the Modest concentrates on predicting in the majority class rather than the minority class.

4.5.3 Specificity comparisons

In order to investigate the performance and stability of the classifiers in the positive class ('Alive') before and after applying pre-processing, specificity is used. The 200 iterations of the specificity of Real, Gentle and Modest classifiers using the 5-year breast cancer survivability data set are shown in Figures 4.7 and 4.8.

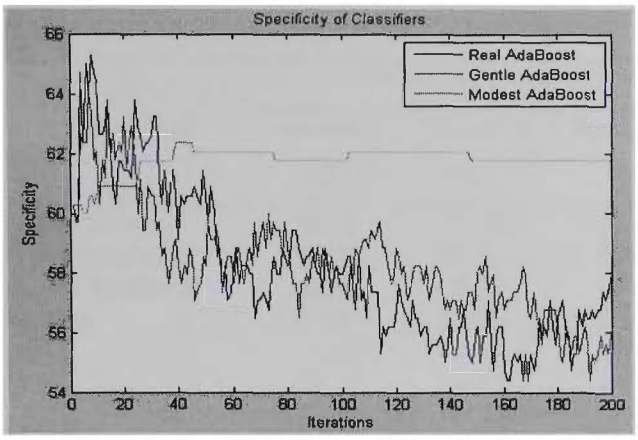


Figure 4.7: Specificity before pre-processing

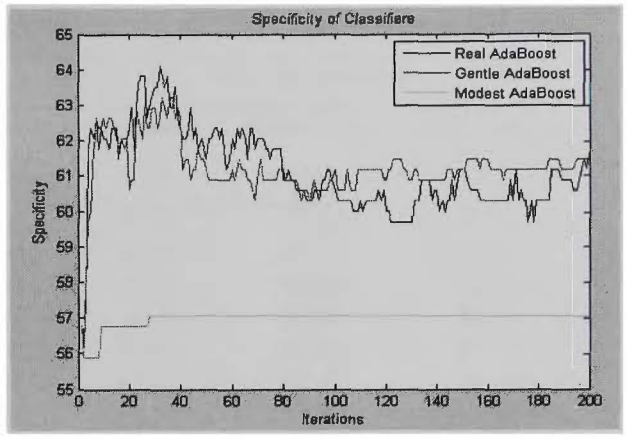


Figure 4.8: Specificity after pre-processing

Figures 4.7 and 4.8 demonstrate the specificity of the Real, Gentle and Modest classifiers. The experimental results show that Real and Gentle classifiers are better than Modest at 10 iterations and then they decrease hastily until 200 iterations. On the other hand, Modest slightly improves from the beginning before pre-processing. However, after applying pre-processing, the specificity of the Real and Gentle classifiers rapidly improves and then slightly decreases while the specificity of Modest slightly improves from the beginning and then becomes stable. This may be due to the fact that the Modest classifier is limited in classifying the instances in the minority class within an imbalanced data set.

4.5.4 Comparisons of well-known classifiers

In order to present the performance of AdaBoost classifiers and compare it to those of well-known classifiers including Bagging, C4.5, Support Vector Machine (SVM) and Random Forests, accuracy, sensitivity and specificity of classifiers are used. In this section, Bagging is based on a fast decision tree learner (REPTree) while SVM, used in this section, is based on a C-Support Vector Classification type with a radial basis kernel. The results of this experiment are shown in Table 4.3.

Table 4.3: Accuracy, sensitivity and specificity of classifiers

Classifiers	Before pre-processing			After pre-processing		
	Accuracy (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)
Real AdaBoost	66.39	70.38	61.80	67.77	73.59	60.09
Gentle AdaBoost	66.74	69.99	63.05	67.56	73.41	60.85
Modest AdaBoost	***68.58***	***75.79***	60.03	***68.63***	***79.95***	55.70
Bagging	67.67	69.08	66.18	67.12	66.95	67.81
C4.5	67.53	67.64	***68.37***	67.67	67.81	67.92
SVM	*61.37*	*62.77*	*60.08*	*68.22*	*67.12*	***70.44***
Random Forests	57.53	59.90	54.79	64.52	67.00	61.78

Table 4.3 illustrates not only the average of accuracy, sensitivity and specificity of Real, Gentle and Modest classifiers using 200 iterations but also the average of accuracy, sensitivity and specificity of Bagging, C4.5, SVM and Random Forests classifiers before and after pre-processing using stratified 10-fold cross-validation. Results show that the Modest classifier provides the highest sensitivity; on the other hand, it provides the lowest specificity after pre-processing. This may be due to the fact that it misleads and doesn't classify the correct classes in noisy data and imbalanced data. However, the accuracy of Modest not only increases after pre-processing but also provides the highest level of accuracy among six classifiers. This demonstrates that the Modest technique is superior to Real, Gentle, Bagging, C4.5, SVM and Random Forests techniques in building a 5-year breast cancer survivability prediction model from 5-year breast cancer survivability data at Srinagarind Hospital in Thailand.

4.5.5 Discussion

In this chapter, a *k*-means algorithm and RELIEF were utilised to improve breast cancer survivability data quality. Subsequently, the Modest AdaBoost algorithm is employed to build a 5-year breast cancer survivability prediction model at Srinagarind Hospital in Thailand. Accuracy, sensitivity and specificity are utilised to measure the performance

and effectiveness of these prediction models. Consequently, several findings are discussed in relation to the performance and stability of Real, Gentle and Modest AdaBooost prediction models.

Firstly, the accuracy of the prediction model generated from Modest is found to provide more accurate and stable results than the prediction model conducted from Real and Gentle. Similarly, results presented by Jinbo, Li and Wenhuang [164] showed that the Modest prediction model is superior to Gentle in customer risks prediction models. However, Qahwaji, Al-Omari, Colak and Ipson [24] demonstrated that models generated from Gentle are more accurate than Modest in weather forecasting data.

Secondly, the accuracy of the AdaBoost models including Real, Gentle and Modest slightly improves after applying the *k*-means algorithm and RELIEF in pre-processing. This may be due to the fact that selecting the relevant attributes can lead to improving the performance of prediction models. In the same way, Borges and Nievola [189] demonstrated that using an attribute selection method in pre-processing leads to improving the accuracy of their prediction models.

Thirdly, the stability of Real, Gentle and Modest prediction models improves after applying pre-processing. This may be due to the fact that reducing dimensions of instances together with selecting the appropriate attributes can lead to improving the stability of AdaBoost models. Likewise, Kalousis, Prados and Hilario [171] applied attribute selection before building their model leading to improving the stability of a linear SVM prediction model.

Finally, we conclude that the proposed pre-processing framework not only improved the accuracy and sensitivity of models generated from Real, Gentle and Modest, but also improved the performance of models generated from C4.5, SVM and Random Forests.

4.6 Chapter summary

In this chapter, we successfully utilised pre-processing by combining *k*-means and RELIEF to improve data quality in order to enhance the accuracy and stability of the 5-year breast cancer survivability prediction model using Modest AdaBoost algorithm. However, there are alternative ways to improve data quality such as outliers handling and re-sampling approaches. In order to find alternative ways to improve the data quality, the next chapter investigates outlier handling approaches due to the fact that outliers are one of the problems which affect the performance of prediction models [147] [153].

Chapter 5

Identifying and Eliminating Outliers via C-Support Vector Classification Filtering

In Chapter 4, a 5-year breast cancer survivability prediction model was successfully developed using pre-processing in data mining and AdaBoost Algorithms. Pre-processing is used to transform a numerical attribute into discrete attributes using the *k*-means algorithm, and selecting suitable attributes using RELIEF in order to enhance the performance and stability of these prediction models.

In this chapter, the problem of identifying and eliminating outliers to improve the quality of data is addressed in order to enhance the performance and effectiveness of classifiers. This work has been published in the following paper:

- J. Thongkam, G. Xu, Y. Zhang and F Huang, Support vector machines for outlier detection in cancer survivability prediction, in Proceedings of the International Workshop on Health Data Management (IWHDM'08), pp. 99-109, April 28, Shenyang, China, 2008.

In addressing this problem, the C-Support Vector Classification Filtering (C-SVCF) approach is proposed to identify and eliminate outliers from a 5-year breast cancer survivability data set. In order to evaluate the capability and effectiveness of the proposed approach, accuracy and Area Under the receiver operating characteristic Curve (AUC)

of prediction models are used. Moreover, this proposed approach is compared, with AdaBoost Filtering (ABF), Bagging Filtering (BF), AdaBoost with Support Vector Machine Filtering (ABSVMF) and Bagging with Support Vector Machine Filtering (BSVMF). Following this, the experimental results are illustrated and discussed in order to provide support for this proposed approach and the chapter summary is presented.

5.1 Problems of outliers

Outliers commonly refer to instances in a current data set that do not comply with general behaviour of a model [53] [146] [147]. For example, in this context patients who have breast cancer in stage I and are less than 30 years of age should be categorised as 'Alive'. However, these patients have been categorised as 'Dead' in the current data set, possibly due to having died of other causes. These outliers decrease the performance and effectiveness of the model [147] [154]. Consequently, several approaches have been introduced in order to detect outliers. These include statistical tests that assume a distribution or probability model for the data and distance measures in which instances that are a substantial distance from any other cluster are considered as outliers. However, in relation to pattern recognition and instance-based learner fields, they utilise a learning algorithm to identify outliers in order to improve the performance of classifiers [53] [25].

There are three main outlier handling approaches including robust, outlier filtering, and outlier correction (see Section 3.6.2). The *robust approach* is used to build a complex control mechanism in order to avoid overfitting in the training data and generalise well in the unseen data [154]. Unlike robust, the *outlier filtering approach* involves a learning algorithm to identify and eliminate misclassified instances (called outliers) from a

data set [25]. On the other hand, the *outlier correction approach* is built upon the assumption that each attribute in the data set is correlated with others, and can be reliably predicted [150]. Unfortunately, outlier correction is usually more computationally expensive than robust and outlier filtering, and is unstable in correcting and cleaning unwanted instances [25]. Therefore, in this chapter the outlier filtering approach is selected to identify and eliminate outliers to improve the quality of the breast cancer survivability data set.

Recently, Support Vector Machine (SVM) has been emerging as a popular classification technique in machine learning [53]. This novel classification technique is based on neural networks technology using a statistical learning theory to classify the class attribute [111]. Moreover, the SVM family provides several types of classification techniques including C-Support Vector, nu-Support Vector and One-class Support Vector Classifications. However, C-Support Vector Classification (C-SVC) is the simplest and most powerful among the other SVM in binary class classification problems [53]. For instance, Yin, Yin, Sun and Wu [190] intensively investigated C-SVC with a radial basis function to identify classification problems in handwritten Chinese characters. Their results showed that C-SVC with radial basis function had the highest accuracy in the predicting task. Although several research studies have used the Support Vector Machine technique for filtering e-mail spam and patterns recognition, few research studies have employed C-SVC to filter outliers in medical data.

Since a patient's information, illnesses and treatments in medical databases are of interest for developing prediction models, they commonly contain outliers which result in an overfitting problem and lead to the prediction model having high error rates [147]. This problem refers to a prediction model that performs the prediction well using the seen

training set, but performs poorly using the unseen test set [26]. This chapter proposed the C-Support Vector Classification Filtering (C-SVCF) approach to identify and eliminate misclassified instances to improve the data quality in order to improve the performance and effectiveness of the 5-year breast cancer survivability prediction model. The capability and effectiveness of the proposed approach is evaluated using the accuracy and AUC score of prediction models generated from several learning algorithms including C4.5, conjunctive rule, Naïve Bayes, Nearest Neighbour classifier (NN-classifier), Random Committee (RC), Random Forests (RF), and a Radial Basis Function Network (RBFNetwork).

5.2 Outlier filtering approaches and framework

In order to improve the quality of data in a data set, an overview of *C*-Support Vector Classification (C-SVC) and a C-Support Vector Classification Filter (C-SVCF) approach are given. Moreover, an outlier filtering framework used in this chapter is introduced.

5.2.1 C-support vector classification

C-Support Vector Classification (C-SVC) [111] is a classification technique which belongs to the Support Vector Machine family, and is a new generation of learning algorithms used in machine learning and pattern recognition [190]. It not only provides a better performance, but is also suited to classifying a binary classification problem. The C-SVC algorithm is reviewed in Algorithm 5.1 below.

Algorithm 5.1: C-support vector classification

Input:

S: a training set (x_i,y_i), i=1,2,...,n; $x_i \in S^n$; $y_i \in \{+1,-1\}$;

Output:

$f(x)$: output space;

(1) Find an optimal separating hyper-plane with the maximum margin with Equation 5.1 below:

$$\text{minimising } \frac{1}{2} \cdot w^2 \tag{5.1}$$

$$\text{Subject to : } y_i ((w \cdot x_i)+b) \text{ -1>= 0}$$

(2) Apply language multipliers (α) in Equation 5.2 as follows:

$$f(x) = sign(\sum_{i=1}^{n} \alpha_i y_i (x_i \cdot x) + b) \tag{5.2}$$

Here, n represents the number of the training samples used to define the decision frontier vectors. In case of the vectors x_i for $(\alpha_i) \neq 0$ which refer to the separation region by the support vectors [191];

(3) Assume that an optimal separating hyper-plane does not exist. This algorithm solves the problem by inserting non-negative slack variables ξ to reduce the optimisation problem which is given in Equation 5.3 following:

$$\text{minimising : } \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{m} \xi_i \tag{5.3}$$

$$\text{Subject to : } y_i \cdot ((W \cdot x_i) + b) \geq 1 - \xi_i, i = 1,...,n$$

Here ξ_i refers to the slack variables, and C refers to a penalty parameter to be chosen by the user; and

(4) Return $f(x)$.

Algorithm 5.1 shows processes of C-Support Vector Classification for finding an optimal separating hyper-plane and returns the classified vectors. Several research studies have employed C-SVC to conduct the prediction models or classifiers. For example, Coussement and Poel [113] employed C-SVC to construct a customer behaviour prediction model. Their results indicated that a model generated from C-SVC with a radial basis kernel function provided greater accuracy and AUC scores than a model generated from Logistic Regression using a marketing data set. On the other hand, Yi and Fuyong [39] performed C-SVC to select suitable attributes in breast cancer diagnosis data sets. Their results showed a significant improvement of accuracy in selected attributes. As

C-SVC is a powerful technique not only used to develop prediction models but also to select the most suitable attributes, this present study uses it to identify outliers in order to improve the performance of prediction models.

5.2.2 C-support vector classification filtering approach

The filtering approach basically refers to an approach that is used to improve data quality by filtering out insignificant outliers from a data set, since classification techniques have a poor result when the data set contains the specific value of outliers. This approach involves a learning algorithm for eliminating outliers from a whole data set, and then the prediction model is generated using the remaining instances to reduce the complexity and increase the performance to the model [25] [26] [192]. Consequently, finding a suitable learning algorithm for removing outliers is a challenging task. In this study, C-Support Vector Classification Filtering (C-SVCF) is proposed to identify outliers in a 5-year breast cancer survivability data set. The C-SVCF algorithm [148] is given in Algorithm 5.2.

Algorithm 5.2: C-support vector classification filter

Input:
D: a training data set;
N: number of instances;
Output:
F: a filtered data set;
O: an Outlier data set;

(1) Empty *F and O;*
(2) Train (*T*) using C-SVC(*D*);
(3) **For** i = 1 **to** N **do**
(4) **If** $D_{(i)} \in T$
(5) Insert $D_{(i)}$ to *F*;
(6) **Else**
(7) Insert $D_{(i)}$ to *O*;
(8) **End if**
(9) **End for**
(10) Return *F,O.*

Algorithm 5.2 shows that the C-SVCF algorithm provides the function to identify and eliminate outliers from training data (D) and return a set of filtered data (F) and a set of outliers (O). In this way, the data set can allow the learning algorithm not only to build an accurate model from significant instances, but also provide a corrected interpretation for domain experts and users.

5.2.3 Outlier filtering research framework

The outlier filtering framework starts generally based on the original data set without outliers [26] [193] [194]. Then the different levels of outliers are added into the original and evaluated through learning algorithms to identify and remove misclassified instances. Similarly, Verbaeten and Assche [26] have shown that the decision tree algorithm is affected by outliers. Their experiment started with an outlier-free data set, followed by adding the different levels of outliers in the data set, and evaluating effectiveness using a decision tree. Their results demonstrated that the accuracy of decision tree decreased rapidly after increasing outliers. In contrast, our framework started from data with outliers, and reduced outliers by applying learning algorithms to identify and randomly eliminate outliers. For example, an algorithm marks an instance as misclassified if it is classified wrongly, whereas this algorithm marks an instance as classified if it is classified correctly. Following this, the misclassified instances are randomly eliminated from the original data set by 5% each time until 20% of elimination is reached, to simulate the ability of the outlier filtering approach. This outlier filtering framework is illustrated in Figure 5.1.

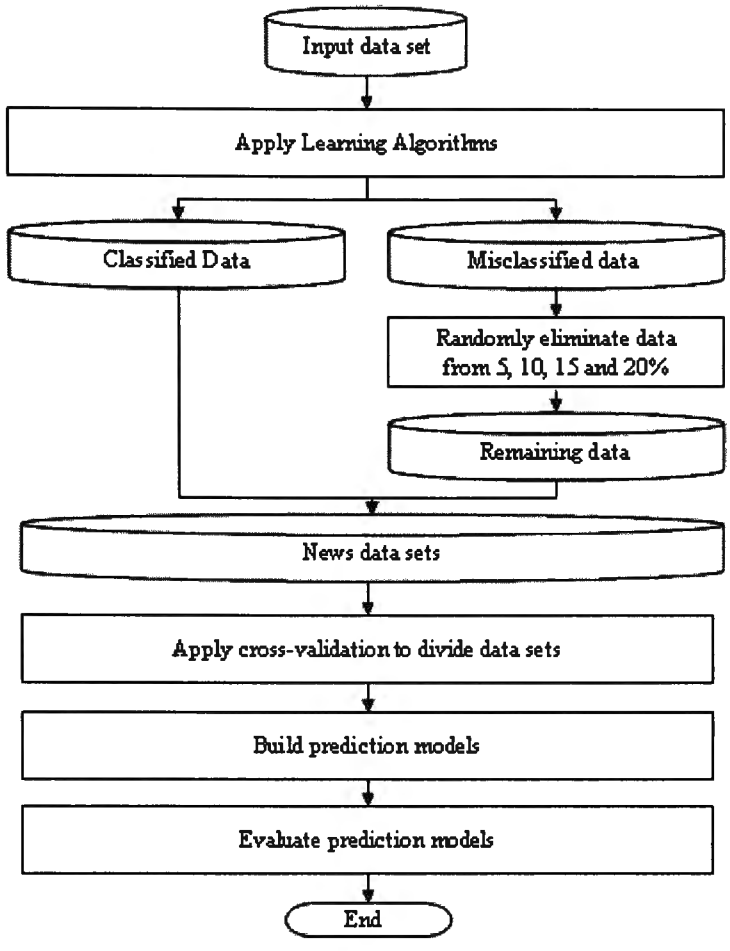


Figure 5.1: Outlier filtering research framework

Figure 5.1 shows the outlier elimination framework used to evaluate the capability and effectiveness of the outlier filtering approaches. This framework is a suitable for finding an appropriate outlier filtering approach which results in increasing the performance of prediction models whilst eliminating the same number of outliers.

5.3 Data sets

In order to conduct the experiment, breast cancer survivability data were obtained at Srinagarind Hospital in Thailand. The data include patient information and the choices of treatments for patients diagnosed with breast cancer from 1990 to 2001. The data consist of 12 attributes which are presented in Table 5.1.

Table 5.1: Input attributes of 5-year breast cancer survivability data

No.	Attributes	Types of attribute	Attribute Values
1	Age	Number	-
2	Marital status	Category	3
3	Occupation	Category	27
4	Basis of diagnosis	Category	6
5	Topography	Category	9
6	Morphology	Category	14
7	Extent	Category	4
8	Stage	Category	4
9	Received surgery	Category	2
10	Received radiation	Category	2
11	Received chemotherapy	Category	2
12	Survivability status (class attribute)	Category	2

Table 5.1 shows the input attributes including 'Age', 'Marital status', 'Occupation', 'Basis of diagnosis', 'Topography', 'Morphology', 'Extent', 'Stage', 'Received surgery', 'Received radiation', 'Received chemotherapy', and 'Survivability status'. In order to build the breast cancer survivability prediction models, uncompleted and duplicated data are deleted. The class attribute is divided into 'Dead' and 'Alive' classes. The 'Dead' class refers to patients who died within the five years period prior to the first diagnosis, while the 'Alive' class refers to patients who were still alive for five years or more after the first diagnosis. In this way, data consist of 738 instances. The 'Dead' class is composed of 394 instances, and the 'Alive' class is composed of 342 patients.

5.4 Validations

In order to validate the capability and effectiveness of the proposed filtering outliers approach, both accuracy (see Section 2.2.3.1) and Area Under the receiver operating characteristic Curve (AUC) (see Section 2.2.3.3) of prediction models including C4.5, Conjunctive Rule, Naïve Bayes, Nearest Neighbour classifier (NN-classifier), Random Committee, Random Forests and a Radial Basis Function Network (RBFNetwork), are used. *A*ccuracy is used to present the percentage of correctness of outcome among the

test sets. On the other hand, *AUC* is used to evaluate the predictive ability of learning algorithms by calculating the area under the Receiver Operating Characteristic (ROC) curve. Moreover, stratified 10-fold cross-validation is used to divide the data set into a training set and a test set. The training set is used to generate the prediction model and the test set is used to evaluate the prediction model.

5.5 Performance and effectiveness of classifiers

In this section, WEKA version 3.5.6 [92] and LIBSVM [28] data mining environments are selected. The WEKA environment is a well-defined framework, and offers a variety of learning algorithms. Several numbers of experiments are conducted to compare the performance and effectiveness of our proposed approach. The default baseline learner used in these experiments is first presented. Then the number of identifying and eliminating instances using our approaches (C-Support Vector Classification Filtering (C-SVCF)) and other approaches including AdaBoost Filtering (ABF), Bagging Filtering (BF), AdaBoost with SVM Filtering (ABSVMF) and Bagging with SVM Filtering (BSVMF) are exhibited. Finally, the results of these experiments are presented and discussed.

5.5.1 Default baseline learners

In order to investigate the capability and performance of outlier filtering approaches, five algorithms are used. Firstly C-Support Vector Classification Filtering (C-SVCF)) uses C-Support Vector Classification (C-SVC) with a radial basis function as a base learner to solve a quadratic optimisation problem. Secondly, AdaBoost Filtering (ABF) employs AdaBoost with Decision Stump as a base learner to build a decision tree. Thirdly, Bagging Filtering (BF) utilises Bagging with a fast decision tree learner (REP-

Tree) as a base learner to also build a decision tree. Fourthly, AdaBoost with SVM Filtering (ABSVMF) exploits AdaBoost with C-SVC and a radial basis function as a base learner. Lastly, Bagging with SVM Filtering (BSVMF) makes use of Bagging with C-SVC and a Radial Basis Function as a base learner.

5.5.2 Outliers identification

In order to identify outliers, the number of outliers from misclassified instances is used. In this section, the number of outliers using the C-Support Vector Classification Filtering (C-SVCF) approach is compared with the number of outliers using AdaBoost Filtering (ABF), Bagging Filtering (BF), AdaBoost with SVM Filtering (ABSVMF), and Bagging with SVM Filtering (BSVMF), respectively. The number of outliers is shown in Table 5.2 below.

Table 5.2: Number of outliers in the data set

Filtering Approaches	Filtered		Outliers		Total outliers	Percentage of outliers
	'Dead'	'Alive'	'Dead'	'Alive'		
C-SVCF	322	248	72	94	166	22.55
ABF	316	185	78	157	235	31.93
BF	283	241	111	101	212	28.80
ABSVMF	*338*	*292*	*56*	*50*	*106*	*14.02*
BSVMF	320	221	74	121	195	26.49

Table 5.2 exploits the number of outliers using C-SVCF, ABF, BF, ABSVMF, and BSVMF. The results of this experiment show that ABSVMF can identify outliers by 14.02%, followed by C-SVCF (22.55%), BSVMF (26.49%), BF (28.80%), and ABF (31.93%), respectively. This indicates that the model generated from the filtered data using ABSVMF may build more knowledge than other models. However, in order to evaluate the capability and effectiveness of these outlier filtering approaches in this study, the accuracy and AUC scores of the prediction models generated from the filtered data at 5%, 10%, 15% and 20% are validated and compared.

5.5.3 Accuracy of classifiers

In order to evaluate the capability and effectiveness of outlier filtering approaches including C-SVCF, ABF, BF, ABSVMF and BSVMF, the average accuracy of seven classifiers including C4.5, Conjunctive Rule, Naïve Bayes, Nearest Neighbour classifier (NN-classifier), Random Committee, Random Forests, and a Radial Basis Function Network (RBFNetwork), are used. The average accuracy of each classifier is calculated using 10-fold cross-validation. The experimental results are displayed in Tables 5.3, 5.4, 5.5, 5.6 and 5.7.

Table 5.3: Accuracy of classifiers using C-Support Classification Filtering

Classifiers	Accuracy of classifiers after removing outliers					Accuracy Improvement (%)
	0%	5%	10%	15%	20%	
C4.5	68.07	69.81	72.66	77.96	81.49	13.42
Conjunctive Rule	63.86	68.38	70.24	72.84	76.23	12.37
Naïve Bayes	68.21	70.82	74.77	77.16	81.15	12.94
NN-Classifier	**57.88**	**62.09**	**67.37**	**70.13**	**82.00**	**24.12**
Random Committee	59.51	63.09	68.88	72.84	83.53	24.02
Random Forests	60.73	64.81	69.34	74.76	83.53	22.80
RBF Network	67.12	69.96	73.41	77.32	78.95	11.83
Average	63.63	*66.99*	*70.95*	74.72	*80.98*	*17.36*

Table 5.4: Accuracy of classifiers using AdaBoost Filtering

Classifiers	Accuracy of classifiers after removing outliers					Accuracy improvement (%)
	0%	5%	10%	15%	20%	
C4.5	68.07	72.68	76.13	80.35	85.40	17.33
Conjunctive Rule	63.86	66.09	73.41	77.48	82.00	18.14
Naïve Bayes	68.21	72.39	75.23	79.55	84.38	16.17
NN-classifiers	57.88	58.23	62.84	67.89	73.34	15.46
Random Committee	59.51	60.66	65.41	69.33	76.91	17.40
Random Forests	**60.73**	**63.81**	**65.41**	**73.00**	**78.10**	**17.37**
RBF Network	67.12	72.10	74.32	77.80	84.04	16.92
Average	63.63	66.57	70.39	75.06	80.60	16.97

Table 5.5: Accuracy of classifiers using Bagging Filtering

Classifiers	Accuracy of classifiers after removing outliers					Accuracy improvement (%)
	0%	5%	10%	15%	20%	
C4.5	68.07	72.25	75.08	80.51	85.91	17.84
Conjunctive Rule	63.86	68.81	70.85	74.60	77.76	13.90
Naïve Bayes	68.21	71.67	74.92	78.12	82.17	13.96
NN-classifiers	57.88	58.94	64.05	69.17	77.08	19.20
Random Committee	59.51	60.23	67.07	72.04	79.46	19.95
Random Forests	**60.73**	**61.52**	**69.03**	**73.32**	**82.00**	**21.27**
RBF Network	67.12	70.67	74.62	77.00	81.49	14.37
Average	63.63	66.30	70.80	74.97	80.84	17.21

Table 5.6: Accuracy of classifiers using AdaBoost with SVM Filtering

Classifiers	Accuracy of classifiers after removing outliers					Accuracy improvement (%)
	0%	5%	10%	15%	20%	
C4.5	68.07	69.67	69.64	-	-	1.57
Conjunctive Rule	63.86	66.67	70.39	-	-	6.53
Naïve Bayes	68.21	70.39	73.11	-	-	4.90
NN-classifiers	57.88	59.66	66.01	-	-	8.13
Random Committee	**59.51**	**62.66**	**69.03**	-	-	**9.52**
Random Forests	60.73	62.95	69.64	-	-	8.91
RBF Network	67.12	68.53	72.36	-	-	5.24
Average	63.63	65.79	70.03	-	-	6.40

Table 5.7: Accuracy of classifiers using Bagging with SVM Filtering

Classifiers	Accuracy of classifiers after removing outliers					Accuracy improvement (%)
	0%	5%	10%	15%	20%	
C4.5	68.07	69.24	72.05	76.04	79.46	11.39
Conjunctive Rule	63.86	64.23	70.24	72.20	75.21	11.35
Naïve Bayes	68.21	70.96	73.41	76.04	79.46	11.25
NN-classifiers	57.88	59.08	61.78	69.01	74.36	16.48
Random Committee	**59.51**	**61.09**	**63.90**	**71.41**	**77.93**	**18.42**
Random Forests	60.73	61.23	67.98	74.12	78.27	17.54
RBF Network	67.12	69.81	71.75	75.88	79.63	12.51
Average	63.63	65.09	68.73	73.53	77.76	14.13

Tables of 5.3, 5.4, 5.5, 5.6 and 5.7 show the accuracy of classifiers after applying C-SVCF, ABF, BF, ABSVMF and BSVMF respectively, to identify and eliminate out-

liers. The results of this experiment show that the accuracy of NN-classifier achieves better results than other classifiers after applying C-SVCF while the accuracy of Random Forests improves more than other classifiers after applying ABF and BF. On the other hand, the accuracy of Random Committee improves more than other classifiers after applying ABSVF and BSVMF. This indicates that NN-classifier, Random Committee and Random Forests are sensitive to outliers. However, the average accuracy of classifiers improves up to 17.36%, 16.97%, 17.21%, 6.40% and 14.13% after applying C-SVCF, ABF, BF, ABSVMF and BSVMF, respectively. This proves that C-SVCF is capable of identifying and eliminating outliers from the data set in order to improve the accuracy of the prediction models.

5.5.4 AUC of classifiers

The Area Under the receiver operating characteristic Curve (AUC) is commonly used to evaluate the performance and effectiveness of prediction models (classifiers). In this context, it is employed to evaluate the capability of the proposed filtering approach, and compared with four outlier filtering approaches including AdaBoost Filtering (ABF), Bagging Filtering (BF), AdaBoost with Support Vector Machine Filtering (ABSVMF) and Bagging with Support Vector Machine Filtering (BSVMF), respectively. The overall AUC scores of seven classifiers are exhibited in Tables 5.8, 5.9, 5.10, 5.11 and 5.12.

Table 5.8: AUC scores of classifiers using C-Support Vector Classification Filtering

Classifiers	AUC scores of classifiers after removing outliers					AUC scores improvement (%)
	0%	5%	10%	15%	20%	
C4.5	69.67	72.60	74.90	79.90	80.40	10.73
Conjunctive Rule	63.86	64.70	71.30	69.70	72.90	9.04
Naïve Bayes	68.21	77.30	80.70	82.90	87.10	18.89
NN-classifiers	57.88	61.70	67.00	69.80	81.60	23.72
Random Committee	**59.51**	**65.80**	**70.40**	**76.70**	**89.20**	**29.69**
Random Forests	60.73	68.90	75.00	81.50	90.80	30.07
RBF Network	67.12	76.10	78.80	82.50	86.90	19.78
Average	63.85	*69.59*	*74.01*	77.57	*84.13*	*20.27*

Table 5.9: AUC scores of classifiers using AdaBoost Filtering

Classifiers	AUC scores of classifiers after removing outliers					AUC scores improvement (%)
	0%	5%	10%	15%	20%	
C4.5	69.67	73.50	75.80	79.80	85.30	15.63
Conjunctive Rule	63.86	67.40	70.30	74.80	80.50	16.64
Naïve Bayes	68.21	77.90	80.20	83.80	89.10	20.89
NN-classifiers	57.88	57.70	62.10	66.90	72.20	14.32
Random Committee	59.51	60.60	65.60	71.40	78.60	19.09
Random Forests	**60.73**	**66.00**	**69.50**	**75.20**	**82.80**	**22.07**
RBF Network	67.12	76.90	78.80	81.80	89.00	21.88
Average	63.85	68.57	71.76	76.24	82.50	18.65

Table 5.10: AUC scores of classifiers using Bagging Filtering

Classifiers	AUC scores of classifiers after removing outliers					AUC scores improvement (%)
	0%	5%	10%	15%	20%	
C4.5	69.67	70.00	76.30	80.90	85.60	15.93
Conjunctive Rule	63.86	69.50	68.50	72.60	73.90	10.04
Naïve Bayes	68.21	77.70	80.90	84.70	88.00	19.79
NN-classifiers	57.88	58.60	63.80	68.80	76.90	19.02
Random Committee	59.51	60.20	69.50	74.20	81.70	22.19
Random Forests	**60.73**	**66.20**	**74.20**	**78.60**	**86.90**	**26.17**
RBF Network	67.12	76.00	80.10	83.80	87.70	20.58
Average	63.85	68.31	73.33	77.66	82.96	19.10

Table 5.11: AUC scores of classifiers using AdaBoost with SVM Filtering

Classifiers	AUC scores of classifiers after removing outliers					AUC scores improvement (%)
	0%	5%	10%	15%	20%	
C4.5	69.67	70.70	76.00	-	-	6.33
Conjunctive Rule	63.86	68.90	68.60	-	-	4.74
Naïve Bayes	68.21	76.90	79.90	-	-	11.69
NN-classifiers	57.88	59.20	65.70	-	-	7.82
Random Committee	**59.51**	**64.10**	**75.60**	-	-	**16.09**
Random Forests	60.73	67.40	75.90	-	-	15.17
RBF Network	67.12	75.10	79.40	-	-	12.28
Average	63.85	68.90	74.44	-	-	10.59

Table 5.12: AUC scores of classifiers using Bagging with SVM Filtering

Classifiers	AUC scores of classifiers after removing outliers					AUC scores improvement (%)
	0%	5%	10%	15%	20%	
C4.5	69.67	70.20	72.80	75.30	80.60	10.93
Conjunctive Rule	63.86	66.40	68.90	68.40	72.20	8.34
Naïve Bayes	68.21	77.30	79.60	82.30	85.10	16.89
NN-classifiers	57.88	58.70	61.50	68.20	73.70	15.82
Random Committee	59.51	62.50	65.60	75.20	81.60	22.09
Random Forests	**60.73**	**65.40**	**71.50**	**78.80**	**85.00**	**24.27**
RBF Network	67.12	75.40	78.50	80.10	85.20	18.08
Average	63.85	67.99	71.20	75.47	80.49	16.63

Tables 5.8, 5.9, 5.10, 5.11 and 5.12 show the AUC scores of classifiers including C4.5, conjunctive rule, Naïve Bayes, Nearest Neighbour classifier (NN-classifier), Random Committee, Random Forests, and a Radial Basis Function Network (RBFNetwork) after applying outlier filtering approaches including C-SVCF, ABF, BF, ABSVMF and BSVMF respectively, to identify and eliminate outliers. The experimental results show that Random Committee provides a higher AUC score than other classifiers after applying C-SVCF or ABSVMF while the AUC score of Random Forests improves more than other classifiers after applying ABF, BF and BSVMF. This indicates that Random Committee and Random Forests are sensitive to outliers. However, the average AUC scores of classifiers improves up to 20.27%, 18.65%, 19.10%, 10.59% and 16.63% after

applying C-SVCF, ABF, BF, ABSVMF and BSVMF, respectively. As a result, C-SVCF is superior to ABF, BF, ABSVMF and BSVMF based on the elimination of 20% of outliers.

5.5.5 Discussion of classifier results

In medical databases, raw data contains outliers that do not follow the model behaviour [53]. Therefore, in this chapter a novel approach using C-Support Vector Classification Filtering (C-SVCF) is proposed and applied to a 5-year breast cancer survivability analysis. There are several findings of these results discussed below.

Firstly, the C-SVCF algorithm is used to identify and eliminate outliers from the 5-year breast cancer survivability data set. Results show that removing outliers can improve the prediction results and this has been especially evidenced in the better result achieved by using C-SVC in comparison with ABF, BF, ABSVMF and BSVMF. Likewise, Khoshgoftaar, Seliya and Gao [195] showed that the performance of classifiers improves after removing outliers from data sets. However, his results were based on a rule-based algorithm and a different level of attributes.

Secondly, the average prediction accuracy is improved by 17.35% and the average AUC scores improved by 20.28% after removing 20% of outliers using C-SVC filtering. Although these experimental results disagree with the experimental results of Brodley and Friedl [25], who demonstrated that removing 10% or less of outliers is insignificant in prediction results, in this context removing outliers from 5% to 10% significantly improved the prediction results of classifiers. For example, the average accuracy of the prediction models has been improved by 6.76%, 7.17%, 6.40%, 5.10% and 7.32%, after removing 10% of outliers using AdaBoost, Bagging, AdaBoost with SVM, Bagging with SVM, and C-SVC, respectively.

Finally, C-SVCF provides a better accuracy and AUC score of predictions than that made by ABF, BF, ABSVMF and BSVMF. This may be due to the fact that C-SVCF is based on the C-Support Vector Classification technique which is suited to resolving the binary classification problems [39]. Unlike C-SVCF, AdaBoost filtering (ABF) is based on a Decision Stump while Bagging Filtering (BF) is based on a Fast decision tree learner (REPTree). Although ABSVMF and BSVMF also use C-SVC with a radial basis function as a base learner, the results of the identifying are affected by the weighing mechanism of AdaBoost and Bagging which can be misleading in separating the class labels.

5.6 Chapter summary

In this chapter, the C-SVCF algorithm has been proposed to identify and eliminate outliers in order to improve the quality of a 5-year breast cancer survivability data set. The experimental results showed that C-SVCF can generated a better data quality than that made by ABF, BF, ABSVM and BSVM based on the accuracy and AUC of seven classifiers including C4.5, Conjunctive Rule, Naïve Bayes, NN-classifier, Random Committee, Random Forests and RBFNetwork. Although using outlier filtering approach to eliminate misclassified instances from the full data set can increase the performance of classifiers, the bias may arise in the results. In the next chapter, an alternative approach to improving the data quality related to outliers and imbalanced data problems will be investigated.

Chapter 6

Improving Data Space via Combining Outlier Filtering and Over-Sampling

In Chapter 5, the C-Support Vector Classification Filtering (C-SVCF) approach was proposed to improve the quality of 5-year breast cancer survivability data in order to build accurate and reliable prediction models. Accuracy and Area Under the receiver operating characteristic Curve (AUC) of well-known classifiers were used to evaluate the capability and effectiveness of the proposed approach. The experimental results indicated that this approach is superior to AdaBoost, Bagging, AdaBoost with Support Vector Machine and Bagging with Support Vector Machine filtering approaches.

In this chapter, problems of outliers filtering and imbalanced data to improve the quality of data are addressed in order to further improve the performance and effectiveness of classifiers. Work introduced in this chapter has been accepted for publication as follows:

- J. Thongkam, G. Xu, Y. Zhang and F Huang, Toward breast cancer survivability prediction models through improving training space, Expert Systems with Applications, 2009.

In addressing this problem, a combination of Outlier filtering and Over-Sampling (OOS) approaches is proposed to further improvements in breast cancer survivability data quality. In order to evaluate the capability and effectiveness of the proposed

approach, not only accuracy and Area Under the receiver operating characteristic Curve (AUC) are used, but also sensitivity, specificity, and *F*-measure are employed. Moreover, the capability and effectiveness of the proposed approach is compared with outlier filtering and over-sampling approaches. The results of experiments are presented and discussed. The chapter then concludes with a summary.

6.1 Overview and approaches

In order to evaluate a prediction model in classification problems, a data set is commonly divided into training and test sets. The training set is used to build a prediction model while the test set is utilised to evaluate the model. However, most learning algorithms rarely handle both outlier and imbalanced data problems which commonly occur in medical data sets [152] [196] [30].

Outlier refers to an instance which does not follow the common rules [146] [147]. This kind of data affects the performance of prediction models. There are three main outlier handing approaches including robust algorithm, outlier filtering and outlier correction. However, outlier filtering is the simplest approach and cost of computation is inexpensive.

Imbalanced data refers to one class in a data set which outnumbers instances in the other classes [22] [156]. This problem commonly occurs in many fields including medical credit card fraud detection and software defect prediction [157] [158]. Usually, a learning algorithm provides a poor performance when utilised in imbalanced data, thus resulting in bias towards the majority class [30]. In relation to imbalanced data problems, much research has used the re-sampling approach including under-sampling and over-sampling [156]. Under-sampling is used to decrease the size of the majority

class as the same size of the minority class, whereas over-sampling is used to increase the size of the minority class as the same size of the majority class. However, over-sampling has been demonstrated to handle imbalanced data better than under-sampling [160] [157].

These problems are commonly handled by an outlier filtering approach [147] [55]. On the other hand, over-sampling and under-sampling are used to handle imbalanced data [197] [157]. Due to the limitations of learning algorithms, this chapter proposes a hybrid approach using combination of outlier filtering and over-sampling approaches to improve the quality of breast cancer survivability data sets in order to build an accurate breast cancer prediction models to yield better classification results. In order to present the background of the approaches, outlier filtering and over-sampling approaches are revealed, followed by the combined approach and the research framework which are introduced in order to understand the proposed approach.

6.1.1 Outlier filtering

An outlier filtering approach commonly utilises distance measures to detect outsider instances that are at a substantial distance from the others. In order to improve the performance of classifiers, identifying and eliminating is a challenging task in pattern recognition and instance-based learners which generally employ *k*-Nearest Neighbour (*k*-NN) [25] [53]. However, Support Vector Machine (SVM) has also been used to filter out outliers to assist in improving the performance of classifiers. For instance, Moffitt, Phan, Hemby and Wang [198] successfully employed a support vector machine technique to filter out outliers from gene marker section data. Their results demonstrated that removing outliers resulted in an enhanced prediction model. Similarly, Marquez, Paredes and Garcia-Gabin [199] utilised SVM to improve the quality of images. Their

results indicated that after using SVM the visual quality of images was enhanced. However, few research studies have used SVM to identify outliers in the medical field. Therefore, this chapter utilises Support Vector Machine to filter out outliers, based on experimental results in chapter five.

6.1.2 Over-sampling

An over-sampling approach is commonly used in the problem of imbalanced data, due to the fact that it significantly improves the performance of classifiers [22] [156] [159]. This approach is a non-heuristic method used to balance the class distribution through the random replication of a minority class [156]. It is used to increase the training data size and enhance the performance of classifiers [161]. In this way, both the majority and minority classes are able to achieve a similar size, which has less effect on the performance of the classifiers in relation to predicting the unseen data.

Recently, the Synthetic Minority Over-sampling TEchnique (SMOTE) is commonly used to resize the imbalanced data. SMOTE uses a synthetic minority over-sampling technique to match the majority class by taking minority class instances and introducing synthetic instances [197]. For example, He, Han and Wang [157] employed SMOTE to ensemble classifiers trained on data sets. Their results indicated that SMOTE can enhance the C4.5 classifier performance. Similarly, Pelayo and Dick [161] employed SMOTE to resize the minority class to match the majority class. Their results highlighted that SMOTE improved the accuracy of four NASA benchmark models. However, producing synthetic instances using SMOTE did not seem to fit well, due to new instances that could lead to misinterpretation of patterns. Therefore, this chapter generally employs a simple over-sampling approach to compare with the proposed approach.

6.1.3 Combined approaches

In medical databases, raw data normally contains outliers [53] and imbalanced data which affect the performance of prediction models (classifiers) [156] [20] [200]. In relation to improving the quality of data which are contained in outliers and imbalanced data problems, much research has combined outlier filtering and re-sampling approaches in fraud detection data sets. However, a few researchers have applied this in relation to medical data. For instance, Padmaja, Dhulipalla, Bapi and Krishna [30] employed *k*-Nearest Neighbour (*k*-NN) to eliminate outliers in a minority class, and applied over-sampling to increase the size of this minority class while applying under-sampling to reduce the size of the majority class. In contrast, in this chapter the framework starts with cleaned data sets (without duplicated and missing data). Then an outlier filtering approach using C-Support Vector Classification (C-SVC) with radial basis function to identify outliers from both classes is applied. Following this, an over-sampling approach is utilised to resize the minority class to match the size of the majority class. The framework for combining the Outlier filtering and Over-Sampling approaches (called OOS) [201] is illustrated in Figure 6.1.

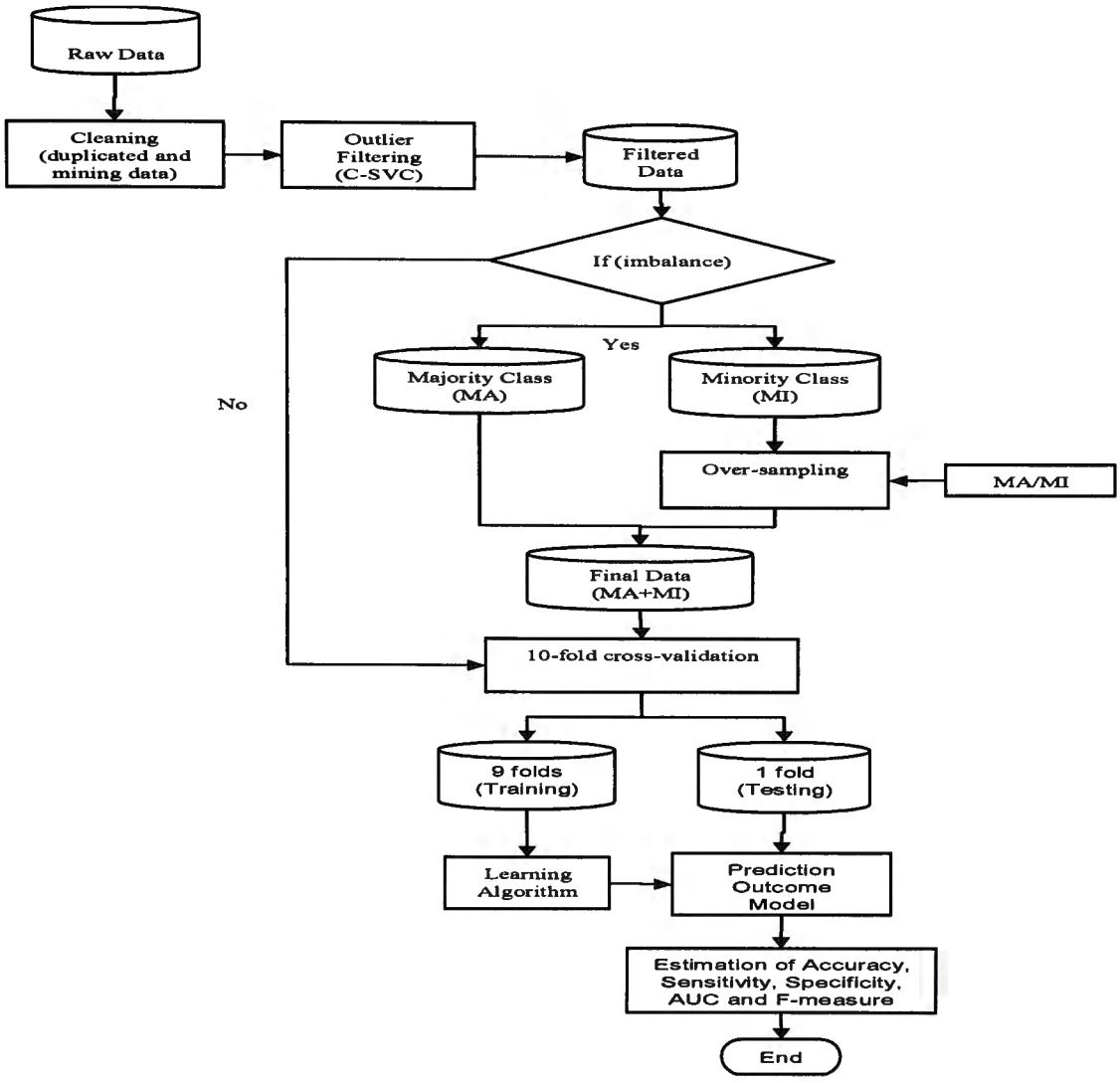


Figure 6.1: Combined outlier filtering and over-sampling framework

Figure 6.1 shows the OOS framework which is summarised in four main steps as follows:

Step 1: the C-Support Vector Classification filtering (C-SVCF) approach is used to identify and eliminate outliers from both 'Dead' and 'Alive' classes in the original data sets;

Step 2: the filtered data sets are divided into minority and majority classes;

Step 3: the over-sampling approach is utilised to increase the size of the minority class to the same size as the majority class by using the ratio between majority and minority classes; and

Step 4: the majority and minority classes are combined into a new data set which becomes a balanced data set.

In this way, the quality of data would be improved and suited to building an accurate and reliable prediction model.

6.2 Breast cancer survivability data sets

In order to evaluate the performance and effectiveness of the proposed approach, the breast cancer survivability data and survival periods are expanded using data from 1985 to 2006. The data contain 13 attributes and a class attribute from both patient information and the treatment choice of patients diagnosed with breast cancer. The input attributes are presented in Table 6.1 below.

Table 6.1: Input attributes

No.	Attributes	Attribute Types	Values
1	Age	Number	-
2	Marital status	Category	3
3	Basis of diagnosis	Category	6
4	Topography	Category	9
5	Morphology	Category	14
6	Extent	Category	4
7	Stage	Category	4
8	Received surgery	Category	2
9	Received radiation	Category	2
10	Received chemotherapy	Category	2
11	Received hormonal therapy	Category	2
12	Received supportive therapy	Category	2
13	Received other therapy	Category	2
14	Survivability status (class attribute)	Category	2

Table 6.1 shows 14 attributes used in this chapter. These attributes include 'Age', 'Marital status', 'Basis of diagnosis', 'Morphology', 'Extent', 'Stage', 'Received surgery', 'Received radiation', 'Received chemotherapy', 'Received hormonal therapy', 'Received supportive therapy', 'Received other therapy' and 'Survivability status'. In order to obtain 1-, 2-, 3-, 4-, 5-, 6-, 7-, 8-, 9- and 10-year breast cancer survivability data

sets, 10 periods of patient's survival are applied. As a result, each data set has unique numbers of instances, as displayed in Table 6.2.

Table 6.2: The number of instances in original data sets

Data Sets	Years	'Dead'	'Alive'	Total	'Dead' (% Positive Class)	'Alive' (%Negative Class)
1-year	1985-2006	351	1128	1479	23.73	76.27
2-year	1985-2005	455	846	1301	34.97	65.03
3-year	1985-2004	485	654	1139	42.58	57.42
4-year	1985-2003	488	495	983	49.64	50.36
5-year	1985-2002	466	392	858	54.31	45.69
6-year	1985-2001	437	304	741	58.97	41.03
7-year	1985-2000	351	198	549	63.93	36.07
8-year	1985-1999	276	130	406	67.98	32.02
9-year	1985-1998	248	103	351	70.66	29.34
10-year	1985-1997	221	90	311	71.06	28.94

Table 6.2 shows the percentage of imbalanced data in each data set. Each data set involves two classes including 'Dead' and 'Alive'. In the case of a 1-year breast cancer survivability data set obtained from 1958 to 2006, the 'Dead' class refers to patients who die within one year after the first diagnosis, while the 'Alive' class refers to patients who are still alive for one year or more after the first diagnosis. Similarly, the 'Dead' class in 2-, 3-, 4-, 5-, 6-, 7-, 8-, 9- and 10-year breast cancer survivability data sets refer to patients who die within two, three, four, five, six, seven, eight, nine and 10 years respectively, after the first diagnosis. Unlike the 'Dead' class, the 'Alive' class in 2-, 3-, 4-, 5-, 6-, 7-, 8-, 9- and 10-year breast cancer survivability data sets refer to patients who are alive for two or more, three or more, four or more, five or more, six or more, seven or more, eight or more, nine or more and 10 years or more, respectively after the first diagnosis. Such data sets have imbalanced data problems which affect the performance of prediction models. This indicates a need to fully understand the data set before constructing prediction models.

6.3 Evaluation approaches

In order to evaluate the Outlier filtering and Over-Sampling (OOS) approach, five evaluation methods including accuracy, sensitivity, specificity (see Section 2.2.3.1), Area Under the receiver operating characteristic Curve (AUC) (see Section 2.2.3.3) and *F*-measure (see Section 2.2.3.4) of prediction models are applied. *Accuracy* refers to the percentage of the correctness of outcomes among the test sets. *Sensitivity* refers to the true positive rate while *specificity* refers to the true negative rate. *AUC* uses the area under true positive and false positive rates to calculate the score in order to evaluate the predictive ability of learning algorithms while *F-measure* is an evaluation method based on recall and precision. The experiment procedures are performed using 10 iterations of the stratified 10-fold cross-validation to reduce the bias and variance associated with the classification results.

6.4 Experimental results

In order to evaluate the capability and effectiveness of a hybrid of the Outlier filtering and Over-Sampling (OOS) approaches, the WEKA experimenter version 3.5.6 [92] is selected. This is due to the fact that it provides a variety of learning algorithms used in data mining, pattern recognition, and machine learning. Four well-known algorithms including AdaBoost, Bagging, C4.5 and Support Vector Machine (SVM) are employed to present the capability and effectiveness of OOS approach. The best results are presented after repeating the simple over-sampling approach on the data sets many times until the insignificant difference in the results is reached. The default algorithm setting used in these experiments is first presented. Then the ratio of imbalanced data in major-

ity and minority classes is discussed. Finally, the experimental results are presented and discussed.

6.4.1 Default algorithm setting

In order to perform these experiments, there are four defaults of algorithm setting including:

1) AdaBoost uses a Decision Stump as a base learner;
2) Bagging utilises a fast decision tree learner as a base learner;
3) C4.5 uses 0.25 confidence factor for pruning; and
4) SVM uses the C-Support Vector Classifiers (C-SVC) type with a radial basis function.

6.4.2 Imbalance data

In order to define the imbalanced data in this section, the ratio between the majority class and the minority class is used. The numbers of remaining instance using an OOS approach are compared with outlier filtering and over-sampling. Results are exhibited in Table 6.3.

Table 6.3: The number of instances using outlier filtering , over-sampling and OOS approaches

Data Sets	Original			Outlier filtering			Over-sampling			OOS		
	'Dead'	'Alive'	Ratio (%) (MA/MI)[1]	'Dead'	'Alive'	Ratio (%) (MA/MI)	'Dead'	'Alive'	Ratio (%) (MA/MI)	'Dead'	'Alive'	Ratio (%) (MA/MI)
1-year	351 (MI)	1128 (MA)	321	87 (MI)	1093 (MA)	1256	1126	1128	1.00	1092	1093	1.00
2-year	455 (MI)	846 (MA)	186	159 (MI)	799 (MA)	503	846	846	1.00	798	799	1.00
3-year	485 (MI)	654 (MA)	135	270 (MI)	544 (MA)	201	653	654	1.00	542	544	1.00
4-year	488 (MI)	495 (MA)	101	349 (MI)	355 (MA)	102	492	495	0.99	355	355	1.00
5-year	466 (MA)	392 (MI)	119	368 (MA)	250 (MI)	147	466	466	1.00	368	367	1.00
6-year	437 (MA)	304 (MI)	144	378 (MA)	153 (MI)	247	437	437	1.00	378	377	1.00
7-year	351 (MA)	198 (MI)	177	316 (MA)	77 (MI)	410	351	350	1.00	316	315	1.00
8-year	276 (MA)	130 (MI)	212	265 (MA)	36 (MI)	736	276	275	1.00	265	264	1.00
9-year	248 (MA)	103 (MI)	241	238 (MA)	24 (MI)	992	248	248	1.00	238	238	1.00
10-year	221 (MA)	90 (MI)	246	212 (MA)	26 (MI)	815	221	220	1.00	212	211	1.00

[1] MA refers to the majority class and MI refers to the minority class

Table 6.3 shows the numbers of instances that incurred the imbalanced data problems. Results present that 4- and 5-year breast cancer survivability data sets are slightly more balanced than 1-, 2-, 3-, 6-, 7-, 8-, 9- and 10-year breast cancer survivability data sets. This problem of imbalanced data significantly increased after applying the outlier filtering approach, especially in 1-, 7-, 8-, 9- and 10-year breast cancer survivability data sets. This may be due to the fact that more patients with breast cancer are alive than die after the first four years of the first diagnosis whereas there are more dead patients with breast cancer than live after the first five years of the first diagnosis. Although applying an over-sampling approach can handle the imbalance problem, it cannot reduce outliers. However, the OOS approach can reduce outliers as well as handle the imbalanced problem in data sets.

6.4.3 Accuracy, sensitivity and specificity results

The capability and effectiveness of the OOS approach are evaluated, using the average of accuracy, sensitivity and specificity of four classifiers including AdaBoost, Bagging, C4.5 and Support Vector Machine (SVM), and are compared with outlier filtering and over-sampling approaches. The results of this experiment are shown in Tables 6.4, 6.5, 6.6 and 6.7, respectively.

Table 6.4: Accuracy, sensitivity and specificity of AdaBoost

Data Sets	Accuracy (%)				Sensitivity (%)				Specificity (%)			
	Raw	Outlier	OS[2]	OOS[3]	Raw	Outlier	OS	OOS	Raw	Outlier	OS	OOS
1-year	78.48	96.07	68.74	93.00	25.67	72.90	61.68	93.01	94.91	97.92	75.78	93.00
2-year	69.70	90.60	63.85	92.15	38.59	59.54	69.87	93.65	86.43	96.78	57.84	90.66
3-year	67.83	86.99	67.38	88.24	48.57	76.11	84.18	86.68	82.09	92.39	50.61	89.80
4-year	66.27	84.42	68.50	87.44	80.89	91.87	68.99	91.54	51.85	77.12	68.02	83.34
5-year	68.88	87.77	66.75	87.25	82.92	95.84	65.33	95.87	52.19	75.88	68.19	78.61
6-year	69.41	88.26	68.09	89.37	81.10	92.64	78.20	90.32	52.61	77.42	57.99	88.41
7-year	68.32	93.69	63.72	92.55	82.23	98.19	67.69	92.91	43.72	75.38	59.74	92.19
8-year	68.97	92.64	66.48	92.17	87.56	94.41	67.00	85.30	29.54	80.00	65.96	99.05
9-year	71.54	97.64	67.13	95.67	92.46	99.45	64.88	92.61	21.04	80.33	69.38	98.74
10-year	71.64	96.69	66.33	96.62	92.18	98.11	68.18	95.56	21.22	85.33	64.45	97.68
Average	70.10	**91.48**	66.70	91.45	71.22	87.91	69.60	**91.75**	53.56	83.86	63.80	**91.15**

Table 6.5: Accuracy, sensitivity and specificity of Bagging

Data Sets	Accuracy (%)				Sensitivity (%)				Specificity (%)			
	Raw	Outlier	OS	OOS	Raw	Outlier	OS	OOS	Raw	Outlier	OS	OOS
1-year	77.45	96.50	76.85	98.01	26.38	72.86	81.03	100.00	93.34	98.41	72.68	96.02
2-year	67.86	94.24	71.30	96.99	37.33	81.53	74.87	98.93	84.28	96.77	67.73	95.04
3-year	65.64	92.86	71.75	95.48	54.35	91.04	76.78	97.60	73.99	93.77	66.73	93.36
4-year	65.67	90.74	71.09	91.46	69.28	90.20	74.87	92.31	62.09	91.27	67.33	90.62
5-year	67.04	92.14	70.31	94.39	74.31	93.97	68.70	92.98	58.39	89.44	71.91	95.80
6-year	68.83	92.08	69.58	94.17	78.63	94.99	69.23	92.33	54.75	84.89	69.92	96.03
7-year	66.63	94.68	71.17	96.15	80.54	96.68	68.50	95.29	41.98	86.55	73.86	97.02
8-year	69.86	93.81	71.20	97.33	87.38	95.70	67.33	94.69	32.69	79.67	75.07	100.00
9-year	70.89	97.63	74.09	99.14	89.03	98.69	72.55	98.28	27.07	87.67	75.66	100.00
10-year	69.87	95.46	70.72	97.85	88.10	96.18	66.09	96.23	25.11	90.33	75.36	99.48
Average	68.97	94.01	71.81	**96.10**	68.53	91.18	72.00	**95.86**	55.37	89.88	71.63	**96.34**

Table 6.6: Accuracy, sensitivity and specificity of C4.5

Data Sets	Accuracy (%)				Sensitivity (%)				Specificity (%)			
	Raw	Outlier	OS	OOS	Raw	Outlier	OS	OOS	Raw	Outlier	OS	OOS
1-year	78.38	95.96	75.95	98.53	25.61	62.00	77.62	100.00	94.80	98.68	74.29	97.05
2-year	69.32	94.13	68.93	97.49	32.80	79.33	73.80	99.00	88.96	97.08	64.06	96.00
3-year	67.26	92.08	72.04	95.17	61.47	88.19	74.72	96.59	71.55	94.01	69.36	93.75
4-year	67.15	90.19	70.88	91.76	76.53	88.57	78.69	90.99	57.89	91.78	63.11	92.53
5-year	68.64	92.64	70.36	94.12	78.32	93.48	74.46	93.29	57.12	91.40	66.26	94.96
6-year	68.84	93.17	69.06	94.93	78.58	95.39	73.25	94.02	54.85	87.68	64.87	95.84
7-year	67.58	95.39	67.96	97.67	78.52	97.49	66.79	96.90	48.17	86.91	69.14	98.44
8-year	68.93	95.21	67.61	97.49	89.97	97.01	63.54	94.99	24.31	81.67	71.71	100.00
9-year	71.68	97.56	73.61	99.37	92.82	98.61	71.35	98.74	20.75	87.83	75.86	100.00
10-year	66.72	96.17	70.52	98.84	88.52	97.97	67.46	97.97	13.22	81.83	73.59	99.71
Average	69.45	94.25	70.69	**96.54**	70.31	89.80	72.17	**96.25**	53.16	89.89	69.23	**96.83**

2 OS refers to Over-Sampling

3 OOS refers to Outlier and Over-Sampling

Table 6.7: Accuracy, sensitivity and specificity of SVM

Data Sets	Accuracy (%)				Sensitivity (%)				Specificity (%)			
	Raw	Outlier	OS	OOS	Raw	Outlier	OS	OOS	Raw	Outlier	OS	OOS
1-year	77.16	98.18	73.89	99.76	16.67	79.46	73.25	100.00	95.98	99.68	74.53	99.51
2-year	69.35	96.46	69.37	98.56	28.18	86.36	71.39	99.95	91.49	98.47	67.37	97.17
3-year	66.06	94.69	68.85	96.93	48.77	89.67	71.45	98.19	78.86	97.19	66.24	95.68
4-year	63.65	93.87	69.56	94.41	63.72	93.95	68.92	94.37	63.58	93.78	70.19	94.45
5-year	64.84	93.90	68.80	97.46	73.77	96.31	68.21	96.22	54.21	90.36	69.38	98.69
6-year	65.71	95.29	67.23	98.03	81.65	98.70	71.65	97.30	42.79	86.86	62.81	98.76
7-year	64.43	95.86	66.32	98.64	86.10	99.40	62.83	97.72	26.04	81.30	69.83	99.56
8-year	70.08	97.41	72.74	98.68	92.65	99.43	72.30	97.34	22.15	82.33	73.19	100.00
9-year	72.23	98.48	73.26	99.56	94.68	100.00	67.63	99.12	18.09	84.17	78.90	100.00
10-year	71.93	96.90	70.45	99.27	94.21	99.29	66.70	98.53	17.22	78.67	74.23	100.00
Average	68.54	96.10	70.05	**98.13**	68.04	94.26	69.43	**97.87**	51.04	89.28	70.67	**98.38**

Tables 6.4, 6.5, 6.6 and 6.7 show the accuracy, sensitivity and specificity of AdaBoost, Bagging, C4.5 and SVM, respectively. Although the average accuracy of AdaBoost based on outlier filtering is slightly better than the average accuracy of AdaBoost based on OOS, the average of the sensitivity and specificity of AdaBoost using the OOS approach is much better than using the outlier filtering approach. Nonetheless, the problem of being low in sensitivity towards 1-, 2- and 3-year breast cancer survivability data sets remained in AdaBoost after applying the outlier filtering approach. Similarly, the specificity of 6-, 7-, 8- and 10-year survivability data sets is unable to achieve a high performance after applying the outlier filtering approach. Moreover, the sensitivity of a 5-year survivability data set of all four classifiers using outlier filtering gives similar results to the OOS approach, and the specificity of a 5-year breast cancer survivability data set of all four classifiers using outlier filtering gives lower results than the OOS approach. This is may be due to the fact that basic AdaBoost concentrates too much on misclassified instances. This means that the AdaBoost model is much affected by imbalanced data leading to an overfitting problem.

Although using the over-sampling approach slightly improved the performance of the minority class, low overall performance remained in both sensitivity and specificity.

This means that using the over-sampling approach alone can only slightly improve the average of accuracy, sensitivity and specificity. This proves that the combination of the outlier filtering and over-sampling approach is suitable for improving the overall accuracy, sensitivity and specificity of classifiers in breast cancer survivability data sets. This proves that an OOS approach provides better quality data than outlier filtering and over-sampling approaches.

6.4.4 AUC results

In this section, Area Under the receiver operating characteristic Curve (AUC) of four classifiers including AdaBoost, Bagging, C4.5 and Support Vector Machine (SVM) is utilised to evaluate the capability and effectiveness of the proposed approach and compare it with the outlier filtering and over-sampling approach. The AUC scores are shown in Tables 6.8, 6.9, 6.10 and 6.11, respectively.

Table 6.8: AUC scores of AdaBoost

Data Sets	Raw Data	Outlier filtering	Over-sampling	OOS
1-year	75.13	97.83	75.15	98.24
2-year	72.50	96.03	69.61	96.89
3-year	73.22	94.90	72.12	94.99
4-year	70.56	92.96	74.58	94.17
5-year	72.41	93.44	73.49	93.36
6-year	73.17	93.36	71.87	93.44
7-year	68.68	95.05	70.44	96.49
8-year	67.41	96.33	73.69	96.46
9-year	68.15	98.28	73.55	99.34
10-year	67.62	98.15	71.40	98.98
Average	70.89	95.63	72.59	**96.24**

Table 6.9: AUC scores of Bagging

Data Sets	Raw Data	Outlier filtering	Over-sampling	OOS
1-year	72.80	96.16	84.53	99.02
2-year	69.92	96.82	78.09	98.60
3-year	72.59	97.27	79.44	98.38
4-year	71.72	95.86	78.88	96.31
5-year	71.62	95.96	77.29	96.78
6-year	71.27	95.21	76.17	97.61
7-year	68.06	93.22	77.57	98.59
8-year	68.04	98.16	78.57	99.21
9-year	67.61	97.57	82.08	99.59
10-year	67.16	97.36	78.61	99.37
Average	70.08	96.36	79.12	**98.35**

Table 6.10: AUC scores of C4.5

Data Sets	Raw Data	Outlier filtering	Over-sampling	OOS
1-year	60.86	84.80	81.47	98.62
2-year	69.52	93.38	73.36	98.07
3-year	72.16	93.41	76.63	96.24
4-year	71.10	91.17	76.14	93.18
5-year	71.59	93.04	74.10	94.24
6-year	70.68	93.32	71.63	96.14
7-year	67.78	90.53	70.46	97.51
8-year	67.17	93.64	71.94	98.94
9-year	66.95	94.77	78.06	99.67
10-year	55.32	90.51	75.34	98.71
Average	67.31	91.86	74.91	**97.13**

Table 6.11: AUC scores of SVM

Data Sets	Raw Data	Outlier filtering	Over-sampling	OOS
1-year	56.33	89.57	73.89	99.76
2-year	59.84	92.42	69.38	98.56
3-year	63.82	93.43	68.85	96.94
4-year	63.65	93.87	69.55	94.41
5-year	63.99	93.33	68.80	97.45
6-year	62.22	92.78	67.23	98.03
7-year	56.07	90.35	66.33	98.64
8-year	57.40	90.88	72.74	98.67
9-year	56.39	92.08	73.26	99.56
10-year	55.72	88.98	70.46	99.27
Average	59.54	91.77	70.05	**98.13**

Tables 6.8, 6.9, 6.10 and 6.11 display the overall average of AUC scores of AdaBoost, Bagging, C4.5 and SVM in order to evaluate the capability and effectiveness of the proposed approach (OOS). The results of this experiment show that the overall average of

AUC scores of AdaBoost, Bagging, C4.5 and SVM improves by up to 25.35%, 28.27%, 29.22% and 38.59%, respectively after applying the OOS approach. Likewise, the overall average of AUC scores of AdaBoost, Bagging, C4.5 and SVM improves 24.74%, 26.28%, 24.55% and 32.23%, using the outlier filtering approach. However, the overall average of AUC scores of AdaBoost, Bagging, C4.5 and SVM using the random over-sampling approach only somewhat improves at 1.7%, 9.04%, 7.6% and 10.51%, respectively. This proves that AUC scores of classifiers improve more after applying the OOS approach than after applying outlier and over-sampling alone.

6.4.5 *F*-measure results

In order to measure the capability and effectiveness of the proposed approach, *F*-measure of four classifiers including AdaBoost, Bagging, C4.5 and Support Vector Machine (SVM) is utilised. The *F*-measure results are illustrated in Tables 6.12, 6.13, 6.14 and 6.15, respectively.

Table 6.12: *F*-measure of AdaBoost

Data Sets	Raw Data	Outlier filtering	Over-sampling	OOS
1-year	35.78	73.01	66.27	92.99
2-year	46.84	67.35	64.72	92.25
3-year	55.68	79.04	71.93	87.99
4-year	70.17	85.48	68.39	87.96
5-year	74.30	90.34	66.10	88.32
6-year	75.74	91.80	70.87	89.45
7-year	76.73	96.17	64.99	92.57
8-year	79.18	95.72	66.56	91.46
9-year	82.06	98.71	66.05	95.44
10-year	82.15	98.13	66.79	96.52
Average	67.86	87.58	67.27	**91.50**

Table 6.13: *F*-measure of Bagging

Data Sets	Raw Data	Outlier filtering	Over-sampling	OOS
1-year	35.58	75.01	77.74	98.06
2-year	44.66	82.26	72.25	97.05
3-year	57.22	89.43	73.08	95.58
4-year	66.62	90.60	72.02	91.52
5-year	70.96	93.44	69.75	94.31
6-year	74.78	94.44	69.34	94.04
7-year	75.46	96.69	70.27	96.10
8-year	79.67	96.43	70.01	97.21
9-year	81.15	98.69	73.54	99.12
10-year	80.49	97.38	69.13	97.79
Average	66.66	91.44	71.71	**96.08**

Table 6.14: *F*-measure of C4.5

Data Sets	Raw Data	Outlier filtering	Over-sampling	OOS
1-year	35.52	68.80	76.31	98.55
2-year	42.39	81.57	70.28	97.54
3-year	61.26	88.03	72.75	95.24
4-year	69.77	89.90	72.83	91.68
5-year	73.00	93.78	71.44	94.07
6-year	74.81	95.20	70.21	94.88
7-year	75.53	97.13	67.50	97.64
8-year	79.56	97.26	66.05	97.38
9-year	82.20	98.65	72.83	99.35
10-year	78.76	97.85	69.46	98.82
Average	67.28	90.82	70.97	**96.52**

Table 6.15: *F*-measure of SVM

Data Sets	Raw Data	Outlier filtering	Over-sampling	OOS
1-year	25.55	86.20	73.68	99.76
2-year	38.83	88.92	69.94	98.59
3-year	54.83	91.72	69.57	96.98
4-year	63.44	93.82	69.23	94.39
5-year	69.44	94.95	68.58	97.41
6-year	73.69	96.77	68.54	98.01
7-year	75.52	97.49	64.98	98.62
8-year	80.78	98.55	72.63	98.63
9-year	82.77	99.17	71.42	99.55
10-year	82.63	98.28	69.23	99.24
Average	64.75	94.59	69.78	**98.12**

Tables 6.12, 6.13, 6.14 and 6.15 show the *F*-measure of four classifiers including AdaBoost, Bagging, C4.5 and SVM. The experimental results present that the OOS approach improves the *F*-measure of AdaBoost, Bagging, C4.5 and SVM up to 3.92%,

4.64%, 5.7% and 3.53% respectively, when compared to the *F*-measure of AdaBoost, Bagging, C4.5 and SVM after applying the outlier filtering approach. Besides, the OOS approach improves the *F*-measure of AdaBoost, Bagging, C4.5 and SVM by up to 24.23%, 24.37%, 25.55% and 28.34% respectively, when compared to the *F*-measure of AdaBoost, Bagging, C4.5 and SVM after applying the over-sampling approach. This proves that the proposed approach is suitable for improving the quality of breast cancer survivability data.

6.4.6 Discussion of experimental results

Breast cancer survivability prediction models are used to assist medical practitioners to enhance the provision of care in medical prognoses. In order to refine the performance of these models, a hybrid approach combining outlier filtering and over-sampling approaches was proposed to improve the quality of the data set. Accuracy, sensitivity, specificity, AUC scores and *F*-measure of four learning algorithms including AdaBoost, Bagging, C4.5 and Support Vector Machine (SVM) were used to evaluate the prediction models. As a result, three main findings from experimental results are discussed below. Firstly, outliers and imbalanced data are found to be a direct effect of the performance and effectiveness of classifiers. A possible explanation for this might be that the performance of well-known classifiers is improved by eliminating a number of outliers from both the minority and majority classes, and by increasing the size of the minority class to the same size as the majority class. These findings are consistent with those of Padmaja, Dhulipalla, Bapi and Krishna [30] who found that the performance improvement of classifiers occurs only after firstly eliminating outliers in a minority class, then increasing the size of the minority class, and lastly decreasing the size of the majority class of fraud detection databases.

Secondly, the finding that AdaBoost is less affected by the problem of imbalance than in Bagging, C4.5 and SVM, is interesting. This finding may be evidenced by the improvement of accuracy after applying the outlier filtering and OOS approaches (see Table 4), where the accuracy of AdaBoost decreases by 0.03% and the accuracy of Bagging improves by 2.09%. Also, the accuracy of C4.5 improves by 2.29% and the accuracy of SVM improves by 2.03%, and the AUC score improves by 0.61%, 1.99%, 5.27% and 6.36% in the cases of AdaBoost, Bagging, C4.5 and SVM using OOS and Outlier filtering approaches. This may be due to the fact that the AdaBoost algorithm applies direct weight to instances in order to generate a separation line [32], while C4.5 and SVM use statistical calculations to separate the binary classes [111] [96].

Finally, SVM is found to be more accurate than AdaBoost, Bagging and C4.5. This may be due to C-SVC (one type of SVM algorithms) being utilised to eliminate outliers from data sets. In addition, the OOS approach is suited to improving the quality of data sets in order to enhance the prediction results of classifiers.

6.5 Chapter summary

The combining of Outlier filtering and Over-Sampling (OOS) approaches has been proposed to improve the data quality in order to develop accurate breast cancer survivability models from real-world data sets. The experimental results explicitly pointed out that the OOS approach is able to decrease insignificant outliers, as well as significantly increase the performance of classifiers. Moreover, applying the combination of outlier filtering and over-sampling, the average of accuracy, sensitivity, specificity, AUC score and *F*-measure of SVM were improved by 29.83%, 29.83%, 47.34%, 38.59% and 33.38%, respectively. In the next chapter, the combination of AdaBoost and Random

Forests used to develop breast cancer survivability prediction models will be investigated in order to select the suitable classifiers.

Chapter 7

Breast Cancer Survivability Prediction Models

In Chapter 6, the quality of data was improved using a combined outlier filtering and over-sampling (OOS) approach. In order to evaluate the capability and effectiveness of this approach, accuracy, sensitivity, specificity, Area Under the receiver operating characteristic Curve (AUC) and *F*-measure of classifiers have been analysed. The OOS approach has been proven to be superior to both outlier filtering and over-sampling approaches.

Developing accurate, stable and effective breast cancer survivability prediction models using data mining processes is a challenging task. Work introduced in this chapter has been published as follows:

- J. Thongkam, G. Xu and Y. Zhang, AdaBoost algorithm with random forests for predicting breast cancer survivability, in Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN2008), pp. 1-8, Jun. 1-6, Hong Kong, 2008.

In this chapter, a hybrid of AdaBoost and Random Forests is proposed to build breast cancer survivability prediction models. In order to understand the basic concept of algorithms used in this present study, AdaBoost, Random Forests and the combining of AdaBoost and Random Forests algorithms are discussed. In order to evaluate this approach, this study is divided into two parts. *Part-I* involves the development of a

5-year breast cancer survivability prediction model using only filtered data from 1990 to 2001. In order to evaluate the performance and effectiveness of this model, accuracy, sensitivity and specificity are used and also compared with 9 classifiers. The experimental results of Part-I are presented and discussed. *Part-II* contains the development of 3-, 5-, 8- and 10-year breast cancer survivability prediction models using filtered and balanced data from 1985 to 2006. In order to evaluate the performance and effectiveness of the proposed approach, accuracy, AUC score, *F*-measure and Kappa statistics are utilised and compared with 19 classifiers. In order to show this, experimental results in Part-II are presented and discussed. Finally, the summary concludes the chapter.

7.1 Overview and motivation

AdaBoost has become an attractive ensemble method in machine learning since it is low in error rates and performs well in low noise data sets [31] [32]. It is used to combine a weak learner to form a model with higher prediction outcomes [31] [202] [34] [203] [164]. Although this weak learner is also called a weak classifier and a base learner, for the rest of this thesis the term *weak learner* will be used. Generally, AdaBoost is combined with a Decision Stump as a weak learner [163] [204]. Recently, several research studies have merged AdaBoost with other weak learners to improve the accuracy of classifiers. For example, Szarvas, Farkas and Kocsor [205] successfully applied the combination of AdaBoost with the C4.5 decision tree in a multilingual named entity recognition system. Their results displayed that AdaBoost with C4.5 achieved *F*-measure up to 94.77%. Unlike Szarvas, Farkas and Kocsor; Sun, Wang and Wong [204] combined AdaBoost and associative classification using UCI Machine Learning

Repository databases. Their results indicated that this combined technique improved accuracy and reduced the construction time of the models. On the other hand, Li, Wang and Sung [34] combined AdaBoost with Support Vector Machine (SVM) to demonstrate that the combination of AdaBoost and SVM has better generalisation error rates than both basic AdaBoost and SVM alone. Zhang and Ren [16] supported the study of Li, Wang and Sung [34] by combining AdaBoost with SVM to improve the accuracy of the SVM classifier. They presented that this combined method is better than original SVM using five data sets at UCI Machine Learning Repository databases. Therefore, finding a suitable weak learner is a challenging task for improving the classification result not only in machine learning and pattern recognition but also in data mining.

This chapter proposes the combination of AdaBoost and Random Forests to build the breast cancer survivability prediction models due to the fact that it provides low error rates and few researchers have applied Random Forests as a weak learner in medical data sets. In order to evaluate the performance and effectiveness of this proposed method, several evaluation methods including accuracy, sensitivity and specificity, AUC, *F*-measure and Kappa statistics are exploited.

7.2 Related and a hybrid method

In order to build the breast cancer survivability prediction models, the theoretical background of AdaBoost, Random Forests and the proposed method are briefly described.

7.2.1 Basic AdaBoost

The AdaBoost algorithm is the most popular new, ensemble method for generating a better classifier [32] [180] [179]. It is flexible not only for combining with Decision

Stump but also with other weak learners to improve the performance and effectiveness of the prediction models [180] [179] [206]. Moreover, it requires less input parameters and less knowledge of computing background in improving the accuracy of prediction models [180] [179]. In relation to AdaBoost, Gentle AdaBoost is one of the attractive algorithms which has a high performance in several data sets and is available for combining with diverse classifiers [24] [92]. Therefore, in this chapter, Gentle AdaBoost [206] is utilised to build the prediction models. Gentle AdaBoost is displayed in Algorithm 7.1 below.

Algorithm 7.1: Gentle AdaBoost

Input:

S: Training set;

K: Iterations number;

Output:

$H_{(x)}$: Final hypothesis;

(1) Assign S sample $(x_1,y_1),..,(x_n,y_n)$; $x_i \in X, y_i \in \{-1,+1\}$;

(2) Initialise the weights of $D_1(i)=1/n$, $(i=1,\ldots,n)$;

(3) **For** $k=1$ **to** K **do**

(4) Call WeakLearn, providing it with the distribution D_k;

(5) Get weak hypothesis $h_k: X \rightarrow \{-1,+1\}$ with its error: $\varepsilon_k = \sum_{i=h_k(x_i)\neq y_i} D_k(i)$;

(6) Update distribution D_k: $D_{k+1}(i) = \frac{D_k(i)\exp(-\alpha_k y_k k_k(x_k))}{z_k}$;

(7) **End for**

(8) Output : $H_{(x)} = sign\left(\sum_{k=1}^{K} \alpha_k h_k(x)\right)$.

In Algorithm 7.1 above, S refers to a training set consisting of $(x_1,y_1),\ldots(x_n,y_n)$, where each x_i belongs to instance space X and each label y_i is in the label set Y, which is equal to the set of $\{-1,+1\}$. It assigns the weight on the training example i on round k as $D_k(i)$. The same weight is set at the starting point $(D_k(i)=1/N, i=1,\ldots,N)$. The weight of the misclassified example from the weak learner (called weak hypothesis) then increases to concentrate on hard-to-classify instances in the training set of each round. In step (6), Z_k is the normalisation constant (chosen so that D_{k+1} is a distribution), while, α_k is exer-

cised to improve the generalising result and also solves the overfitting and noise sensitive problems [183]. Therefore, α_k is defined as Equation 7.1.

$$\alpha_k = \frac{1}{2}\ln\left(\frac{W_{+1} - W_{-1}}{W_{-1} + W_{-1}}\right) \tag{7.1}$$

where W refers to the class probability estimate to construct the real value of $\alpha_k h_k(x)$. Therefore, the final hypothesis $H_{(x)}$ is a weighted majority vote of the K weak hypotheses in which it is the weight assigned to h_k. In addition, AdaBoost does handle both the binary class and the numerical class for prediction purposes [179]. Several research studies have successfully applied the AdaBoost algorithm to solve classification problems in object detection, including face recognition, video sequences and signal processing systems. For instance, Renno, Makris and Jones [162] employed an basic AdaBoost algorithm to develop AdaBoost classifiers using visual surveillance data. Their results demonstrated that the AdaBoost algorithm was suitable to build accurate classifiers. Similarly, Ho and Tay [207] also utilised AdaBoost to distinguish the spatially similar face and text in data sets. Unfortunately, their results showed that the basic AdaBoost lacked the performance of the model. This is due to the fact that they provided insufficient positive instances which resulted in a low hit rate (67.65% in fixed text and 13.89% in variable text).

7.2.2 Random Forests

Random Forests (RF) [208] is one of the most successful ensemble learning techniques in pattern recognition and machine learning [209] and is suited to imbalanced data classification problems [208]. It constructs a collection of individual decision tree classifiers utilising the Classification And Regression Tree (CART) algorithm [101]. CART is a rule-based method that generates a binary tree through a binary recursive partitioning

process that splits a node based on the 'yes' and 'no' answer of the predictors. CART makes use of the Gini Index to measure the impurity of a data partition in a training set [53]. Although this Gini Index is used to maximise the difference of heterogeneity, the real world data sets present an overfitting problem that causes the CART classifier to provide a high error rate in unseen data sets. In order to avoid this problem, Random Forests applies a bagging mechanism to increase the creation of classifiers in high dimensional data to reduce the error rate [208] [209]. The parameters of Random Forests involve the number of b trees and a random vector (S_b), using bootstrap samples generated independently from random vectors, but with the same distribution. The Random Forests algorithm is shown in Algorithm 7.2.

Algorithm 7.2: Random Forests

Input:
S: Training set;
f: Number of input instance;
B: Number of generated trees in Random Forests;
Output:
E: Classifier;

(1) E is empty;
(2) **For** b=1 **to** B **do**
(3) S_b = Bootstrap Sample(S);
(4) C_b = Build Random Tree Classifiers(S_b,f);
(5) E=$E \cup (C_b)$;
(6) **end for**
(7) Return E.

Many research studies have applied the Random Forests algorithm to construct a model. For instance, Kim, Lee and Park [210] extensively utilised this algorithm to build the lightweight Intrusion Detection Classifier. Their results showed that this classifier outperformed Support Vector Machines (SVM) and Artificial Neural Networks (ANN). However, this classifier is weak in high noise data which could cause an overfitting problem and reduce the accuracy of models in a test set. It also suffers from the overgrowth problem as an un-pruned tree [209].

7.2.3 Hybrid AdaBoost and Random Forests

The hybrid AdaBoost and Random Forests (ABRF) is used to generate accurate and reliable prediction models [33] due to the fact that it provides a low error rate and less overfitting problems in less outliers in the data set [33] [211]. This hybrid method is shown in Algorithm 7.3.

Algorithm 7.3: The hybrid AdaBoost and Random Forests

Input:

S: Training set;
K: Iterations number ;
f: Number of input instances to be used at each of the tree;
B: Number of generated trees in Random Forests;

Output:

$H_{(x)}$: Final hypothesis;

(1) Assign N sample $(x_1,y_1),..,(x_n,y_n)$; $x_i \in X, y_i \in \{-1,+1\}$
(2) Initialise the weights of $D_1(i)=1/n$, $i=1,\ldots,n$)
(3) **For** $k=1$ **to** K **do**
(4) Empty E with the distribution D_k ;
(5) **For** $b=1$ **to** B **do**
(6) S_b = booststrapSample(S);
(7) C_b = BuildRandomTreeClassifiers(S_b,f);
(8) $E=E \cup \{C_b\}$;
(9) **End for**
(10) Get weak hypothesis $h_k: X \rightarrow \{-1,+1\}$ with its error: $\varepsilon_k = \sum_{i=h_k(x_i)\neq y_i} D_k(i)$;
(11) Update distribution D_k: $D_{k+1}(i) = \frac{D_k(i)\exp(-\alpha_k y_k k_k(x_k))}{z_k}$;
(12) **End for**
(13) Output: $H_{(x)} = sign\left(\sum_{k=1}^{K} \alpha_k h_k(x)\right)$.

In the Algorithm 7.3 above, four input parameters need to be assigned: 1) S referring to training set ($S= x_i(i=1,2,\ldots,n)$, labels $y_i \in Y$); 2) K referring to the iterations number; 3) f referring to the number of input instances to be applied at each of the trees; and 4) B referring to the number of generated trees in Random Forests. It assigns the weight on the training example i on round k as $D_k(i)$. The same weight is set at the starting point ($D_k(i)=1/N$, $i=1,\ldots,N$). The weights of the misclassified instances from Random Forests

then increase to concentrate on hard-to-classify instances in the training set of each round. Therefore, the final hypothesis $H_{(x)}$ is a weighted majority vote of the K weak hypotheses in which it is the weight assigned to a collection of individual decision tree classifiers (h_k). In this way, the model generated from this hybrid method is able to achieve higher performance. Although this hybrid is effective, few research studies have utilised it. For instance, Leshem and Ritov [33] exploited the AdaBoost algorithm and Random Forests algorithm as the base learning algorithms to predict traffic flow. Their results pointed out that this combination had a low error rate which is the basic measurement method used to investigate a weak learner and the strong points of algorithms. Therefore, this method is proposed to build accurate and reliable breast cancer survivability models, due to the fact that it is an effective method and requires few input parameters.

7.3 Part-I: Prediction models of 5-year breast cancer survivability on the outlier-filtered data set

In order to build an accurate and reliable breast cancer survivability model, the 5-year breast cancer survivability data set is used. Accuracy sensitivity and specificity are employed in order to evaluate the proposed method. This section concludes with the experimental results and discussion.

7.3.1 Data of 5-year breast cancer survivability

In this section, the filtered 5-year breast cancer survivability data set from 1990-2001 is utilised. This data set consists of 12 attributes and 570 instances. Attributes are listed in Table 7.1 below.

Table 7.1: Input attributes of breast cancer data

No.	Attributes	Types of attribute	Attribute Values
1	Age	Number	-
2	Marital status	Category	3
3	Occupation	Category	26
4	Basis of diagnosis	Category	6
5	Topography	Category	9
6	Morphology	Category	14
7	Extent	Category	4
8	Stage	Category	4
9	Received surgery therapy	Category	2
10	Received radiation	Category	2
11	Received chemotherapy	Category	2
12	Survivability status (class attribute)	Category	2

Table 7.1 shows 12 attributes including 'Age', 'Marital status', 'Occupation', 'Basis of diagnosis', 'Topography', 'Morphology', 'Extent', 'Stage', 'Received surgery therapy', 'Received radiation', 'Received chemotherapy' and 'Survivability status'. The 'Survivability status' (class attribute) is divided into two classes including 'Dead' and 'Alive'. The 'Dead' class refers to patients who died within five years after the first diagnosis, while the 'Alive' class refers to patients who are still alive for five years or more after the first diagnosis. This 'Dead' class is composed of 322 instances, while the 'Alive' class comprises 248 instances.

7.3.2 Methods for evaluating classifiers

In order to evaluate the proposed approach, accuracy, sensitivity and specificity (see Section 2.2.3.1) are employed. *Accuracy* refers to the percentage of correctness of outcome among the test sets of prediction results of a classifier. *Sensitivity* refers to the true positive rate of prediction results, while, *specificity* refers to the true negative rate of prediction results. Both sensitivity and specificity are used for measuring the factors that affect the performance of classifiers in a binary classification problem. The evaluation procedure is performed using a stratified 10-fold cross-validation to divide a 5-year

breast cancer survivability data set from Srinagarind Hospital in Thailand into training and test sets. The training set is used to build a model and the test set is used to evaluate the model.

7.3.3 Results of classifiers

In this section, the WEKA version 3.5.6 [92] explorer and knowledge flow are used to evaluate the performance and effectiveness of the proposed classifiers (ABRF), and compared with other classifiers. This is due to the fact that they provide a well defined framework and offer a variety of learning algorithms for the development of new data mining and machine learning algorithms. Default parameters used in these experiments are given first. Then the results of the experiments are presented and discussed.

7.3.3.1 Default parameters

In order to compare and establish the performance and effectiveness of prediction models, the default parameters of the algorithms are defined as follows:

1) AdaBoost with Random Forests (ABRF) applies the unlimited depth of the trees, 10 trees, 10 iterations, 100 weight thresholds and one random seed;
2) AdaBoost uses Decision Stump as a based learner, 10 iterations, 100 weight thresholds and one random seed;
3) A Alternative Decision Tree (ADTree) uses 10 iterations and an exhaustive search;
4) Bagging uses a Fast decision tree as a weak learner, 100% size of each bag of the training set size, 10 iterations and one random seed;
5) C4.5 uses 0.25 confidence factor for pruning, two numbers of instances per leaf, three amounts of data used for reduced-error pruning and one random seed;

6) A Conjunctive Rule utilises three folds used for pruning, the rest for growing the rules, two minimum total weights of the instances in a rule, the number of antecedents allowed in the rule and one random seed;
7) Naïve Bayes does not apply the output additional information to the console, but applies a kernel estimator for numeric attributes to convert these numeric attributes into discrete attributes;
8) Nearest-Neighbour classifier (NN-classifier) uses normalised Euclidean distance to find the training instance closest to the given test instance, and predicts the same class as training instances;
9) Random Forests uses 10 trees and one random seed;
10) Repeated Incremental Pruning to Produce Error Reduction (RIPPER) uses an error rate more than 0.5 for stopping criterion, three folds pruning and the rest for growing the rules, two minimum total weight of the instances in a rule, two numbers of optimisation runs and one random seed; and
11) Support Vector Machine uses C-support vector classification, 40 cache size, three degrees of the kernel, 0.0010 tolerances of the termination criterion, radial basis kernel function and 0.1 for the loss function.

7.3.3.2 Performance of classifiers

In order to evaluate the performance and effectiveness of the hybrid AdaBoost and Random Forests (ABRF), accuracy, sensitivity and specificity are employed and compared with 10 single classifiers including AdaBoost, ADTree, Bagging, C4.5, Conjunctive Rules, Naïve Bayes (NB), Nearest Neighbour classifier (NN-classifier), Random Forests (RF), Repeated Incremental Pruning to Produce Error Reduction (RIPPER) and Support Vector Machine (SVM). The experimental results tested in both training and

test sets are displayed in Table 7.2.

Table 7.2: Performance of single classifier on the training and test sets

Classifiers	Training Set (%)			Test Set (%)		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
ABRF	*100.00*	*100.00*	*100.00*	*88.60*	*89.30*	*87.65*
AdaBoost	80.88	78.55	85.28	80.35	77.93	85.05
ADTree	85.09	85.59	84.39	82.28	83.59	80.50
Bagging	91.23	92.24	89.92	83.86	84.64	82.77
C4.5	92.46	93.19	91.50	84.04	87.38	80.08
Conjunctive Rule	77.54	74.74	83.71	77.54	74.74	83.71
Naïve Bayes	84.04	85.54	82.04	83.51	84.97	81.56
NN-classifier	100.00	100.00	100.00	83.86	85.49	81.71
Random Forests	99.65	99.69	99.60	85.79	86.63	84.65
RIPPER	87.54	91.15	83.40	85.79	88.25	82.75
SVM	99.82	99.69	100.00	85.96	86.45	85.29

Table 7.2 shows the accuracy, sensitivities and specificities of classifiers including ABRF, AdaBoost, ADTree, Bagging, C4.5, Conjunctive Rule, Naïve Bayes, NN-classifier, Random Forests, RIPPER and SVM. The results of this experiment show that the accuracy of ABRF achieves 100% when utilising the training set and achieves 88.60% when using the test set. Besides, the average accuracy, sensitivity and specificity of ABRF increases 8.25%, 11.37% and 2.6% respectively, based on the basic AdaBoost using test sets. Likewise, the average accuracy, sensitivity and specificity of ABRF increases up to 2.81%, 2.67% and 3% respectively, based on Random Forests. This proves that ABRF provides a better approximation of the prediction than models made by basic AdaBoost and Random Forests.

7.3.3.3 Performance of AdaBoost with weak learners

In order to evaluate the performance and effectiveness of the hybrid AdaBoost and Random Forests model, accuracy, sensitivity and specificity are also employed and compared with eight AdaBoost and weak learners including ADTree, C4.5, Conjunctive Rule, Decision Stump, Naïve Bayes, NN-classifier, RIPPER and SVM. The default parameters (see Section 7.3.3.1) of each weak learner are used to generate models. The

experimental results provide an insight into the increasing iterations of AdaBoost from 5 to 100 iterations. The results are shown in Figures 7.1, 7.2 and 7.3 below.

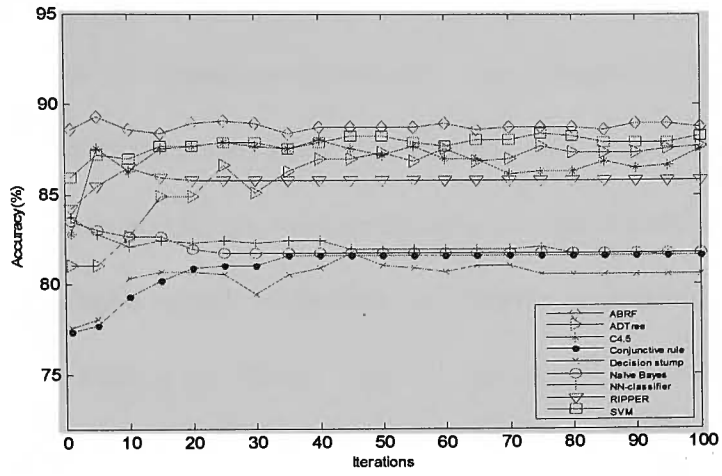


Figure 7.1: Accuracy comparisons

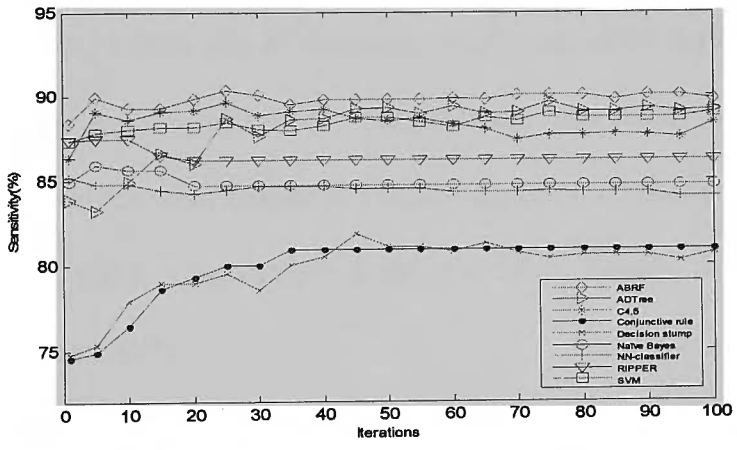


Figure 7.2: Sensitivity comparisons

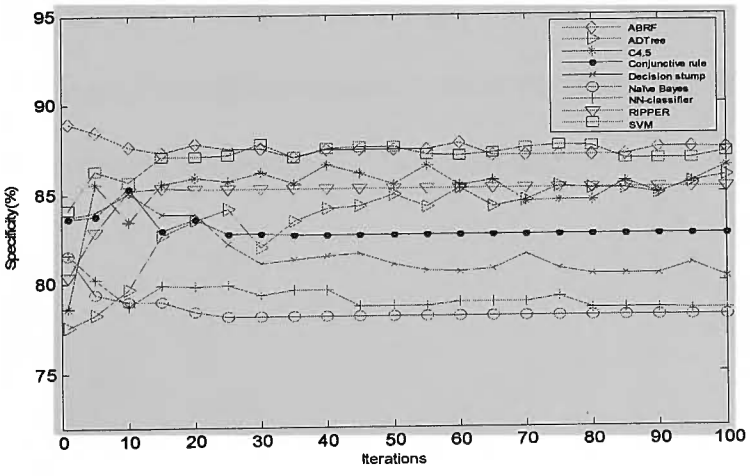


Figure 7.3: Specificity comparisons

Figures 7.1, 7.2 and 7.3 show the accuracy, sensitivity and specificity of AdaBoost with weak learners. In relation to the accuracy of ensemble classifiers, it seems that the accuracy of ensemble classifiers including ADTree, C4.5, Conjunctive Rule and Decision Stump, increases after 10 iterations of re-boost. In contrast, the accuracy of Naïve Bayes and NN-classifier decreases after five rounds. Although the sensitivity of ensemble classifiers seems stable, the specificity of ensemble classifiers seems uncertain in the results. It might be related to the fact that AdaBoost concentrates on improving the majority class, which is the 'Dead' class in this case. However, the accuracy of these prediction models is mostly stable after running for 45 iterations. The results show that the average accuracy, sensitivity and specificity of ABRF prediction models increases up to 7.85%, 9.93% and 4.54% respectively, based on the basic AdaBoost in the same test set after running 100 times. This indicates that ABRF is not only superior to basic AdaBoost and Random Forests but is also superior to other ensemble classifiers including ADTree, C4.5, Conjunctive Rule, Decision Stump, Naïve Bayes, NN-classifier, RIPPER and SVM.

7.3.3.4 Statistical analysis of multiple classifiers

In order to investigate the stability of prediction models, statistics analysis, including the minimum, maximum, average and variance of the accuracy, sensitivity and specificity of the proposed method is employed. The analysis results are illustrated in Tables 7.3, 7.4 and 7.5.

Table 7.3: Statistics of accuracy of ensemble classifiers on test sets

Classifiers	Minimum	Maximum	Average	Variance
ABRF	88.42	89.30	88.79	**0.05**
AD Tree	81.05	87.72	86.07	4.34
C4.5	82.81	88.07	86.95	1.25
Conjunctive Rule	77.37	81.58	80.94	1.61
Decision Stump	77.54	81.75	80.40	0.92
Naïve Bayes	81.75	83.51	81.98	0.25
NN-classifier	81.58	83.68	82.15	0.23
RIPPER	84.21	86.49	85.74	0.15
SVM	85.96	88.42	87.79	0.29

Table 7.4: Statistics of sensitivity of ensemble classifiers on test sets

Classifiers	Minimum	Maximum	Average	Variance
ABRF	88.36	90.37	89.79	0.18
AD Tree	83.23	89.62	88.02	3.67
C4.5	86.36	89.66	88.35	0.64
Conjunctive Rule	74.55	80.91	79.85	4.08
Decision Stump	74.74	81.87	79.74	3.38
Naïve Bayes	84.71	85.94	84.87	**0.13**
NN-classifier	84.01	85.23	84.47	0.08
RIPPER	86.19	87.46	86.38	0.20
SVM	87.35	89.02	88.40	0.18

Table 7.5: Statistics of specificity of ensemble classifiers on test sets

Classifiers	Minimum	Maximum	Average	Variance
ABRF	*86.99*	*88.94*	*87.48*	**0.22**
AD Tree	77.56	85.89	83.56	0.51
C4.5	78.63	86.59	85.15	2.80
Conjunctive Rule	82.65	85.33	82.94	0.43
Decision Stump	80.18	85.05	81.69	2.15
Naïve Bayes	78.13	81.56	78.44	0.64
NN-classifier	78.49	81.63	79.19	0.60
RIPPER	80.38	85.29	84.89	1.33
SVM	84.15	87.76	86.98	0.65

Tables 7.3, 7.4 and 7.5 show the minimum, maximum, average and variance of the accuracy, sensitivity and specificity of the ensemble classifiers. The results show that ABRF increases 8.39% of the approximation of the prediction when compared to the basic AdaBoost in the same test set after running the test 100 times. Furthermore, the variance of accuracy and specificity of ABRF being the lowest indicates that ABRF is

keeping stable even after increasing the iteration. Therefore, the computation cost can be reduced by applying a few iterations. This proves that ABRF provides a lower error rate and variance than made by other ensemble classifiers including ADTree, C4.5, Decision Stump, Conjunctive Rule, Naïve Bayes, NN-classifier, RIPPER and SVM.

7.3.3.5 Discussion of a 5-year breast cancer survivability prediction model

A 5-year breast cancer survivability prediction model has been developed from Srinagarind Hospital's cancer data sets from 1990 to 2001. These data have similar behaviour in survival proportions analysis (See detail in 3.7). The performance and effectiveness of this model generated from the AdaBoost with Random Forests are presented in terms of accuracy, sensitivity and specificity. In order to point out the significance of the results, several findings are discussed.

Firstly, ABRF provides more accuracy of prediction models than the models made by ADTree, Bagging, C4.5, Conjunctive Rule, Naïve Bayes, NN-classifier, RIPPER and SVM in both training and test sets. This may be due to the ability to select the important instance of the AdaBoost algorithm to reduce the error rates [172] [188]. Similarly, Leshem and Ritov [33] demonstrated that ABRF is superior to AdaBoost with Decision Stump in terms of error rates by evaluating using both training and test sets.

Secondly, the experimental results indicate that most hybrid classifiers including ADTree, C4.5, Conjunctive Rule and Decision Stump improve their accuracy after 10 rounds of re-boost. In contrast, Naïve Bayes and NN-classifier decrease their accuracy after 10 rounds. However, Leshem and Ritov's [33] work indicated that the error rates of AdaBoost with Random Forests decreased after 80 rounds in traffic flow data.

Finally, the results of the present study agree with Bartlett and Traskin's [202] results that AdaBoost is only consistent when applying the suitable iterations. As a result, the

experiment results in Part-I concluded that the accuracy of AdaBoost is found to be stable after running the algorithm for 35 iterations.

7.4 Part-II: Breast cancer survivability prediction models on outliers-filtered and balanced data sets

Since breast cancer is the second most frequent cause of cancer incidence among women in Thailand [2], building 3-, 5-, 8- and 10-year breast cancer survivability prediction models is helpful in increasing the provision of care by medical practitioners. In order to evaluate the performance and effectiveness of prediction models, four evaluation methods including accuracy, Area Under the receiver operating characteristic Curve (AUC), *F*-measure and Kappa statistics are employed, followed by the experimental results and discussion.

7.4.1 Data sets of 3-, 5-, 8- and 10-year breast cancer survivability

The breast cancer survivability data sets consist of 14 attributes obtained at Srinagarind Hospital in Thailand from 1985-2006. These attributes are listed in Table 7.6.

Table 7.6: The list of attributes

No.	Attributes	Attribute Types
1	Age	Number
2	Marital status	Category
3	Basis of diagnosis	Category
4	Topography	Category
5	Morphology	Category
6	Extent	Category
7	Stage	Category
8	Received surgery	Category
9	Received radiation	Category
10	Received chemotherapy	Category
11	Received hormonal therapy	Category
12	Received supportive therapy	Category
13	Received other therapy	Category
14	Survivability status (class attribute)	Category

Table 7.6 presents 14 attributes including 'Age', 'Marital status', 'Basis of diagnosis', 'Morphology', 'Extent', 'Stage', 'Received surgery', 'Received radiation', 'Received chemotherapy', 'Received hormonal therapy', 'Received supportive therapy', 'Received other therapy' and 'Survivability status'. In order to build prediction models, 3-, 5-, 8- and 10-year survival periods of patients surviving from breast cancer are used. After that the combination of Outlier filtering and Over-Sampling (OOS) are utilised to improve the data quality. Hence, the number of instances corresponding to 3-, 5-, 8- and 10-year breast cancer survivability are tabulated in Table 7.7.

Table 7.7: The number of instances in data sets

Data Sets	Years	'Dead'	'Alive'	Total
3-year	1985-2004	542	544	1086
5-year	1985-2002	368	367	735
8-year	1985-1999	265	264	529
10-year	1985-1997	212	211	423

Table 7.7 displays the number of instances of 3-, 5-, 8- and 10-year breast cancer survivability data sets which involve two classes including 'Dead' ('0') and 'Alive' ('1'). For instance, 3-year breast cancer survivability data from 1985 to 2004 consist of 'Dead' and 'Alive' classes. The 'Dead' class refers to patients who died within three years after the first diagnosis, while the 'Alive' class refers to patients who are still alive for three years or more after the first diagnosis. Similarly, the 'Dead' class in 5-, 8- and 10-year breast cancer survivability data refers to patients who died within five, eight and 10 years after the first diagnosis respectively, and the 'Alive' class refers to patients who are still alive for five or more, eight or more and 10 years or more after the first diagnosis.

7.4.2 Performance evaluation methods

In order to evaluate the performance of prediction models, four evaluation methods including accuracy, Area Under the receiver operating characteristic Curve (AUC), *F*-measure, and Kappa statistics are utilised. *A*ccuracy (see Section 2.2.3.1) of classifiers presents the basic performance as the percentage of correctness of outcome among the test sets. *AUC* (see Section 2.2.3.3) is an area under the ROC curve which is the relation between true positive and false negative rates [80]. *F-measure* (see Section 2.2.3.4) is used to measure the performance of prediction models based on recall and precision. *Kappa statistics* (see Section 2.2.3.5) are used to evaluate the intra-class correlation coefficients based on the individual members of the population [86] [88] [87]. These experiments are performed using a stratified 10-fold cross-validation to divide 3-, 5-, 8- and 10-year breast cancer survivability data sets into training and test sets. In addition, the statistic t-test [212] with 0.05 statistical significance improvement is used to illustrate the significant difference of the performance of prediction models generated from AdaBoost with Random Forests (ABRF) to the other classifiers using an asterisk (*).

7.4.3 Experimental evaluation results

WEKA version 3.5.6 [92] experimental is used to evaluate the performance and effectiveness of prediction models using a combination of AdaBoost and Random technique. This combination technique is proposed to generate breast cancer survivability prediction models prior to comparing it with 19 classifiers including: Bayes Network, Naïve Bayes; Logistic regression, Radial Basis Function Network (RBFNetwork), Sequential Minimal Optimisation (SMO), AdaBoost with Decision Stump, Bagging; Random Committee, Random Sub-Space, Alternative Decision Tree (ADTree), Best-First deci-

sion tree (BFTree), C4.5, Naive Bayes Tree (NBTree), Random Forests, Fast decision tree (REPTree), Conjunctive Rule, Decision table, Repeated Incremental Pruning to Produce Error Reduction (RIPPER) and a PART decision list. Default parameters used in these experiments are given. Then results of these experiments are presented and discussed.

7.4.3.1 Parameters setting

In order to demonstrate the performance and effectiveness of the proposed approach, the default parameters of each algorithm are defined as follows:

1) AdaBoost with Random Forests (ABRF) applies: the unlimited depth of the trees, 5 trees to be generated, 35 iterations, 100 weight thresholds and one random seed.
2) A Bayes Network uses: the simple estimator; and a *k2* for search purposed.
3) Naïve Bayes does not apply the output additional information to the console, but uses a kernel estimator for numeric attributes to convert into nominal ones.
4) Logistic regression uses: a maximum number of iterations to perform; and the 1.0E -8 of ridge value in the log-likelihood.
5) Radial Basis Function Network (RBFNetwork) applies: one random seed to pass on to *k*-means; sets the maximum number of iterations for the logistic regression to perform to one into discrete class problems; sets the minimum standard deviation for the clusters to -1; and sets the number of clusters for *k*-means to two to generate and set the Ridge value for the logistic or linear regression to 1.0E -8.
6) Sequential Minimal Optimisation (SMO) utilises: one the complexity parameter C; 1.0E -12 for the epsilon to round off an error; normalised training data for fil-

tering; Poly kernel function; cross-validation to generate training data; one random seed; and 0.0010 for the tolerance parameter.

7) AdaBoost uses: Decision Stump as a weak learner; 10 iterations, 100 weight thresholds; and one random seed.
8) Bagging uses: a Fast decision tree (REPTree) as a weak learner; 100% size of each bag of the training set size; 10 iterations; and one random seed.
9) A Random Committee uses: Random Tree as a weak learner; 10 iterations; and one random seed.
10) Random Sub-Space uses: Fast decision tree (REPTree) as a weak learner; 10 iterations; one random seed; and 0.5% of the number of attributes in subspace.
11) An Alternative Decision Tree (ADTree) uses 10 iterations and exhaustive search.
12) A Best-First decision Tree (BFTree) uses: heuristic search for a binary split for nominal attributes; two minimal number of instances at the terminal nodes; five folds in internal cross-validation; post pruning to prune the tree; one random seed; a error rate as an error estimate; and the Gini Index for splitting criterion.
13) C4.5 uses: 0.25 confidence factor for pruning; two numbers of instances per leaf; three amounts of data used for reduced-error pruning; and one random seed.
14) A Naive Bayes Tree (NBTree) uses Naive Bayes classifiers to generate a decision tree at the first leaves.
15) Random Forests uses 10 trees and one random seed.
16) A fast decision tree (REPTree) uses: no restriction for the maximum depth tree; two minimum total weights of the instances in a leaf; 0.0010 minimum propor-

tion of the variance on all the data using three folds of regression trees; and one random seed.

17) A Conjunctive Rule uses: three folds for pruning; the rest for growing the rules; two minimum total weights of the instances in a rule; the number of antecedents allowed in the rule; and one random seed.

18) A Decision table uses: leave one out data selection; accuracy to evaluate the performance of attributes combinations; and best first search.

19) Repeated Incremental Pruning to Produce Error Reduction (RIPPER) uses: 0.5 error rates for stopping criterion; three folds pruning and the rest for growing the rules; two minimum total weights of the instances in a rule; two numbers of optimisation runs; and one random seed.

20) A PART decision list uses: 0.25 confidence factor used for pruning; two minimum numbers of instances per rule; three folds for pruning and the rest for growing the rules; and one random seed.

7.4.3.2 Accuracy classifications

In order to evaluate the performance and effectiveness of the prediction model, in this section the accuracy is utilised and compared with 19 classifiers. The results of this experiment are presented in Table 7.8.

Table 7.8: Accuracy of classifiers

Classifiers	3-year	5-year	8-year	10-year	Average
ABRF	96.96 (±1.38)	96.60(±2.14)	98.29(±1.90)	98.35(±1.60)	**97.55**
Random Forests	96.13 (±1.44)	96.60 (±2.32)	98.48 (±1.97)	98.35 (±1.60)	97.39 ↓
BFTree	96.59 (±1.23)	95.11 (±2.31)	97.35 (±2.05)	99.53 (±0.99)	97.15 ↓
Random Committee	96.50 (±0.95)	95.52 (±3.12)	98.29 (±1.90)	98.12 (±1.86)	97.11 ↓
C4.5	95.58 (±1.48)	94.01 (±2.81) *	97.16 (±2.05)	99.06 (±1.22)	96.45 ↓
Bagging	95.76 (±1.80)	94.69 (±2.25) *	96.97 (±2.23)	97.88 (±2.05)	96.33 ↓
PART	94.76 (±2.02)	93.19 (±2.50) *	97.91 (±2.10)	99.30 (±1.13)	96.29 ↓
RIPPER	94.02 (±1.75) *	93.60 (±3.39)	98.11 (±1.55)	98.58 (±1.66)	96.08 ↓
REPTree	94.10 (±2.10) *	93.47 (±4.02)	96.97 (±2.99)	97.88 (±2.05)	95.61 ↓
NBTree	94.11 (±1.64) *	93.33 (±2.51) *	95.84 (±2.49)	97.64 (±1.93)	95.23 ↓
Random Sub-Space	93.83 (±2.65) *	92.52 (±3.06) *	96.22 (±4.45)	95.51 (±4.67)	94.52 ↓
ADTree	90.42 (±3.97) *	89.39 (±2.37) *	96.97 (±1.84)	99.53 (±0.99)	94.08 ↓
Decision Table	89.32 (±2.31) *	88.70 (±2.23) *	98.86 (±1.83)	99.30 (±1.57)	94.05 ↓
Bayes Network	89.32 (±3.37) *	88.57(±3.28) *	97.35(±1.84)	97.88(±2.05)	93.28 ↓
Logistic	85.09 (±3.60) *	89.39(±4.72) *	95.28(±3.36)	96.23(±3.15)	91.50 ↓
AdaBoost	88.67 (±2.90) *	87.20 (±3.67) *	92.24 (±3.96) *	95.51 (±4.11)	90.91 ↓
RBFNetwork	87.20 (±3.61) *	87.88(±3.28) *	92.24(±4.36) *	93.18(±4.94) *	90.13 ↓
SMO	84.62 (±2.79) *	87.34(±4.22) *	93.19(±3.82) *	93.16(±4.20) *	89.58 ↓
Naïve Bayes	82.14 (±3.25) *	87.89(±2.91) *	91.30(±3.83) *	91.04(±4.64) *	88.09 ↓
Conjunctive Rule	76.98 (±4.31) *	82.99 (±3.06) *	85.99 (±7.67) *	79.94 (±8.63) *	81.48 ↓

Table 7.8 provides the accuracy of 20 classifiers using 3-, 5-, 8- and 10-year breast cancer survivability data sets. The experimental results show that the accuracy of ABRF classifier achieves 96.96%, 96.60%, 98.29% and 98.35% based on 3-, 5-, 8- and 10-year of breast cancer survivability prediction models, respectively. It is found that the accuracy of the proposed classifier (ABRF) is significantly different from the accuracy of 13 classifiers in 3-year breast cancer survivability prediction models while the accuracy of ABRF is significantly different from the accuracy of 14 classifiers in 5-year breast cancer survivability prediction models. Although the accuracy of ABRF is only significantly different from the accuracy of five classifiers, its accuracy is better than the accuracy of 10 classifiers in 8-year breast cancer survivability prediction models. Even though the accuracy of ABRF classifier is only significantly different from the accuracy of four classifiers, the accuracy of ABRF classifier is better than the accuracy of eight classifiers in 10-year breast cancer survivability prediction models. This may be due to the fact that the number of instances in 8- and 10-year data sets is too small. However,

average accuracy of ABRF achieves 97.55% and provides better average accuracy than 19 classifiers. This indicates that the ABRF classifier is superior to the other 19 classifiers.

7.4.3.3 Area under the receiver operating characteristic curve classifications

In order to evaluate the performance and effectiveness of prediction models, Area Under the receiver operating characteristic Curve (AUC) is used. The experimental results are displayed in Table 7.9.

Table 7.9: AUC scores of classifiers

Classifiers	3-year	5-year	8-year	10-year	Average	
ABRF	99.59(±0.00)	99.09(±0.01)	99.81(±0.01)	100.00(±0.00)	**99.62**	
Random Committee	99.34(±0.01)	98.84(±0.01)	99.81(±0.01)	100.00(±0.00)	99.50	↓
Random Forests	99.44(±0.01)	98.65(±0.01)	99.81(±0.01)	100.00(±0.00)	99.48	↓
Bagging	98.37(±0.01) *	96.21(±0.03) *	99.24(±0.01)	99.30(±0.01)	98.28	↓
NBTree	97.62(±0.01) *	97.39(±0.01) *	98.63(±0.02)	99.49(±0.01)	98.28	↓
Random Sub-Space	97.38(±0.02) *	96.38(±0.03)	99.30(±0.01)	99.86(±0.00)	98.23	↓
ADTree	96.88(±0.02) *	96.43(±0.02) *	99.07(±0.01)	99.47(±0.01)	97.96	↓
BFTree	97.39(±0.01) *	96.15(±0.02) *	97.93(±0.02)	99.52(±0.01)	97.75	↓
PART	96.52(±0.02) *	95.98(±0.03) *	98.04(±0.03)	99.42(±0.01)	97.49	↓
C4.5	96.84(±0.01) *	94.43(±0.03) *	98.27(±0.02)	98.77(±0.02)	97.08	↓
Decision Table	94.22(±0.03) *	94.72(±0.03) *	99.41(±0.01)	99.23(±0.02)	96.90	↓
REPTree	96.73(±0.01) *	94.36(±0.03) *	97.42(±0.03)	98.57(±0.02)	96.77	↓
RIPPER	94.87(±0.02) *	94.54(±0.04) *	98.57(±0.02)	99.00(±0.01)	96.75	↓
Bayes Network	94.76(±0.01) *	93.05(±0.03) *	99.04(±0.01)	99.20(±0.01)	96.51	↓
AdaBoost	95.40(±0.01) *	93.15(±0.02) *	96.35(±0.02) *	98.97(±0.02) *	95.97	↓
Naïve Bayes	93.88(±0.02) *	92.62(±0.03) *	95.8(±0.03) *	97.49(±0.03) *	94.95	↓
RBFNetwork	94.25(±0.02) *	93.32(±0.04) *	95.18(±0.02) *	96.53(±0.04) *	94.82	↓
Logistic	92.18(±0.02) *	92.34(±0.05) *	95.99(±0.03) *	96.15(±0.03) *	94.17	↓
SMO	84.63(±0.03) *	87.32(±0.04) *	93.20(±0.04) *	93.20(±0.04) *	89.59	↓
Conjunctive Rule	76.98(±0.04) *	82.97(±0.03) *	86.02(±0.08) *	79.87(±0.09) *	81.46	↓

Table 7.9 illustrates the AUC scores of 20 classifiers using 3-, 5-, 8- and 10-year breast cancer survivability data sets. The experimental results present that the AUC scores of ABRF classifier achieve 99.59%, 99.09%, 99.81% and 100% based on 3-, 5-, 8- and 10-year of breast cancer survivability prediction models, respectively. It is found that the AUC scores of the proposed classifier (ABRF) are significantly different from the AUC scores of 17 classifiers based on 3- and 5-year breast cancer survivability prediction models while the AUC scores of ABRF are significantly different from the AUC scores

of six classifiers based on 8- and 10-year breast cancer survivability prediction models. Besides, the AUC scores of ABRF are significantly better than the AUC score of basic AdaBoost on 3-, 5-, 8- and 10-year breast cancer survivability data sets and also better than other classifiers on the average. This shows that the ABRF classifier provides better AUC scores than those that made by the other 19 classifiers.

7.4.3.4 *F*-measure classifications

In order to evaluate a trade-off between precision and recall of prediction models, *F*-measure is employed. The results of this experiment are shown in Table 7.10.

Table 7.10: *F*-measure of classifiers

Classifiers	3-year	5-year	8-year	10-year	Average	
ABRF	97.02(±0.01)	96.52(±0.02)	98.24(±0.02)	98.30(±0.02)	**97.52**	
Random Forests	96.22(±0.01)	96.50(±0.02)	98.44(±0.02)	98.30(±0.02)	97.37	↓
BFTree	96.62(±0.01)	95.00(±0.02)	97.25(±0.02)	99.51(±0.01)	97.10	↓
Random Committee	96.55(±0.01)	95.43(±0.03)	98.23(±0.02)	98.06(±0.02)	97.07	↓
C4.5	95.64(±0.01)	93.92(±0.03) *	97.05(±0.02)	99.04(±0.01)	96.41	↓
Bagging	95.85(±0.02)	94.62(±0.02) *	96.84(±0.02)	97.82(±0.02)	96.28	↓
PART	94.73(±0.02)	93.12(±0.03) *	97.84(±0.02)	99.28(±0.01)	96.24	↓
RIPPER	94.22(±0.02) *	93.38(±0.04)	98.06(±0.02)	98.54(±0.02)	96.05	↓
REPTree	94.24(±0.02) *	93.39(±0.04)	96.88(±0.03)	97.84(±0.02)	95.59	↓
NBTree	94.20(±0.02) *	93.41(±0.02) *	95.61(±0.03)	97.55(±0.02)	95.19	↓
Random Sub-Space	94.04(±0.03) *	92.63(±0.03) *	95.91(±0.05)	95.10(±0.05)	94.42	↓
ADTree	90.41(±0.04) *	89.74(±0.02) *	96.86(±0.02)	99.51(±0.01)	94.13	↓
Decision Table	89.77(±0.02) *	88.64(±0.02) *	98.83(±0.02)	99.28(±0.02)	94.13	↓
Bayes Network	89.51(±0.03) *	89.18(±0.03) *	97.26(±0.02)	97.82(±0.02)	93.44	↓
Logistic	85.60(±0.03) *	89.62(±0.05) *	94.97(±0.04)	96.02(±0.03)	91.55	↓
AdaBoost	88.38(±0.03) *	88.29(±0.03) *	91.48(±0.05) *	95.25(±0.05) *	90.85	↓
RBFNetwork	87.47(±0.04) *	88.44(±0.03) *	91.78(±0.05) *	92.76(±0.05) *	90.11	↓
SMO	85.36(±0.03) *	88.08(±0.04) *	92.54(±0.04) *	92.51(±0.05) *	89.62	↓
Naïve Bayes	81.10(±0.04) *	88.66(±0.03) *	90.34(±0.05) *	90.29(±0.05) *	87.60	↓
Conjunctive Rule	78.48(±0.04) *	85.05(±0.03) *	85.67(±0.08) *	77.31(±0.13) *	81.63	↓

Table 7.10 displays the *F*-measure of 20 classifiers using 3-, 5-, 8- and 10-year breast cancer survivability data sets. The experimental results illustrate that the *F*-measure of ABRF classifier achieves 97.02%, 96.52%, 98.24% and 98.30%, based on 3-, 5-, 8- and 10-year of breast cancer survivability prediction models, respectively. It is demonstrated that the *F*-measure of ABRF classifier is significantly different from the *F*-measure of 14 classifiers in both 3- and 5-year breast cancer survivability prediction

models while the *F*-measure of ABRF classifier is significantly different from the *F*-measure of five classifiers in both 8- and 10-year breast cancer survivability prediction models. Moreover, the *F*-measure of ABRF classifier is better than the *F*-measure of 12 and seven classifiers based on 8- and 10-year breast cancer survivability prediction models, respectively. Furthermore, the average of the *F*-measure of ABRF classifier achieves 97.52% and is better than the average of the *F*-measure of 19 classifiers. This demonstrates that the prediction models generated from the hybrid AdaBoost and Random Forests technique are better than ones made by the other techniques.

7.4.3.5 Kappa statistics classifications

In this section, Kappa statistics are used to evaluate the performance of prediction models using 3-, 5-, 8- and 10-year breast cancer survivability data sets. The experimental results are presented in Table 7.11.

Table 7.11: Kappa statistics of classifiers

Classifiers	3-year	5-year	8-year	10-year	Average	
ABRF	93.92(±0.03)	93.21(±0.04)	96.58(±0.04)	96.70(±0.03)	**95.10**	
Random Forests	92.26(±0.03)	93.21(±0.05)	96.96(±0.04)	96.70(±0.03)	94.78	↓
BFTree	93.19(±0.02)	90.21(±0.05)	94.70(±0.04)	99.06(±0.02)	94.29	↓
Random Committee	93.00(±0.02)	91.03(±0.06)	96.58(±0.04)	96.24(±0.04)	94.21	↓
C4.5	91.17(±0.03)	88.02(±0.06) *	94.32(±0.04)	98.12(±0.02)	92.91	↓
Bagging	91.52(±0.04)	89.39(±0.05) *	93.94(±0.04)	95.77(±0.04)	92.66	↓
PART	89.51(±0.04)	86.38(±0.05) *	95.83(±0.04)	98.59(±0.02)	92.58	↓
RIPPER	88.03(±0.03) *	87.20(±0.07)	96.21(±0.03)	97.16(±0.03)	92.15	↓
REPTree	88.21(±0.04) *	86.95(±0.08)	93.94(±0.06)	95.77(±0.04)	91.22	↓
NBTree	88.21(±0.03) *	86.67(±0.05) *	91.68(±0.05) *	95.27(±0.04) *	90.46	↓
Random Sub-Space	87.66(±0.05) *	85.04(±0.06) *	92.45(±0.09)	91.04(±0.09)	89.05	↓
ADTree	80.84(±0.08) *	78.77(±0.05) *	93.94(±0.04)	99.06(±0.02)	88.15	↓
Decision Table	78.63(±0.05) *	77.41(±0.04) *	97.72(±0.04)	98.61(±0.03)	88.09	↓
Bayes Network	78.64(±0.07) *	77.13(±0.07) *	94.70(±0.04)	95.77(±0.04)	86.56	↓
Logistic	70.17(±0.07) *	78.79(±0.09) *	90.56(±0.07)	92.48(±0.06)	83.00	↓
AdaBoost M1	77.34(±0.06) *	74.39(±0.07) *	84.47(±0.08) *	91.02(±0.08) *	81.81	↓
RBFNetwork	74.40(±0.07) *	75.74(±0.07) *	84.48(±0.09) *	86.38(±0.10) *	80.25	↓
SMO	69.25(±0.06) *	74.67(±0.08) *	86.37(±0.08) *	86.34(±0.08) *	79.16	↓
Naïve Bayes	64.26(±0.06) *	75.77(±0.06) *	82.59(±0.08) *	82.09(±0.09) *	76.18	↓
Conjunctive Rule	53.94(±0.09) *	65.96(±0.06) *	72.00(±0.15) *	59.79(±0.17) *	62.92	↓

Table 7.11 illustrates the Kappa statistics of 20 classifiers using 3-, 5-, 8- and 10-year breast cancer survivability data sets. The experimental results show that the Kappa statistics of the ABRF classifier achieve 93.92%, 93.21%, 96.58% and 96.70% in 3-, 5-, 8- and 10-year of breast cancer survivability prediction models, respectively. Besides, the Kappa statistics of the ABRF classifier are significantly different from 13 classifiers using a 3-year breast cancer survivability data set while significantly different from 15 classifiers using a 5-year breast cancer survivability data set. Although the Kappa statistics of the ABRF classifier are significantly different from six classifiers, it is better than 10 and six classifiers, using 8- and 10-year breast cancer survivability data sets. Therefore, using the ABRF can increase the Kappa statistics of basic AdaBoost up to 16.58%, 18.82%, 12.11% and 5.68% using 3-, 5-, 8- and 10-year breast cancer survivability data sets. This indicates that the hybrid AdaBoost with Random Forests are better than basic AdaBoost based on Kappa statistics.

7.4.3.6 Discussion of classification results

Survival analysis in the field of medical prognosis mainly uses various applications and methods from historical data for predicting the survival of particular patients suffering from diseases over particular time periods [129]. Nonetheless, period analysis is a new method of survival analysis used to monitor survival rates and provide up-to-date estimations of long-term survival [4] [129]. This method is commonly utilised to build a 5-year breast cancer survivability prediction model [4] [136], possibly due to the fact that few problems affect classifier performance. However, in this section, not only a 5-year breast cancer prediction model but also 3-, 8- and 10-year breast cancer prediction models are built using AdaBoost with Random Forests (ABRF). Moreover, accuracy, AUC,

F-measure and Kappa statistics are employed to evaluate these prediction models. There are several possible explanations for these impressive results.

Firstly, ABRF is found to significantly improve the performance of basic AdaBoost among 3-, 5-, 8- and 10-year breast cancer prediction models. This may be due to the fact that ABRF can generate several trees to select the best trees, but basic AdaBoost uses the decision stump to generate the separation line which can mislead into classifying wrongly. These results are congruous with those of Leshem and Ritov [33], who found that AdaBoost with Random Forests is better than basic AdaBoost based on error rates.

Secondly, the ABRF classifier is significantly better than most classifiers using 3- and 5-year breast cancer survivability data sets. This may be due to the fact that the ABRF classifier is less sensitive to outlier and imbalanced data than the other classifiers especially Naïve Bayes which is sensitive to outliers and imbalanced data [147].

Thirdly, in relation to cancer survivability prediction models, a 5-year breast cancer survivability prediction model using our approach (OOS) achieved 94.01% accuracy when generated from C4.5 whereas Jonsdottir et al. [20] attained 80.00% of the accuracy using C4.5. On the other hand, Delen and Patil [65] accomplished 90.00% accuracy of prostate cancer survivability prediction model utilizing the CART technique.

Lastly, Random Forests alone is found to achieve similar results using a 5-year breast cancer survivability data set. This may be due to the fact that this data set contains less outliers and balanced data that reduce the overfitting problem in most of the classifiers. However, ABRF provides the accuracy of 3-, 5-, 8- and 10-year breast cancer prediction model up to 96.96%, 96.60%, 98.29% and 98.35%. On the other hand, Delen, Walker and Kadam [4] provided 93.6% of the accuracy of a 5-year breast cancer survivability

prediction model using C5 in SEER databases, whereas Jonsdottir et al. [20] presented that the accuracy of three classifiers, including Naïve Bayes, C4.5 and a PART decision list, was up to 80% using a 5-year breast cancer outcomes data set. Therefore, the proposed approach used in this thesis is suitable to build the accurate and reliable prediction models.

7.5 Chapter summary

In this chapter, a combination of the AdaBoost and Random Forests (ABRF) algorithms has been proposed for constructing breast cancer survivability prediction models. The development of breast cancer survivability prediction models consisted of two sections. Firstly, a 5-year breast cancer survivability prediction model was developed and the capability and effectiveness of the proposed models was evaluated using accuracy, sensitivity and specificity. The experimental results in the first part indicated that the ABRF model achieved 88.60% accuracy and reduces the overfitting in Random Forests. Moreover, it outperforms several single and combined classifiers.

Secondly, 3-, 5-, 8- and 10-year breast cancer survivability prediction models were developed and the capability and effectiveness of these models was evaluated using accuracy, AUC scores, *F*-measure and Kappa statistics. The performance and effectiveness of the proposed method using, accuracy, AUC, *F*-measure and Kappa statistics have been illustrated. The results showed that the proposed method improved the accuracy up to 96.96%, 96.60%, 98.29% and 98.35% while the AUC scores were improved up to 99.59%, 99.09%, 99.81% and 100%. As a result, it seems that AdaBoost with Random Forests demonstrated promising results when compared to basic AdaBoost. These prediction models are used to predict the class label of the new cases to enhance the deci-

sion-making systems. In the next chapter the interpretation of breast cancer survivability prediction models will be provided in the form of decision trees and decision rules, which are the most widely used and easy to understand models.

Chapter 8

Breast Cancer Survivability Outcomes

In Chapter 7, the hybrid algorithm of AdaBoost and Random Forests was proposed to develop accurate breast cancer survivability prediction models. The performance and effectiveness of these prediction models were evaluated using five measurement criteria including accuracy, sensitivity, specificity, Area Under the receiver operating characteristic Curve (AUC), *F*-measure and Kappa statistics. Results showed that the combined AdaBoost and Random Forests algorithm was superior to other classifiers.

In this chapter, C4.5 and C4.5rules are used to discover the knowledge from 3-, 5-, 8- and 10-year breast cancer survivability data sets. C4.5 is used to present decision trees with error rates while C4.5rules is used to exhibit decision rules with the accuracy of each decision rule. The chapter is organised as follows: Section 8.1 reviews decision trees and decision rules and also provides a brief overview of C4.5 and C4.5rules methods; Section 8.2 presents the data sets used in this chapter; Sections 8.3 and 8.4 illustrate and interpret 3-, 5-, 8- and 10-year breast cancer survivability decision trees and rules; Section 8.5 discusses the results of decision trees and rules; and finally, a summary chapter is presented in Section 8.6.

8.1 Overview and techniques

The interpretation of the models is an important step in data mining in order to enhance user ability and understanding of the models. Such methods, including decision tree

and decision rule, are commonly used to build the models which are easy to interpret [81].

Decision Tree refers to a tree structure for classifying instances in classification problems [53]. This method provides the most promising results and is easy to use. For this reason, this method has been used to build the prediction model and present the tree structure. One of the most widely used decision trees is C4.5. For example, Bellaachia and Guven [19] utilised C4.5 to build a 5-year breast cancer survivability prediction model from SEER databases. Their results presented that the accuracy of C4.5 decision tree was superior to Neural Networks and Naïve Bayes.

Decision Rule refers to a set of 'If- Then' rules for presenting information and knowledge in databases and interpreting the models [64] [53]. This can be made more comprehensible by reducing the number of conditions in the original rules [114]. In relation to the medical field, common rules are usually decided by practitioners and recorded as linguistic knowledge [115]. However, as decision rules generated from a decision rule algorithm are easily understood [44] [50], they are used to combine with previous practitioner knowledge to expand the knowledge-base for more accurate decision making. C4.5rules is a well-known and effective method used to generate decision rules. For instance, Zhou and Jiang [98] successfully applied C4.5rules to produce rules from a data set that was generated using an Artificial Neural Network. Similarly, Nettleton, Calderon-Benavides and Baeza-Yates [213] also successfully employed C4.5rules to identify document profiles which relate to theoretical user behaviour and documents (URL).

In this chapter, well-known C4.5 and C4.5rules are used to build 3-, 5-, 8- and 10-year breast cancer survivability decision trees and decision rules due to the fact that they

provide good prediction results and are easy to interpret.

8.1.1 C4.5 decision tree

C4.5 [96] (see Section 2.3.1) is a classic decision tree algorithm in machine learning. It is used to build a tree structure for classifying a training set related to a class attribute consisting of nodes and leaves [53] [97] [98]. Nodes represent rules which categorise data according to attributes and leaves represent the condition in each rule. It also provides the resubstitution error rate in each leaf. This error rate is the relationship between the number of the incorrect cases (E) and training cases covered by the leaf (N). The resubstitution error rate is shown in Equation 8.1.

$$\text{Resubstitution error rate} = E/N. \tag{8.1}$$

In this way, the C4.5 decision tree model is easy to interpret from a tree structure [98] [99], provides a short computation time for building a model [98] [99] [100] and presents the resubstitution error rate in helping users for their decision making.

8.1.2 C4.5rules decision rule

C4.5rules [96] is a method used to generate decision rules from the C4.5 decision tree. The objective of C4.5rules is to identify specific higher precision rules [213]. Six main steps of C4.5rules include: 1) train a C4.5 decision tree directly from a data set; 2) convert every path from the root to a leaf; 3) remove each initial rule antecedent for distinguishing a specific class from other classes; 4) combine rules into rule sets; 5) sort rule sets into an ascending order of false positive error rates; and 6) create a default rule to deal with instances that are not covered by any of the generated rules. Although the generalisation ability of C4.5rules is better than the C4.5 decision tree in some data sets

[96], C4.5rules has limitations in providing a good global precision of rules model [213].

8.2 Breast cancer survivability data sets

Breast cancer survivability data were obtained from breast cancer databases at Srinagarind Hospital in Thailand from 1985-2004. The input attributes are exhibited in Table 8.1.

Table 8.1: Input attributes

No.	Attributes	Attribute types
1	Age (age)	Number
2	Marital status (status)	Category
3	Basis of diagnosis (basic)	Category
4	Topography (top)	Category
5	Morphology (mor)	Category
6	Extent (ext)	Category
7	Stage (stage)	Category
8	Received surgery (surg)	Category
9	Received radiation (radi)	Category
10	Received chemotherapy (chem)	Category
11	Received hormonal therapy (horm)	Category
12	Received supportive therapy (supt)	Category
13	Received other therapy (other)	Category
14	Survivability (Class)	Category

Table 8.1 displays input attributes which consist of a number, 12 categories and a class attribute. In order to build prediction models, four data sets include 3-, 5-, 8- and 10-year survival periods of patients surviving from breast cancer. The number of instances is tabulated in Table 8.2.

Table 8.2: Instances within data sets

Data Sets	Years	Instances		
		'Dead'	'Alive'	Total
3-year	1985-2004	542	544	1086
5-year	1985-2002	368	367	735
8-year	1985-1999	265	264	529
10-year	1985-1997	212	211	423

Table 8.2 shows the number of instances of 3-, 5-, 8- and 10-year breast cancer survivability data sets which involve two classes including 'Dead' and 'Alive'. The 'Dead' class refers to patients who died within three years following diagnosis, while the 'Alive' class refers to patients who have survived for three years or more post diagnosis. Similarly, the 'Dead' class in 5-, 8- and 10-year breast cancer survivability data sets refers to patients who died within five, eight and 10 years, respectively after the first diagnosis and the 'Alive' class refers to patients who have survived for five or more, eight or more and 10 years or more, from the first diagnosis. These data sets have few outliers and the data are balanced in order to yield high prediction results.

8.3 Breast cancer survivability decision tree models

In order to interpret 3-, 5-, 8- and 10-year breast cancer survivability decision tree models, C4.5 is utilised. Each decision tree contains nodes, leaves, predicted class and resubstitution error rates. In this way, the 3-, 5-, 8- and 10-year breast cancer survivability decision trees are presented with a resubstitution error rate in section 8.3.1, 8.3.2, 8.3.3 and 8.3.4, respectively.

8.3.1 Decision tree for predicting 3-year breast cancer survivability

In order to interpret the 3-year breast cancer survivability model, a decision tree is generated using C4.5 as shown in Figure 8.1.

```
ext = 1: 1 (1.0/0.8)
ext = 2: 1 (43.0/1.4)
ext = 3:
| supt = 2: 1 (278.0/7.4)
| supt = 1:
| | mor = 8000: 1 (0.0)
| | mor = 8001: 1 (0.0)
| | mor = 8010: 0 (5.0/1.2)
| | mor = 8020: 1 (0.0)
| | mor = 8041: 1 (0.0)
| | mor = 8070: 1 (0.0)
| | mor = 8140: 1 (0.0)
| | mor = 8260: 1 (0.0)
| | mor = 8480: 1 (0.0)
| | mor = 8500: 1 (7.0/1.3)
| | mor = 8501: 1 (0.0)
| | mor = 8510: 1 (0.0)
| | mor = 8520: 1 (0.0)
| | mor = 8522: 1 (0.0)
| | mor = 8523: 1 (0.0)
| | mor = 8530: 1 (0.0)
| | mor = 8541: 1 (0.0)
| | mor = 8800: 1 (0.0)
| | mor = 9020: 1 (0.0)
ext = 4:
| age <= 45 :
| | mor = 8000: 0 (18.0/2.5)
| | mor = 8001: 1 (0.0)
| | mor = 8010: 1 (0.0)
| | mor = 8020: 1 (0.0)
| | mor = 8041: 1 (0.0)
| | mor = 8070: 1 (0.0)
| | mor = 8140: 1 (4.0/1.2)
| | mor = 8260: 1 (0.0)
| | mor = 8480: 1 (2.0/1.0)
| | mor = 8500: 1 (84.0/1.4)
| | mor = 8501: 1 (2.0/1.0)
| | mor = 8510: 1 (0.0)
| | mor = 8520: 1 (0.0)
| | mor = 8522: 1 (1.0/0.8)
| | mor = 8523: 1 (0.0)
| | mor = 8530: 0 (3.0/1.1)
| | mor = 8541: 1 (1.0/0.8)
| | mor = 8800: 1 (0.0)
| | mor = 9020: 0 (3.0/1.1)
| age > 45 :
| | age <= 55 :
| | | top = 500: 0 (3.0/1.1)
| | | top = 501: 0 (0.0)
| | | top = 503: 0 (0.0)
| | | top = 504: 1 (5.0/1.2)
| | | top = 505: 1 (1.0/0.8)
| | | top = 508: 0 (0.0)
| | | top = 502:
| | | | age <= 51 : 1 (2.0/1.0)
| | | | age > 51 : 0 (2.0/1.0)
| | | top = 509:
| | | | mor = 8000: 0 (12.0/1.3)
```

```
| | | | mor = 8001: 0 (0.0)
| | | | mor = 8010: 0 (0.0)
| | | | mor = 8020: 0 (0.0)
| | | | mor = 8041: 0 (0.0)
| | | | mor = 8070: 0 (0.0)
| | | | mor = 8140: 0 (6.0/1.2)
| | | | mor = 8260: 0 (0.0)
| | | | mor = 8480: 0 (0.0)
| | | | mor = 8501: 0 (0.0)
| | | | mor = 8510: 1 (2.0/1.0)
| | | | mor = 8520: 1 (2.0/1.0)
| | | | mor = 8522: 0 (0.0)
| | | | mor = 8523: 0 (0.0)
| | | | mor = 8530: 0 (1.0/0.8)
| | | | mor = 8541: 1 (1.0/0.8)
| | | | mor = 8800: 0 (0.0)
| | | | mor = 9020: 0 (0.0)
| | | | mor = 8500:
| | | | | age <= 54 : 0 (120.0/3.8)
| | | | | age > 54 :
| | | | | | radi = 1: 0 (6.0/1.2)
| | | | | | radi = 2: 1 (5.0/1.2)
| | age > 55 :
| | | mor = 8001: 1 (0.0)
| | | mor = 8010: 0 (1.0/0.8)
| | | mor = 8020: 1 (0.0)
| | | mor = 8041: 1 (0.0)
| | | mor = 8070: 1 (0.0)
| | | mor = 8140: 1 (0.0)
| | | mor = 8260: 1 (0.0)
| | | mor = 8480: 1 (1.0/0.8)
| | | mor = 8500: 1 (58.0/3.8)
| | | mor = 8501: 1 (1.0/0.8)
| | | mor = 8510: 1 (0.0)
| | | mor = 8520: 1 (0.0)
| | | mor = 8522: 1 (0.0)
| | | mor = 8523: 1 (0.0)
| | | mor = 8530: 0 (6.0/1.2)
| | | mor = 8541: 1 (1.0/0.8)
| | | mor = 8800: 0 (2.0/1.0)
| | | mor = 9020: 1 (0.0)
| | | mor = 8000:
| | | | radi = 1: 0 (8.0/1.3)
| | | | radi = 2: 1 (3.0/1.1)
ext = 5:
| top = 500: 0 (0.0)
| top = 501: 1 (1.0/0.8)
| top = 502: 1 (1.0/0.8)
| top = 503: 0 (0.0)
| top = 504: 1 (9.0/4.5)
| top = 505: 0 (2.0/1.0)
| top = 508: 0 (0.0)
| top = 509:
| | age > 43 : 0 (272.0/16.2)
| | age <= 43 :
| | | mor = 8000: 0 (40.0/2.6)
| | | mor = 8001: 0 (0.0)
| | | mor = 8010: 0 (1.0/0.8)
```

```
| | | mor = 8020: 0 (0.0)
| | | mor = 8041: 0 (0.0)
| | | mor = 8070: 0 (0.0)
| | | mor = 8140: 0 (6.0/2.3)
| | | mor = 8260: 0 (0.0)
| | | mor = 8480: 0 (0.0)
| | | mor = 8501: 0 (0.0)
| | | mor = 8510: 0 (0.0)
| | | mor = 8520: 0 (0.0)
| | | mor = 8522: 0 (0.0)
| | | mor = 8523: 0 (0.0)
| | | mor = 8530: 0 (0.0)
| | | mor = 8541: 0 (0.0)
| | | mor = 8800: 0 (0.0)
| | | mor = 9020: 0 (5.0/1.2)
| | | mor = 8500:
| | | | age > 40 : 1 (11.0/1.3)
| | | | age <= 40 :
| | | | | age > 36 : 0 (26.0/2.6)
| | | | | age <= 36 :
| | | | | | chem = 1: 1 (8.0/1.3)
| | | | | | chem = 2: 0 (3.0/1.1)
```

Figure 8.1: Decision tree model for predicting 3-year breast cancer survivability

Figure 8.1 illustrates the 3-year breast cancer survivability decision tree used to predict a class label. This decision tree model starts with the extent of breast cancer as the root and moving through it until a leaf is found. There are five conditions referring to the extent of the breast cancer (see Appendix A.6) including: 1) in situ refers to '1'; 2) localised refers to '2'; 3) direct extension refers to '3'; 4) regional lymph nodes refers to '4'; and 5) distant metastases refers to '5'. The examples of interpreting the decision tree above are presented as follows.

1) If the extent of breast cancer of a patient is an in situ ('ext' = '1'), then this patient is predicted to survive for three years or more after the first diagnosis with 1.0/0.8 error rates. Moreover, if a patient has a localised ('ext' = '2'), then this patient is predicted to survive for three years or more after the first diagnosis with 4.3/1.4 error rates.

2) If a patient with direct extension of breast cancer ('ext' = '3'), receives a supportive therapy ('supt' = '1') and the morphology is an epithelial tumour ('mor' = '8010'), then this patient is predicted to live less than three years after the first diagnosis with 5.0/1/2 resubstitution error rates. Whereas if morphology is infiltrating duct carcinoma in the third condition ('mor' = '8500'), then this patient is predicted to live for three years or more after the first diagnosis with 7.0/1/3 resubstitution error rates.
3) If a patient has the regional lymph nodes ('ext' = '4'), and the age of the patient at the first diagnosis is 45 years or younger ('age' <= 45) and morphology is equal to neoplasm ('mor' = '8000'), then this patient is predicted to live less than three years after the first diagnosis with 18/2.5 resubstitution error rates. Whereas if morphology is equal to Adenocarcinoma ('mor' = '8140'), then this patient is predicted to live for three years or more after the first diagnosis with 4.0/1.2 resubstitution error rates.
4) If a patient has distant metastases ('ext' = '5') and central portion of breast ('top' = '501') at the first diagnosis, then this patient is predicted to live for three years or longer after the first diagnosis with 1.0/8.0 resubstitution error rates. But if a patient has distant metastases ('ext' = '5') and lower-outer quadrant of b ('top' = '505') at the first diagnosis, then this patient is predicted to die less than three years after the first diagnosis with 2.0/1.0 resubstitution error rates.

As a result, in the case of a patient who is predicted to live for less than three years following the first diagnosis with low resubstitution error rates, medical practitioners should pay more attention to such a patient to improve the situation and evaluate decisions for further treatment.

8.3.2 Decision tree for predicting 5-year breast cancer survivability

The 5-year breast cancer survivability decision tree is created using a C4.5 algorithm.

This decision tree is displayed in Figure 8.2.

```
ext = 2: 1 (37.0/3.8)
ext = 3: 1 (268.0/16.2)
ext = 4:
| age > 44 : 0 (148.0/7.3)
| age <= 44 :
| | surg = 2: 0 (3.0/1.1)
| | surg = 1:
| | | mor = 8001: 1 (0.0)
| | | mor = 8010: 1 (0.0)
| | | mor = 8041: 1 (0.0)
| | | mor = 8070: 1 (0.0)
| | | mor = 8140: 1 (5.0/2.3)
| | | mor = 8480: 1 (1.0/0.8)
| | | mor = 8501: 1 (0.0)
| | | mor = 8510: 0 (1.0/0.8)
| | | mor = 8520: 1 (0.0)
| | | mor = 8530: 0 (1.0/0.8)
| | | mor = 8541: 1 (0.0)
| | | mor = 8800: 1 (0.0)
| | | mor = 9020: 0 (1.0/0.8)
| | | mor = 8000:
| | | | age <= 34 : 1 (3.0/1.1)
| | | | age > 34 : 0 (7.0/1.3)
| | | mor = 8500:
| | | | age <= 35 : 1 (15.0/1.3)
| | | | age > 35 :
| | mor = 8480: 0 (1.0/0.8)
| | mor = 8501: 0 (0.0)
| | mor = 8510: 0 (2.0/1.0)
| | mor = 8520: 0 (0.0)
| | mor = 8530: 0 (0.0)
| | | | | age > 40 : 1 (24.0/3.7)
| | | | | age <= 40 :
| | | | | | top = 500: 0 (0.0)
| | | | | | top = 501: 0 (0.0)
| | | | | | top = 502: 1 (2.0/1.0)
| | | | | | top = 503: 0 (0.0)
| | | | | | top = 504: 1 (6.0/1.2)
| | | | | | top = 505: 0 (0.0)
| | | | | | top = 506: 0 (0.0)
| | | | | | top = 508: 0 (1.0/0.8)
| | | | | | top = 509: 0 (19.0/1.3)
ext = 5:
| top = 500: 0 (0.0)
| top = 501: 0 (0.0)
| top = 502: 0 (0.0)
| top = 503: 0 (0.0)
```

```
| top = 504: 1 (6.0/1.2)
| top = 505: 0 (1.0/0.8)
| top = 506: 0 (0.0)
| top = 508: 0 (0.0)
| top = 509:
| | mor = 8001: 0 (1.0/0.8)
| | mor = 8041: 0 (0.0)
| | mor = 8070: 0 (0.0)
| | mor = 8140: 0 (13.0/1.3)
| | mor = 8541: 0 (0.0)
| | mor = 8800: 0 (0.0)
| | mor = 9020: 0 (2.0/1.0)
| | mor = 8000:
| | | age > 33 : 0 (45.0/1.4)
| | | age <= 33 :
| | | | radi = 1: 0 (3.0/1.1)
| | | | radi = 2: 1 (3.0/1.1)
| | mor = 8010:
| | | radi = 1: 0 (3.0/1.1)
| | | radi = 2:
| | | | age <= 52 : 1 (8.0/1.3)
| | | | age > 52 : 0 (2.0/1.0)
| | mor = 8500:
| | | age <= 66 : 0 (97.0/1.4)
| | | age > 66 :
| | | | age <= 74 : 0 (4.0/1.2)
| | | | age > 74 : 1 (2.0/1.0)
```

Figure 8.2: Decision tree model for predicting 5-year breast cancer survivability

Figure 8.2 illustrates a 5-year breast cancer survivability decision tree generated from the C4.5 algorithm using data from 1985 to 2002. The main root of this decision tree is also the extent of breast cancer including: localised ('2'), direct extension ('3'), regional lymph nodes ('4') and distant metastases ('5'). The interpretation of a 5-year breast cancer survivability decision tree is shown below.

1) If a patient has localised extent ('ext' = '2') at the first diagnosis then this patient is predicted to live for five years or longer after the first diagnosis with 37.0/3.8 resubstitution error rates. Similarly, if a patient has direct extension ('ext' = '3') then this patient is also predicted to live for five years or more after the first diagnosis with 268/16.2 resubstitution error rates.
2) If a patient has regional lymph nodes ('ext' = '4') and is older than 44 years ('age' > 44) at the first diagnosis, then this patient is predicted to live less than

five years after the first diagnosis with 148.0/7.3 resubstitution error rates. Similarly, if a patient has the regional lymph nodes ('ext' = '4'), is aged 44 years or younger ('age' <= 44), at the first diagnosis and did not receive surgery treatment ('surg' = '2'), then this patient is predicted to live less than five years following the first diagnosis with 3.0/1.1 resubstitution error rates. On the other hand, if a patient has regional lymph nodes ('ext' = '4'), is aged 44 years or younger ('age' <= 44), received a surgery treatment ('surg' = '1') and had adenocarcinoma ('mor' = '8140'), then this patient is predicted to live for five years or longer after the first diagnosis with 5.0/2.3 resubstitution error rates.

3) If a patient has distant metastases ('ext' = '5') and upper-outer quadrant of b ('top' = '504') at the first diagnosis, then this patient is predicted to live for five years or more after the first diagnosis with 6.0/1.2 resubstitution error rates. On the other hand, if a patient has distant metastases ('ext' = '5') and lower-outer quadrant of b ('top' = '505'), then this patient is predicted to live less than five years after the first diagnosis with 1.0/.8 resubstitution error rates.

These decisions are different from the 3-year breast cancer survivability decision tree due to the fact that they are independent of each other.

8.3.3 Decision tree for predicting 8-year breast cancer survivability

The 8-year breast cancer survivability decision tree is produced from a C4.5 algorithm. The decision tree is presented in Figure 8.3 below.

```
stage = 3: 0 (107.0/1.4)
stage = 4: 0 (114.0/1.4)
stage = 1:
| age <= 57 : 1 (32.0/1.4)
| age > 57 : 0 (2.0/1.0)
stage = 2:
| age <= 47 :
| | mor = 8000: 0 (1.0/0.8)
| | mor = 8001: 0 (1.0/0.8)
| | mor = 8010: 1 (0.0)
| | mor = 8070: 0 (1.0/0.8)
| | mor = 8140: 1 (8.0/2.4)
| | mor = 8200: 1 (0.0)
| | mor = 8480: 1 (0.0)
| | mor = 8501: 1 (0.0)
| | mor = 8510: 0 (1.0/0.8)
| | mor = 8541: 1 (0.0)
| | mor = 8800: 1 (7.0/1.3)
| | mor = 8500:
| | | age > 40 : 1 (157.0/1.4)
| | | age <= 40 :
| | | | age > 36 : 0 (6.0/1.2)
| | | | age <= 36 :
| | | | | age > 32 : 1 (36.0/2.6)
| | | | | age <= 32 :
| | | | | | basis = 1: 1 (0.0)
| | | | | | basis = 2: 1 (0.0)
| | | | | | basis = 3: 1 (0.0)
| | | | | | basis = 5: 1 (10.0/1.3)
| | | | | | basis = 6: 1 (0.0)
| | | | | | basis = 7:
| | | | | | | status = 1: 1 (5.0/1.2)
| | | | | | | status = 2: 0 (4.0/1.2)
| | | | | | | status = 3: 1 (0.0)
| age > 47 :
| | surg = 1: 0 (26.0/1.3)
| | surg = 2: 1 (11.0/1.3)
```

Figure 8.3: Decision tree model for predicting 8-year breast cancer survivability

Figure 8.3 illustrates the 8-year breast cancer survivability decision tree generated from the C4.5 algorithm using data from 1985 to 1999. The main root of this decision tree is the stage of breast cancer including: 1) Stage I; 2) Stage II; 3) Stage III; and 4) Stage IV. The examples of the interpretation of the 8-year breast cancer survivability decision tree are displayed below.

1) If a patient has breast cancer in stage III ('stage' = '3') at the first diagnosis, this patient is predicted to live less than eight years with 107.0/1.4 resubstitution error rates. Likewise, if a patient has breast cancer in stages IV ('stage' = '4') at the first diagnosis, this patient is predicted to live less than eight years with 114.0/1.4 resubstitution error rates.
2) If a patient has breast cancer in stage I ('stage' = '1') and is aged 57 years or younger ('age' <= '57') at the first diagnosis, then this patient is predicted to live for eight years or more after the first diagnosis with 32.0/1.4 resubstitution error rates. On the other hand, if a patient has breast cancer in stage I ('stage' = '1') and is older than 57 ('age' > '57') at the first diagnosis, this patient is predicted to live less than eight years after the first diagnosis with 2.0/1.0 resubstitution error rates.
3) If a patient has breast cancer in stage II ('stage' = '2'), is aged 47 years or younger ('age' <= 47) and had adenocarcinoma ('mor' = '8140') at the first diagnosis, then this patient is predicted to live for eight years or more after the first diagnosis with 8.0/2.4 resubstitution error rates. However, if a patient has breast cancer in stage II ('stage' = '2'), is older than 47 years ('age' > 47) at the first diagnosis and received a surgery treatment ('surg' = '1'), then this patient is predicted to live less than eight years after the first diagnosis with 26.0/1.3 resubstitution error rates. However, if she has not received surgery treatment ('surg' = '2') then this patient is predicted to live for eight years or more after the first diagnosis with 11.0/1.3 resubstitution error rates. This means that patients aged 47 years and over are more suited to receiving other treatments as opposed to surgery.

8.3.4 Decision tree for predicting 10-year breast cancer survivability

The 10-year breast cancer survivability decision tree is conducted from a C4.5 algorithm. The decision tree is shown in Figure 8.4.

```
stage = 3: 0 (86.0/1.4)
stage = 4: 0 (93.0/1.4)
stage = 1:
| age <= 46 : 1 (37.0/1.4)
| age > 46 : 0 (3.0/1.1)
stage = 2:
| age <= 46 :
| | age > 33 : 1 (162.0/2.6)
| | age <= 33 :
| | | basis = 1: 1 (0.0)
| | | basis = 2: 1 (0.0)
| | | basis = 3: 1 (0.0)
| | | basis = 5: 1 (8.0/1.3)
| | | basis = 7: 0 (6.0/1.2)
| age > 46 :
| | age <= 73 : 0 (23.0/1.3)
| | age > 73 : 1 (5.0/1.2)
```

Figure 8.4: Decision tree model for predicting 10-year breast cancer survivability

Figure 8.4 shows the 10-year breast cancer survivability decision tree model generated from the C4.5 algorithm using data from 1985 to 1997. The main root of this decision tree is the stage of breast cancer including Stage I, Stage II, Stage III and Stage IV. The interpretation of the 10-year breast cancer survivability decision tree is presented below.

1) If a patient is diagnosed with stage III ('stage' = '3') of breast cancer at the first diagnosis then this patient is predicted to live less than 10 years after the first diagnosis with 86.0/1.4 resubstitution error rates. Similarly, if a patient is diagnosed with stage IV ('stage' = '4') of breast cancer at the first diagnosis then this patient is predicted to live less than 10 years after the first diagnosis with 93.0/1.4 resubstitution error rates.
2) If a patient is diagnosed with breast cancer at stage I ('stage' = '1') and is 46 years old or younger ('age' <= '46') at the first diagnosis then this patient is predicted to live for 10 years or more after the first diagnosis with 37.0/1.4 resubsti-

tution error rates. On the other hand, if a patient is diagnosed with breast cancer in stage I ('stage' = '1') and is older than 46 ('age' > '46') at the first diagnosis, then this patient is predicted to live less than 10 years after the first diagnosis with 3.0/1.1 resubstitution error rates.

3) If a patient has a breast cancer in stage II ('stage' = '2'), is aged more than 33 and less than or equal to 46 years ('age' > '33' and 'age' <= '46') at the first diagnosis, then this patient is predicted to live for 10 years or more after the first diagnosis with 162.0/2.6 resubstitution error rates. However, if a patient is 33 years or younger ('age' <= 33) and is histology of primary ('basis' = '7') at the first diagnosis, then this patient is predicted to live less than 10 years after the first diagnosis with 6.0/1.2 resubstitution error rates.

As a result, the extent of breast cancer is the main factor for patients who survive more than three and five years after the first diagnosis while the stage of breast cancer is the main factor for patients who survive more than eight and 10 years after the first diagnosis. This demonstrates that C4.5 not only provides factors of breast cancer but also presents resubstitution error rates for the medical practitioners to rely on for their prognosis and decisions.

8.4 Breast cancer survivability decision rules

In order to exhibit 3-, 5-, 8- and 10-year breast cancer survivability rules, the C4.5rules technique is employed to build decision rules. These rules can combine with previous practitioner knowledge to expand the knowledge-base for more accurate decision making.

8.4.1 Decision rules to predict 3-year breast cancer survivability data

The C4.5rules technique is used to generate 3-year breast cancer survivability decision rules up to 27 rules. Each rule comprises a rule number, conditions, the prediction class and the accuracy of the decision rule showed in Table 8.3.

Table 8.3: Rules for predicting 3-year breast cancer survivability

	Conditions						Classes	Accuracy (%)
Rule No.	1	2	3	4	5	6		
Rule 1:	ext = 2						1	96.8
Rule 2:	mor = 8010						0	82.3
Rule 3:	**mor = 8500**	**ext = 3**					**1**	**99.4**
Rule 4:	ext = 3	supt = 2					1	97.4
Rule 5:	**mor = 8500**	**ext = 4**	**age <= 45**				**1**	**98.4**
Rule 6:	mor = 8530						0	89.1
Rule 7:	age <= 51	top = 502	age > 45				1	70.7
Rule 8:	mor = 8000	ext = 4	age <= 55				0	91.5
Rule 9:	age <= 54	top = 509	mor = 8500	stage = 3	age > 45		0	97.8
Rule 10:	**top = 509**	**mor = 8500**	**radi = 1**	**age > 45**	**age <= 55**	**stage = 3**	**0**	**98.4**
Rule 11:	radi = 2	age > 54	mor = 8500	stage = 3			1	96.3
Rule 12:	mor = 8510						1	70.0
Rule 13:	mor = 8520	age <= 55					1	82.0
Rule 14:	top = 509	stage = 4	age > 45	age <= 55			0	97.5
Rule 15:	mor = 8000	Radi = 1					0	96.3
Rule 16:	radi = 2	ext = 4	age > 55				1	93.1
Rule 17:	age <= 67	mor = 8500	ext = 4	age > 55			1	96.9
Rule 18:	age > 69	mor = 8500					1	93.0
Rule 19:	top = 504	mor = 8500					1	94.5
Rule 20:	top = 509	mor = 8000	ext = 5				0	98.0
Rule 21:	**chem = 1**	**age <= 36**	**mor = 8500**				**1**	**98.2**
Rule 22:	top = 509	ext = 5	chem = 2				0	96.3
Rule 23:	age > 36	ext = 5	top = 509	age <= 40			0	93.4
Rule 24:	**age > 40**	**mor = 8500**	**age <= 43**				**1**	**98.0**
Rule 25:	top = 509	mor = 9020					0	84.1
Rule 26:	status = 2	ext = 5	top = 509	age > 43			0	94.1
Rule 27:	age > 55	age <= 59	mor = 8500				1	94.1

Table 8.3 shows the rules of 3-year breast cancer survivability generated from C4.5rules. The interpretation of the top five 3-year breast cancer survivability decision rules including rule numbers 3, 10, 5, 21 and 24 are presented respectively.

Rule 3: if a patient is diagnosed with an infiltrating duct carcinoma ('mor' ='8500') and a direct extension ('ext' = '3') at the first diagnosis then this patient is predicted to survive for three years or more after the first diagnosis with 99.40% accuracy.

Rule 10: if a patient is diagnosed with breast cancer, NOS ('top' = '509'), infiltrating duct carcinoma ('mor' = '8500'), receives radiation ('radi' = '1'), is aged more than 45 and less than or equal to 55 years ('age' > 45 and 'age' <= 55), and is at stage III ('stage' = '3') at the first diagnosis, then this patient is predicted to live less than three years after the first diagnosis with 98.40% accuracy.

Rule 5: if a patient is diagnosed with infiltrating duct carcinoma ('mor' = '8500'), regional lymph nodes ('ext' = '4') and is aged 45 years or younger ('age' <= 45) at the first diagnosis, then this patient is predicted to survive for three years or more after the first diagnosis with 98.40% accuracy.

Rule 21: if a patient receives chemotherapy ('chem' = 1), is aged 36 years or younger ('age' <= 36) and has infiltrating duct carcinoma ('mor' = '8500') at the first diagnosis, then this patient is predicted to survive for three years or more after the first diagnosis with 98.20% accuracy.

Rule 24: if a patient is aged more than 40 and younger than or equal to 43 years old ('age' > 40 and 'age' <= 43) at the first diagnosis and has an infiltrating duct carcinoma ('mor' = '8500'), then this patient is predicted to survive for three years or longer after the first diagnosis with 98.00% accuracy.

In summary, a patient who has breast cancer, NOS, infiltrating duct carcinoma, has received radio therapy, is aged between 45 and 55 years and is at stage III at first diagnosis, is predicted to live less than three years after the first diagnosis with 98.4% accuracy. This rule could help medical practitioners in the field of prognosis to become aware of the therapy that they have given to the patient, with respect to the age of the patient at the first diagnosis in order to improve quality of care.

8.4.2 Decision rules to predict 5-year breast cancer survivability data

The decision rules for predicting 5-year breast cancer survivability are produced using a C4.5rules technique. The 5-year breast cancer survivability decision rules consist of 19 rules with six maximum conditions. The decisions are exhibited in Table 8.4.

Table 8.4: Rules for predicting 5-year breast cancer survivability

	Conditions						Classes	Accuracy (%)
Rule No.	1	2	3	4	5	6		
Rule 1:	ext = 2						1	89.8
Rule 2:	**ext = 3**						**1**	**94.0**
Rule 3:	mor = 8541						0	70.7
Rule 4:	age > 34	mor = 8000					0	86.5
Rule 5:	age <= 35	mor = 8500	ext = 4				1	91.2
Rule 6:	top = 504	age <= 40					1	87.1
Rule 7:	top = 509	ext = 4	age > 35	age <= 40	mor = 8500		0	93.3
Rule 8:	top = 509	ext = 4	age > 40	age <= 44	mor = 8500	surg = 1	1	93.0
Rule 9:	mor = 9020						0	63.0
Rule 10:	ext = 4	surg = 2					0	91.7
Rule 11:	**age > 44**	**age <= 75**	**ext = 4**				**0**	**97.3**
Rule 12:	age > 75	radi = 1					1	63.0
Rule 13:	**radi = 2**	**ext = 4**	**age > 44**				**0**	**96.5**
Rule 14:	top = 504	ext = 5					1	79.4
Rule 15:	**top = 509**	**ext = 5**	**radi = 1**				**0**	**98.2**
Rule 16:	radi = 2	age <= 33	mor = 8000				1	84.1
Rule 17:	age <= 52	mor = 8010	radi = 2				1	84.1
Rule 18:	age > 52	top = 509	ext = 5				0	93.7
Rule 19:	**age <= 74**	**top = 509**	**mor = 8500**	**ext = 5**			**0**	**98.6**

Table 8.4 shows 5-year breast cancer survivability decision rules generated from C4.5rules. The interpretation of the top five rules including rule numbers 19, 15, 11, 13 and 2 are displayed below, respectively;

Rule 19: if a patient is aged 74 years or younger ('age' <= 74), has breast, NOS ('top' = '509'), an infiltrating duct carcinoma ('mor' = '8500') and a distant metastases ('ext' = '5') at the first diagnosis, then this patient is predicted to live less than five years after the first diagnosis with 98.60% accuracy.

Rule 15: if a patient has topography at breast, NOS ('top' = '509'), distant metastases ('ext' = '5') and receives radiation ('radi' = '1') at the first diagnosis, then this patient is predicted to live less then five years after the first diagnosis with 98.20% accuracy.

Rule 11: if a patient is aged more than 44 and younger than or equal to 75 (age > 44 and age <= 75) and has the regional lymph nodes ('ext' = '4') at the first diagnosis, then this patient is predicted to live less than five years after the first diagnosis with 97.30% accuracy.

Rule 13: if a patient does not receive radiation ('radi' = '2'), has regional lymph nodes (ext = 4) and is older than 44 years ('age' > 44) at the first diagnosis, then this patient is predicted to live less than five years after the first diagnosis with 96.50% accuracy.

Rule 2: if a patient had a direct extension ('ext' = '3') at the first diagnosis then this patient is predicted to live for five years or more after the first diagnosis with 94.00% accuracy.

As a result, if a patient had a direct extension then this patient is predicted to live for five years or more yet if a patient has the regional lymph nodes, they need more support from medical practitioners to improve the chance of survival for more than five years.

8.4.3 Decision rules to predict 8-year breast cancer survivability data

In this section, nine decision rules with three maximum number conditions of 8-year breast cancer survivability are generated using the C4.5rules technique. The 8-year breast cancer survivability decision rules are displayed in Table 8.5.

Table 8.5: Rules for predicting 8-year breast cancer survivability

Rule No.	Conditions			Classes	Accuracy (%)
	1	2	3		
Rule 1:	**age <= 57**	**stage = 1**		**1**	**95.8**
Rule 2:	mor = 8000	age <= 47		0	88.2
Rule 3:	age <= 47	top = 509	stage = 2	1	92.2
Rule 4:	status = 2	basis = 7	age <= 32	0	91.2
Rule 5:	**age > 36**	**age <= 40**	**mor = 8500**	**0**	**95.0**
Rule 6:	**surg = 1**	**age > 47**		**0**	**98.9**
Rule 7:	stage = 2	surg = 2		1	88.2
Rule 8:	**stage = 3**			**0**	**98.7**
Rule 9:	**stage = 4**			**0**	**98.8**

Table 8.5 shows the 8-year breast cancer survivability rules generated from C4.5rules. Results of the top five rules involve rule numbers 6, 9, 8, 1 and 5, respectively. These rules are interpreted as follows:

Rule 6: if a patient receives surgery ('surg' = '1') and is older than 47 years ('age' > 47) at the first diagnosis then this patient is predicted to live less than eight years after the first diagnosis with 98.90% accuracy;

Rule 9: if a patient is diagnosed with breast cancer at stage IV ('stage' = '4') at the first diagnosis then this patient is predicted to live less than eight years after the first diagnosis with 98.80% accuracy;

Rule 8: if a patient is diagnosed with breast cancer at stage III ('stage' = '3') at the first time then this patient is predicted to live less than eight years after the first diagnosis with 98.70% accuracy;

Rule 1: if a patient is aged 57 years and younger ('age' <= 57) and has breast cancer at stage I ('stage' = '1') at the first diagnosis then this patient is predicted to live for eight years or more after the first diagnosis with 95.80% accuracy; and

Rule 5: if a patient is aged more than 36 and younger than or equal to 40 years (age > 36 and age <= 40) and has an infiltrating duct carcinoma ('mor' = '8500') at the

first diagnosis then this patient is predicted to live less than eight years after the first diagnosis with 95.00% accuracy.

As a result, 8-year breast cancer survivability rules are uncomplicated due to the fact that they have a common rule which depends on the stage of breast cancer. If a patient was in stages III and IIII at the first diagnosis then this patient is predicted to die within eight years after the first diagnosis.

8.4.4 Decision rules to predict 10-year breast cancer survivability data

The 10-year breast cancer survivability rules generated from C4.5rules include seven in all with two maximum conditions. The rules are illustrated in Table 8.6.

Table 8.6: Rules for predicting 10-year breast cancer survivability

Rule No.	Conditions		Classes	Accuracy (%)
	1	2		
Rule 1:	**age <= 46**	**stage = 1**	**1**	**96.3**
Rule 2:	**age <= 46**	**stage = 2**	**1**	**94.6**
Rule 3:	basis = 7	age <= 33	0	90.6
Rule 4:	basis = 3		0	75.8
Rule 5:	**age > 46**	**age <= 73**	**0**	**98.9**
Rule 6:	**stage = 3**		**0**	**98.4**
Rule 7:	**stage = 4**		**0**	**98.5**

Table 8.6 shows 10-year breast cancer survivability rules generated from a C4.5rules technique. Results of the top five rules including rule numbers 5, 7, 6, 1 and 2 are interpreted, respectively.

Rule 5: if a patient is aged more than 46 and younger than or equal to 73 years old ('age' > 46 and 'age' <= 73) at first diagnosis, then this patient is predicted to live less than 10 years after the first diagnosis with 98.90% accuracy.

Rule 7: if a patient has breast cancer at stage IV ('stage' = '4') at the first diagnosis then this patient is predicted to live less than 10 years after the first diagnosis with 98.50% accuracy.

Rule 6: if a patient has breast cancer at stage III ('stage' = '3') at the first diagnosis then this patient is also predicted to live less than 10 years after the first diagnosis with 98.40% accuracy.

Rule 1: if a patient is aged 46 years or younger ('age' <= 46) and has breast cancer in stage I ('stage' = '1') at the first diagnosis, then this patient is predicted to live for 10 years or longer with 96.3% accuracy.

Rule 2: if a patient is aged 46 years or younger ('age' <= 46) and has breast cancer in stage II ('stage' = '2') at the first diagnosis, then this patient is predicted to live for 10 years or more after the first diagnosis with 94.6% accuracy.

As a result, if a patient was diagnosed with stages III and IV at the first diagnosis, this patient is predicted to die less than 10 years after the first diagnosis. However, the prediction of a patient's survival not only depends on the stage of breast cancer but also on other factors which are beyond the scope of this thesis.

8.5 Discussion of decision trees and rules

Medical data mining is widely used to extract useful patterns and reliable models in the medical field [53]. As a result, these models can be exploited to assist medical practitioners to make accurate decisions and improve health services [136]. Decision tree and decision rule are the most easily understood by medical practitioners [68] [53]. This chapter adopts C4.5 and C4.5rules to generate the decision tree and decision rule. The C4.5 technique generates the decision tree which provides the resubstitution error rate to assist the medical practitioners in their decision making. Likewise, Yao et al. [99] employed C4.5 to construct a decision tree model for predicting the in-patient length of stay. Their results showed that this model can be understood and accepted better by

managers and also in assisting health care organisations to arrange and make full use of hospital data. Similarly, Jonsdottir et al. [20] demonstrated that C4.5 classifier is more accurate in predicting the doctor's risk assessment than the actual outcome from the Med-DS data set. Unlike C4.5 decision tree, C4.5rules generates fewer rules, which reduces the complexity of the decision and it also can be easily updated to the previous rules in the decision-making system. Moreover, Minowa [14] demonstrated that C4.5rules is linguistic information which people can easily comprehend and use.

8.6 Chapter summary

In this chapter, C4.5 and C4.5rules were used to build the decision tree and decision rules which are easy to interpret. The interpretation of the decision tree and rules were also presented to gain better understanding of the decisions in predicting the unseen class for assisting medical practitioners to increase their decision-making processes. The next chapter provides the conclusion of this thesis, and recommends directions for further study.

Chapter 9

Conclusions and Future Work

In this thesis suitable breast cancer survivability models have been derived from data in Thailand to assist medical practitioners with accurate and reliable prediction results. The three major research aims including improved performance, stability and effectiveness of breast cancer survivability prediction models using suitable data mining processes have been achieved. As a result, several approaches including attribute selection, C-support vector classification outlier filtering, over-sampling and hybrid AdaBoost Random Forests, were proposed. The performance, stability and effectiveness of breast cancer survivability prediction models were evaluated using accuracy, sensitivity, specificity, AUC score, *F*-measure and Kappa statistics. The purpose of this final chapter is to summarise the findings and discuss implications for future research directions.

9.1 Summary of results

In order to develop accurate, stable and effective breast cancer survivability prediction models, this dissertation focused on four aspects. The first of these concerns problems with breast cancer survivability data obtained from Srinagarind Hospital databases in Thailand. The second includes the development of approaches in pre-processing to improve data quality in order to enhance the performance, stability and effectiveness of

prediction models. The third involves the combination of the new AdaBoost with Random Forests algorithms to develop superior breast cancer survivability prediction models, and finally breast cancer survivability decision trees and rules have been generated. The results are summarised into four aspects as follows:

1) An understanding of the behaviour of attributes in breast cancer databases in Thailand which affect the performance of prediction models was investigated. Common problems of medical data including missing data, outliers and imbalanced data, are presented in Chapter 3.

2) Three approaches to improving data quality from Srinagarind Hospital in Thailand were developed and introduced in Chapters 4, 5 and 6, respectively.

   Firstly, the *k*-means algorithm was used to transform numeric attribute (age) into a discrete attribute, while RELIEF was utilised to select the suitable attributes. After applying both *k*-means and RELIEF to improve quality of the data set, AdaBoost algorithms were exploited to build 5-year breast cancer survivability prediction models. The capability and effectiveness of this approach were evaluated using accuracy, sensitivity and specificity of prediction models. Results showed that this approach not only improves the performance of prediction models, but also stabilises them.

   Secondly, C-Support Vector Classification Filtering (C-SVCF) was developed to identify and eliminate outliers from misclassified instances in order to improve the quality of a 5-year breast cancer survivability data set. The capability and effectiveness of this approach were evaluated using accuracy and Area Under the receiver operating characteristic Curve (AUC) scores. The experimental re-

sults demonstrated that C-SVCF is superior to AdaBoost filtering, Bagging and SVM ensembles.

Lastly, a combination of Outlier filtering and Over-Sampling (OOS) was utilised to improve the quality of 1-, 2-, 3-, 4-, 5-, 6-, 7-, 8-, 9- and 10-year breast cancer survivability data sets. The capability and effectiveness of this approach were evaluated using accuracy, sensitivity, specificity, AUC scores and *F*-measure. Results illustrated that OOS outperforms C-SVCF and over-sampling.

3) The development of 3-, 5-, 8- and 10-year breast cancer survivability prediction models using a hybrid algorithm, AdaBoost and Random Forests (ABRF), was proposed in Chapter 7. Performance and effectiveness of these prediction models were evaluated using accuracy, sensitivity, specificity, AUC scores, *F*-measure and Kappa statistics. Results exhibited that these combined prediction models are superior to several classifiers such as C4.5, basic AdaBoost and Random Forests.
4) C4.5 and C4.5rules were employed to extract knowledge into decision trees and derive the rules from 3-, 5-, 8- and 10-year breast cancer survivability data sets. These decision rules are used to reinforce with previous practitioner knowledge in order to enhance the decision making in Chapter 8.

9.2 Limitations of the current study

The limitation of the study can be divided into three main issues. Firstly, the current thesis was not specifically designed to evaluate CPU time to build the learning process in the large data sets. This may be due to the limitation of the data size. Secondly, the

finding might not be transferable to other medical data sets. Lastly, the combination of AdaBoost and Random Forests is unable to provide the final tree structures and rules.

9.3 Future research

The burgeoning growth of medical databases and the increasingly huge amount of data to be processed has created massive challenges to medical professionals in Thailand. Appling the outlier filtering approach to the whole data set may introduce the biased problem to the results. Moreover, the combination of AdaBoost and Random Forests used to develop prediction models is time consuming. This is because the AdaBoost algorithm needs to apply several iterations to gain better instances and the Random Forests method needs to produce many trees to achieve high performance. Therefore, future research will need to investigate the improvement of the training time in the combination of AdaBoost and Random Forests while retaining the same performance and effectiveness and using these approaches in the rare event detection. Moreover, such research could develop a method to interpret the combination of AdaBoost and Random Forests prediction models.

References

[1] M. T. Skevofilakas, K. S. Nikita, P. H. Templaleksis, K. N. Birbas, I. G. Kaklamanos, and G. N. Bonatsos, *A decision support system for breast cancer treatment based on data mining technologies and clinical practice guidelines.* Twenty-Seventh Annual International Conference on Medicine and Biology Society, 2005, pp. 2429-2432, IEEE-EMBS.

[2] National Cancer Institute of Thailand, *Cancer in Thailand 1995-1997.* http://www.nci.go.th/cancer_record/. 15 Jul 2007.

[3] T. Srinivasan, A. Chandrasekhar, J. Seshadri, and J. B. S. Jonathan, *Knowledge discovery in clinical databases with neural network evidence combination.* in *Proceedings of the International Conference on Intelligent Sensing and Information Processing.* 2005, pp. 512-517, IEEE.

[4] D. Delen, G. Walker, and A. Kadam, *Predicting breast cancer survivability: a comparison of three data mining methods.* J. Artificial Intelligence in Medicine, 2005, 34(2): pp. 113-127, Elsevier.

[5] Y. U. Ryu, R. Chandrasekaran, and V. S. Jacob, *Breast cancer prediction using the isotonic separation technique.* J. European Operational Research, 2007, 181(2): pp. 842-854, Elsevier.

[6] S. Borovkova, *Analysis of survival data.* http://www.math.leidenuniv.nl/~naw/serie5/deel03/dec2002/pdf/borovkova.pdf.

[7] L. Ohno-Machado, *A comparison of Cox proportional hazards and artificial neural network models for medical prognosis.* J. Computers in Biology and Medicine, 1997, 27(1): pp. 55-65, Elsevier.

[8] S. Tsumoto, *Mining diagnostic rules from clinical databases using rough sets and medical diagnostic model.* J. Information Sciences, 2004, 162(2): pp. 65-80, Elsevier Science Inc. New York, NY, USA

[9] J. Li, A. W.-C. Fu, H. He, J. Chen, and C. Kelman, *Mining risk patterns in medical data.* in *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining.* 2005, pp. 770-775, ACM New York, NY, USA.

[10] K. Revett, S. T. de Magalhaes, and H. M. D. Santos, *Data mining a prostate cancer dataset using rough sets.* in *Proceedings of the Third International IEEE Conference on Intelligent Systems.* 2006, pp. 290-293, IEEE.

[11] D. Shalvi, and N. DeClaris, *An unsupervised neural network approach to medical data mining techniques.* in *Proceedings of the IEEE World Congress on Computational Intelligence.* 1998, 1: pp. 171-176, IEEE.

[12] A. Petrovski, and J. McCall, *Smart problem solving environment for medical decision support.* in *Proceedings of the Workshops on Genetic and Evolutionary Computation.* 2005, pp. 152-158, ACM.

[13] D. Corrigan, N. Harte, and A. Kokaram, *Pathological motion detection for robust missing data treatment in degraded archived media.* in *Proceedings of the IEEE International Conference on Image.* 2006, pp. 621-624, IEEE.

[14] Y. Minowa, *Classification rules discovery from selected trees for thinning with the C4.5 machine learning system* J. Forest Research, 2005, 10(3): pp. 221–231, Springer Japan.

[15] Á. Blanco, A. M. Ricket, and M. Martín-Merino, *Combining SVM classifiers for email anti-spam filtering.* in *Proceedings of the Ninth International Work-Conference on Artificial Neural Networks.* 2007, 4507: pp. 903-910, Springer Berlin.

[16] X. Zhang, and F. Ren, *Improving SVM learning accuracy with Adaboost.* in *Proceedings of the Fourth International Conference on Natural Computation.* 2008, 03: pp. 221-225, IEEE Computer Society.

[17] L. Leyrit, T. Chateau, C. Tournayre, and J. T. Lapreste, *Association of AdaBoost and Kernel based machine learning methods for visual pedestrian recognition.* in *Proceedings of the Intelligent Vehicles Symposium.* 2008, pp. 67-72, IEEE.

[18] L. Ohno-Machado, *Modeling medical prognosis: survival analysis techniques.* J. Biomedical Informatics, 2002, 34(6): pp. 428-439, Biohealthmatics Centers.

[19] A. Bellaachia, and E. Guven, *Predicting breast cancer survivability using data mining techniques.* 2006, pp. 1-4, George Washington University.

[20] T. Jonsdottir, E. T. Hvannberg, H. Sigurdsson, and S. Sigurdsson, *The feasibility of constructing a predictive outcome model for breast cancer using the tools of data mining.* J. Expert Systems with Applications, 2008, 34(1): pp. 108-118, Pergamon Press, Inc. Tarrytown, NY, USA.

[21] S. Tsumoto, *Problems with mining medical data.* Twenty-Fourth Annual International Conference on Computer Software and Applications, 2000, pp. 467-468, IEEE Computer Society, Washington, D.C., USA.

[22] R. Barandelaa, J. S. Sánchez, V. Garcíaa, and E. Rangel, *Strategies for learning in class imbalance problems.* J. Pattern Recognition, 2003, 36(3): pp. 849-851, Elsevier Science B.V..

[23] J.-H. Lin, and P. J. Haug, *Exploiting missing clinical data in Bayesian network modeling for predicting medical problems.* J. Biomedical Informatics, 2007, 41(1): pp. 1-14, Elsevier Inc.

[24] R. Qahwaji, M. Al-Omari, T. Colak, and S. Ipson, *Using the real, gentle and modest AdaBoost learning algorithms to investigate the computerised associations between Coronal Mass Ejections and filaments.* in *Proceedings of the Mosharaka International Conference on Communications, Computers and Applications.* 2008, pp. 37-42, IEEE.

[25] C. E. Brodley, and M. A. Friedl, *Identifying and eliminating mislabeled training instances.* J. Artificial Intelligence Research, 1996, 1: pp. 799-805, CiteSeer.

[26] S. Verbaeten, and A. V. Assche, Ensemble methods for noise elimination in classification problems. in *Multiple Classifier Systems,* Vol. 2709, T. Windeatt, and F. Roli, Eds., 2003, pp. 317-325, Springer Berlin, Heidelberg.

[27] F. Melgani, and L. Bruzzone, *Classification of hyperspectral remote sensing images with support vector machines.* Transactions on Geoscience and Remote Sensing, 2004, 42(8): pp. 1778-1790, IEEE.

[28] C.-C. Chang, and C.-J. Lin, *LIBSVM -- a library for support vector machines.* http://www.csie.ntu.edu.tw/~cjlin/libsvm. 15 Feb 2007.

[29] K. Huang, J. Sun, Y. Hotta, K. Fujimoto, and S. Naoi, *An SVM-based high-accurate recognition approach for handwritten numerals by using difference features.* Ninth International Conference on Document Analysis and Recognition, 2007, 2: pp. 589-593, IEEE.

[30] T. M. Padmaja, N. Dhulipalla, R. S. Bapi, and P. R. Krishna, *Unbalanced data classification using extreme outlier elimination and sampling techniques for fraud detection.* International Conference on Machine Learning and Cybernetics on Advanced Computing and Communications, 2007, pp. 511-516, IEEE.

[31] Y. Ma, and X. Ding, *Robust real-time face detection based on cost-sensitive AdaBoost method.* in *Proceedings of the International Conference on Multimedia and Expo.* 2003, 2: pp. 465-473, IEEE.

[32] A. Vezhnevets, and V. Vezhnevets, *'Modest AdaBoost' - teaching AdaBoost to generalize better.* Graphicon-2005, 2005, pp., Novosibirsk Akademgorodok, Russia.

[33] G. Leshem, and Y. Ritov, *Traffic flow prediction using AdaBoost algorithm with random forests as a weak learner.* International Journal of Intelligent Technology, 2007, pp. 111-116.

[34] X. Li, L. Wang, and E. Sung, *AdaBoost with SVM-based component classifiers.* J. Engineering Applications of Artificial Intelligence, 2007, 21(5): pp. 785-795, Pergamon Press, Inc. Tarrytown, NY, USA

[35] H. M. Chang, and P. Gray, *Introduction to data mining and knowledge discovery.* in *Proceedings of the Thirty-First Hawaii International Conference on System Sciences.* 1998, 5: pp. 244-246, IEEE.

[36] S. L. Dale, and T. Bench-Capon, *A data mining tool for producing characteristic classifications in the legal domain.* in *Proceedings of the Eighth International Workshop on Database and Expert Systems Applications.* 1997, pp. 186-191, IEEE Computer Society, Washington, D.C., USA.

[37] M. Wong, W. Lam, K. Leung, P. Ngan, and J. Cheng, *Discovering knowledge from medical databases using evolutionary algorithms.* J. Engineering in Medicine and Biology Magazine, 2000, 19(4): pp. 45-55, IEEE.

[38] J. C. G. Ramirez, D. J. Cook, L. L. Peterson, and D. M. Peterson, *Temporal pattern discovery in course-of-disease data.* J. Engineering in Medicine and Biology Magazine, 2000, 19(4): pp. 63-71, IEEE.

[39] W. Yi, and W. Fuyong, *Breast cancer diagnosis via support vector machines.* Chinese Control Conference, 2006, pp. 1853-1856, IEEE.

[40] C.-L. Chang, *A study of applying data mining to early intervention for developmentally-delayed children.* J. Expert Systems with Applications, 2007, 33(2): pp. 407-412, Pergamon Press, Inc. Tarrytown, NY, USA.

[41] A. Tseng, I. Petrounias, and P. Chountas, *A complete framework for Web mining.* International Conference on Systems, Man and Cybernetics, 2003, 1: pp. 868-873, IEEE, Los Alamitos, USA.

[42] I.-H. Ting, *Web mining techniques for on-line social networks analysis.* International Conference on Service Systems and Service Management, 2008, pp. 1-5, IEEE.

[43] C. Best, *Web mining for open source intelligence.* in *Proceedings of the Twelfth International Conference Information Visualisation.* 2008, pp. 321-325, IEEE Computer Society, Washington, D.C., USA.

[44] S. K. Madria, C. Raymond, S. Bhowmick, and M. Mohania, *Association rules for web data mining in WHOWEDA*. Kyoto International Conference on Digital Libraries: Research and Practice, 2000, pp. 227-233, IEEE.

[45] C.-R. Yan, J.-Y. Shen, Q.-K. Peng, and D. Pan, *Parallel Web mining for link prediction in cluster server*. in *Proceedings of the International Conference on Machine Learning and Cybernetics*. 2005, 4: pp. 2291-2295 IEEE.

[46] M. W. Berry, *Survey of text mining : clustering, classification, and retrieval*. 2003, Springer, New York.

[47] A. Ekin, *Local information based overlaid text detection by classifier fusion*. in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*. 2006, 2: pp. 753-756, IEEE.

[48] M. Li, and R. C. Staunton, *Optimum Gabor filter design and local binary patterns for texture segmentation*. J. Pattern Recognition Letters, 2008, 29(5): pp. 664-672, Elsevier, Amsterdam, PAYS-BAS.

[49] J. Zhang, and Y. Yang, *Robustness of regularized linear classification methods in text categorization*. in *Proceedings of the Twenty-Sixth Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*. 2003, pp. 190 -197, ACM New York, NY, USA.

[50] A. Olukunle, and S. Ehikioya, *A fast algorithm for mining association rules in medical image data*. Canadian Conference on Electrical and Computer Engineering, 2002, 2: pp. 1181-1187, IEEE.

[51] G. Saarenvirta, *Data mining to improve profitability*. CMA Magazine, 1998, pp. 8-12.

[52] C. Shearer, *The CRISP-DM model: the new blueprint for data mining*. J. Data Warehousing, 2000, 5(4): pp. 13-22.

[53] J. Han, and M. Kamber, *Data mining: concepts and techniques*. 2nd. ed, 2006, Morgan Kaufmann, Elsevier Science, San Francisco.

[54] H. Thearling, *Information about data mining and analytic technologies*. http://www.thearling.com/.

[55] S. Matsumoto, Y. Kamei, A. Monden, and K.-i. Matsumoto, *Comparison of outlier detection methods in fault-proneness models*. The first International Symposium on Empirical Software Engineering and Measurement, 2007, pp. 461-463, ResearchGATE Scientific Network.

[56] D. L. Pham, *Unsupervised tissue classification in medical images using edge-adaptive clustering.* in *Proceedings of the Twenty-Fifth Annual International Conference on Engineering in Medicine and Biology Society.* 2003, 1: pp. 634-637, IEEE.

[57] M. Feist, R. M. McCourt, and H. Cui, *Unsupervised structure discovery for biodiversity information.* in *Proceedings of the Sixth ACM/IEEE-CS Joint Conference on Digital Libraries.* 2006, pp. 382-382, ACM New York, NY, USA.

[58] I. Buciu, C. Kotropoulos, and I. Pitas, *Demonstrating the stability of support vector machines for classification.* J. Signal Processing, 2006, 86(9): pp. 2364-2380, Elsevier North-Holland, Inc. Amsterdam, The Netherlands.

[59] O. Okun, and H. Priisalu, Random forest for gene expression based cancer classification: overlooked issues. in *Pattern Recognition and Image Analysis,* Vol. 4478, J. Marti, Ed., 2007, pp. 483–490, Springer Berlin / Heidelberg.

[60] V. Fragnelli, and S. Moretti, *A game theoretical approach to the classification problem in gene expression data analysis.* J. Computers and Mathematics with Applications, 2008, 55(5): pp. 950-959, Pergamon Press, Inc. Tarrytown, NY, USA.

[61] J. Han, J. Y. Chiang, S. Chee, J. Chen, Q. Chen, S. Cheng, W. Gong, M. Kamber, K. Koperski, G. Liu, Y. Lu, N. Stefanovic, L. Winstone, B. B. Xia, O. R. Zaiane, S. Zhang, and H. Zhu, *DBMiner: a system for data mining in relational databases and data warehouses.* in *Proceedings of the Conference of the Centre for Advanced Studies on Collaborative Research.* 1997, pp. 249-260, IBM Press.

[62] J. Meynet, and J.-P. Thiran, Information theoretic combination of classifiers with application to AdaBoost. in *Multiple Classifier Systems,* Vol. 4472, M. Haindl, J. Kittler, and F. Roli, Eds., 2007, pp. 171-179, Springer Berlin, Heidelberg.

[63] J. Platt, *Support vector machines.* http://research.microsoft.com/users/jplatt/svm.html.

[64] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining.* 2006, Pearson Addison Wesley, Boston.

[65] D. Delen, and N. Patil, *Knowledge extraction from prostate cancer data.* in *Proceedings of the Thirty-Ninth Annual Hawaii International Conference on System Sciences.* 2006, 5: pp. 1-10, IEEE.

[66] E. G. M. de Lacerda, A. C. P. L. F. de Carvalho, and T. B. Ludermir, *A study of cross-validation and bootstrap as objective functions for genetic algorithms.* in *Proceedings of the Seventh Brazilian Symposium on Neural Networks.* 2002, pp. 118-123, IEEE Computer Society, Washington, D.C., USA.

[67] R. Kohavi, *A study of cross-validation and bootstrap for accuracy estimation and model selection.* in *Proceedings of the International Joint Conference on Artificial Intelligence.* 1995, pp. 1137-1143, IJCAI.

[68] B. A. Flores, and J. A. Gonzalez, Data mining with decision trees and neural networks for calcification detection in mammograms. in *Advances in Artificial Intelligence,* Vol. 2972, R. Monroy, Ed., 2004, pp. 232–241, Springer Berlin / Heidelberg.

[69] P. Cabena, P. Hadjinian, R. Stadler, J. Verhees, and A. Zanasi, *Discovering data mining from concept to implementation.* 1998, Prentice Hall, Upper Saddle River, N.J.

[70] J. Kazmierska, and J. Malicki, *Application of the Naïve Bayesian classifier to optimize treatment decisions.* J. Radiotherapy and Oncology, 2008, 86(2): pp. 211-216, PubMed.

[71] S. Wang, C.-I. Chang, S.-C. Yang, G.-C. Hsu, H.-H. Hsu, P.-C. Chung, S.-M. Guo, and S.-K. Lee, *3D ROC analysis for medical imaging diagnosis.* Twenty-Seventh Annual International Conference of the Engineering in Medicine and Biology Society, 2005, pp. 7545-7548, IEEE.

[72] T. Fawcett, *ROC Graphs: notes and practical considerations for data mining researchers.* J. Intelligent Enterprise Technologies Laboratory, 2003, pp. 1-27, HP Laboratories Palo Alto.

[73] R. O. Duda, D. G. Stork, and P. E. Hart, *Pattern classification.* 2nd ed, 2001, Wiley, New York.

[74] S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, and D. Roth, *Generalization bounds for the area under the ROC curve.* J. Machine Learning Research, 2005, 6: pp. 393-425, JMLR.

[75] X. He, and E. C. Frey, *Three-class ROC analysis-the equal error utility assumption and the optimality of three-class ROC surface using the ideal observer.* IEEE Transactions on Medical Imaging, 2006, pp. 979-986, IEEE.

[76] K. Woods, and K. W. Bowyer, *Generating ROC curves for artificial neural networks.* IEEE Transactions on Medical Imaging, 1997, 16(3): pp. 329-337, IEEE.

[77] C. J. Biesheuvel, Y. Vergouwe, E. W. Steyerberg, D. E. Grobbee, and K. G. M. Moons, *Polytomous logistic regression analysis could be applied more often in diagnostic research.* J. Clinical Epidemiology, 2008, 61(2): pp. 125-134, PubMed.

[78] Y. Liao, L. W. Nolte, and L. M. Collins, *Decision fusion of ground-penetrating radar and metal detector algorithms-A Robust Approach.* IEEE Transactions on Geoscience and Remote Sensing, 2007, 45 (2): pp. 398-409, IEEE.

[79] G. I. Webb, and K. M. Ting, *On the application of ROC analysis to predict classification performance under varying class distributions* J. Machine Learning, 2005, 58(1): pp. 25–32, Springer Science and Business Media, Inc. Manufactured in The Netherlands.

[80] D. Hand, H. Mannila, and P. Smyth, *Principles of data mining.* 2001, The MIT Press, Cambridge, Massachusetts London, England.

[81] Y. Jiang, *Uncertainty in the output of artificial neural networks.* in *Proceedings of the International Joint Conference on Neural Networks.* 2007, pp. 2551-2556, IEEE Orlando, Florida, USA.

[82] J. Huang, and C. X. Ling, *Using AUC and accuracy in evaluating learning algorithms.* IEEE Transactions on Knowledge and Data Engineering, 2005, 7(3): pp. 299-310, IEEE.

[83] D. Nakache, E. Metais, and J. F. Timsit, Evaluation and NLP. in *Database and Expert Systems Applications,* Vol. 3588, K. V. Andersen, J. Debenham, and R. Wagner, Eds., 2005, pp. 626-632, Springer Berlin / Heidelberg.

[84] C. H. Li, and S. C. Park, Text categorization based on artificial neural networks. in *Neural Information Processing,* Vol. 4234, I. King, Ed., 2006, pp. 302–311, Springer-Verlag Berlin Heidelberg.

[85] D. R. Musicant, V. Kumar, and A. Ozgur, *Optimizing F-measure with Support Vector Machines.* in *Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference.* 2003, pp. 356-360, Haller AAAI Press.

[86] C. Schuster, and D. A. Smith, *Dispersion-weighted Kappa: An integrative framework for metric and nominal scale agreement coefficients.* J. Psychometrika, 2005, 70(1): pp. 135-146, Springer, 233 Spring Street, New York, NY.

[87] B. D. Eugenio, and M. Glass, *The Kappa statistic: a second look.* J. Computational Linguistics, 2004, 30(1): pp. 95 -101, MIT Press Cambridge, MA, USA.

[88] J. Landis, and G. Koch, *The measurement of observer agreement for categorical data.* J. Biometrics, 1977, 33(1): pp. 159-174, PubMed.

[89] N. O. B. Thomsen, L. H. Olsen, and S. T. Nielsen, *Kappa statistics in the assessment of observer variation: the significance of multiple observers classifying ankle fractures* J. Orthopaedic Science, 2002, 7(2): pp. 163-166, PubMed.

[90] J. Gao, R. Warren, H. Warren-Forward, and J. F. Forbes, *Reproducibility of visual assessment on mammographic density.* J. Breast Cancer Research and Treatment, 2007, 108(1): pp. 121-127, Springer, Dordrecht, PAYS-BAS.

[91] StatSoft Inc, *Data mining techniques*. http://www.statsoft.com/textbook/stdatmin.html.

[92] I. H. Witten, and E. Frank, *Data mining: practical machine learning tools and techniques*. 2 ed, 2005, Morgan Kaufmann, San Francisco.

[93] A. Berson, S. Smith, and K. Thearling, *An overview of data mining techniques*. http://www.thearling.com/text/dmtechniques/dmtechniques.htm. 06/07/2006.

[94] B. M. Thuraisingham, and M. G. Ceruti, *Understanding data mining and applying it to command, control, communications and intelligency enviroments*. Twenty-Fourth Annual International on Computer Software and Applications Conference, 2000, pp. 171-175, IEEE COMPSAC.

[95] S. Mitra, and T. Acharya, *Data mining : multimedia, soft computing, and bioinformatics*. 2003, Wiley Interscience.

[96] J. R. Quinlan, *C4.5: programs for machine learning*. 1993, Morgan Kaufmann, San Mateo, California.

[97] S. Ruggieri, *Efficient C4.5 [classification algorithm]*. IEEE Transactions on Knowledge and Data Engineering, 2002, 14 (2): pp. 438-444, IEEE.

[98] Z.-H. Zhou, and Y. Jiang, *Medical diagnosis with C4.5 rule preceded by artificial neural network ensemble*. IEEE Transactions on Information Technology in Biomedicine, 2003, 7(1): pp. 37-42, IEEE.

[99] Z. Yao, P. Liu, L. Lei, and J. Yin, *R-C4.5 decision tree model and its applications to health care dataset*. in *Proceedings of the International Conference on Services Systems and Services Management*. 2005, 2: pp. 1099-1103, IEEE-ICSSSM.

[100] P. He, L. Chen, and X.-H. Xu, *Fast C4.5*. International Conference on Machine Learning and Cybernetics, 2007, 5: pp. 2841-2846, IEEE.

[101] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and regression trees*. 1984, Belmont, Wadsworth.

[102] L. Breiman, *Statistical modeling: the two cultures*. J. Statistical Science, 2001, 16(3): pp. 199-231, The Institute of Mathematical Statistics.

[103] L. Breiman, *The heuristics of instability and stabilization in model selection*. J. Annals of Statistics, 1996, 24(6): pp. 2350-2383, The Institute of Mathematical Statistics.

[104] H. Chipman, E. I. George, and R. E. McCulloch, *Bayesian CART model search*. J. American Statistical Association, 1998, 93: pp. 935-960, CiteSeerX.

[105] R. Tibshirani, and K. Knight, *Model search and inference by bootstrap "bumping"*. Technical report, 1995, pp. 1-36, University of Toronto.

[106] G. H. John, and P. Langley, *Estimating continuous distributions in bayesian classifiers.* in *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence.* 1995, pp. 338-345, CiteSeerX.

[107] Y. Zhang, L. Zhang, J. Yan, and Z. Li, *Using association features to enhance the performance of Naive Bayes text classifier.* in *Proceedings of the Fifth International Conference on Computational Intelligence and Multimedia Applications.* 2003, pp. 336-341, IEEE.

[108] D. Aha, and D. Kibler, *Instance-based learning algorithms.* J. Machine Learning, 1991, 6(1): pp. 37-66, Springer Netherlands.

[109] Q. Hu, D. Yu, and Z. Xie, *Neighborhood classifiers.* J. Expert Systems with Applications, 2008, 34(2): pp. 866-876, Pergamon Press, Inc.

[110] B. Calvo, P. Larrañaga, and J. A. Lozano, *Learning Bayesian classifiers from positive and unlabeled examples.* J. Pattern Recognition Letters, 2007, 28(16): pp. 2375-2384, Elsevier Science Inc. New York, NY, USA.

[111] V. Vapnik, *Statistical learning theory.* 1998, Wiley, New York.

[112] V. Vapnik, *The nature of statistical learning theory.* 1995, Springer Verlag, New York.

[113] K. Coussement, and D. V. d. Poel, *Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques.* J. Expert Systems with Applications, 2008, 34(1): pp. 313-327, Pergamon Press, Inc. Tarrytown, NY, USA.

[114] P. Pulkkinen, and H. Koivisto, *Identification of interpretable and accurate fuzzy classifiers and function estimators with hybrid methods.* J. Applied Soft Computing, 2007, 7(2): pp. 520-533, Elsevier Science, Amsterdam, The Netherlands.

[115] H. Ishibuchi, T. Nakashima, and T. Murata, *A fuzzy classifier system that generates fuzzy if-then rules for pattern classification problems.* IEEE the International Conference on Evolutionary Computation, 1995, 2: pp. 759-764, IEEE.

[116] W. Yuan-Yuan, and H. Xue-Gang, *A fast algorithm for mining association rules based on concept lattice.* in *Proceedings of the International Conference on Machine Learning and Cybernetics.* 2004, 3: pp. 1687-1691, IEEE.

[117] C. Ordonez, E. Omiecinski, L. d. Braal, C. A. Santana, N. Ezquerra, J. A. Taboada, D. Cooke, E. Krawczynska, and E. V. Garcia, *Mining constrained association rules to predict heart disease.* in *Proceedings of the International Conference on Data Mining.* 2001, pp. 433-440, IEEE Computer Society, Washington, D.C., USA.

[118] A. M. Bensaid, N. Bouhouch, R. Bouhouch, R. Fellat, and R. Amri, *Classification of ECG patterns using fuzzy rules derived from ID3-induced decision trees.* Conference of the North American on Fuzzy Information Processing Society, 1998, pp. 34-38, IEEE.

[119] A. Tsakonas, G. Dounias, J. Jantzen, H. Axer, B. Bjerregaard, and D. G. von Keyserlingk, *Evolving rule-based systems in two medical domains using genetic programming.* J. Artificial Intelligence in Medicine, 2004, 32(3): pp. 195-216, Elsevier Science Publishers Ltd. Essex, UK.

[120] Oxford University Press, *Breast cancer.* 2000.

[121] Clinical Best Practice, *Breast Cancer in Australia: an overview.* http://www.nbcc.org.au/bestpractice/statistics/. Access Date : 7 August 2008.

[122] P. Thongsuksai, V. Chongsuvivatwong, and H. Sriplung, *Delay in breast cancer care: a study in Thai women.* J. Med Care, 2000, 38(1): pp. 108-114, PubMed.

[123] Breastcancer.org, *Stages of breast cancer.* http://www.breastcancer.org/dia_pict_ staging.html. Access Date : 20 June 2006.

[124] CIN, *General cancer information.* http://patient.cancerconsultants.com/centers.aspx?tierID=827&linkID. Access Data : 26 May 2006.

[125] C. Andreetta, and I. Smith, *Adjuvant endocrine therapy for early breast cancer.* J. Cancer Letters, 2007, 251(1): pp. 17-27, Elsevier Ireland Ltd.

[126] E. D. Parker, and A. R. Folsom, *Intentional weight loss and incidence of obesity related cancer: the Iowa women's health study.* J. Nature, 2003, 27: pp. 1447-1452, Nature Publishing Group.

[127] L. Wasserman, S. Flatt, L. Natarajan, G. Laughlin, M. Matusalem, S. Faerber, C. Rock, E. Barrett-Connor, and J. Pierce, *Correlates of obesity in postmenopausal women with breast cancer: comparison of genetic, demographic, disease-related, life history and dietary factors.* J. Nature, 2004, pp. 49-56, Nature Publishing Group.

[128] H. B. Burke, D. B. Rosen, and P. H. Goodman, *Comparing artificial neural networks to other statistical methods for medical outcome prediction.* IEEE International Conference on Neural Networks, 1994, 4: pp. 2213-2216, IEEE.

[129] H. Brenner, O. Gefeller, and T. Hakulinen, *A computer program for period analysis of cancer patient survival.* J. Cancer European, 2002, 38(5): pp. 690-695, PubMed.

[130] H. B. Burke, P. H. Goodman, D. B. Rosen, D. E. Henson, J. N. Weinstein, F. E. H. Jr., J. R. Marks, D. P. Winchester, and D. G. Bostwick, *Artificial neural networks improve the accuracy of cancer survival prediction* J. Cancer, 1997, 79: pp. 857-862, PubMed.

[131] S. A. Boorjian, P. L. Crispen, C. M. Lohse, B. C. Leibovich, and M. L. Blute, *The Impact of temporal presentation on clinical and pathological outcomes for patients with sporadic bilateral renal masses.* J. European Urology, 2008, 54(4): pp. 855-865, Elsevier B.V. .

[132] D. J. D'Ambrosio, K. Ruth, E. M. Horwitz, D. Y. T. Chen, A. Pollack, and M. K. Buyyounouski, *Does transurethral resection of prostate (TURP) affect outcome in patients who subsequently develop prostate cancer?* J. Urology, 2008, 71(5): pp. 938-941, PubMed.

[133] T. Hastie, R. Tibshirani, and J. H. Friedman, *The elements of statistical learning : data mining, inference, and prediction.* 2001, Springer, New York.

[134] J. R. Mertens, A. J. Flisher, D. D. Satre, and C. M. Weisner, *The role of medical conditions and primary care services in 5-year substance use outcomes among chemical dependency treatment patients.* J. Drug and Alcohol Dependence, 2008, 98(1-2): pp. 45-53, PubMed.

[135] A. G. Heidema, and N. Nagelkerke, *Developing a discrimination rule between breast cancer patients and controls using proteomics mass spectrometric data: a three-step approach.* J. Statistical Applications in Genetics and Molecular Biology, 2007, 7(2): pp. Articles 5, PubMed.

[136] X. Xiong, Y. Kim, Y. Baek, D. W. Rhee, and S.-H. Kim., *Analysis of breast cancer using data mining & statistical techniques.* Sixth International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing and the First ACIS International Workshop on Self-Assembling Wireless Networks, 2005, pp. 82-87, IEEE.

[137] B. Kovalerchuk, E. Vityaev, and J. F. Ruiz, *Consistent knowledge discovery in medical diagnosis.* J. Engineering in Medicine and Biology Magazine, 2000, 19(4): pp. 26-37, IEEE.

[138] R. Kohli, R. Krishnamurti, and K. Jedidi, *Subset-conjunctive rules for breast cancer diagnosis.* J. Discrete Applied Mathematics, 2006, 154(7): pp. 1100-1112, Elsevier Science Publishers B. V. Amsterdam, The Netherlands.

[139] N. Bundred, *Prognostic and predictive factors in breast cancer.* J. Cancer Treat Rev., 2001, 27(3): pp. 137-142, PubMed.

[140] P. H. Elkhuizen, B. Kreike, and M. J. v. d. Vijver, Risk functions for local recurrence following conversation therapy in breast cancer. in *Prognostic and predictive factors in*

breast cancer, R. A. Walker, and A. M. Thompson, Eds., 2008, pp. 38-52, Informa Health Care,New York.

[141] M. Lundina, J. Lundina, H. B. Burked, S. Toikkanenb, L. Pylkkänenc, and H. Joensuua, *Artificial neural networks applied to survival prediction in breast cancer.* J. International Journal for Cancer Research and Treatment, 1999, 57(4): pp. 281-286, Oncology.

[142] C.-Y. Wang, C.-G. Wu, Y.-C. Liang, and X.-C. Guo, *Diagnosis of breast cancer tumor based on ICA and LS-SVM.* IEEE International Conference on Machine Learning and Cybernetics, 2006, pp. 2565-2570, IEEE.

[143] G. Richards, V. J. Rayward-Smith, P. H. Sonksen, S. Carey, and C. Weng, *Data mining for indicators of early mortality in a database of clinical records.* J. Artificial Intelligence in Medicine, 2001, 22(3): pp. 215-231.

[144] A. Mussa, and M. Tshilidzi, *The use of genetic algorithms and neural networks to approximate missing data in database.* IEEE Third International Conference on Computational Cybernetics, 2005, pp. 207-212, IEEE.

[145] P. Liu, L. Lei, and N. Wu, *A quantitative study of the effect of missing data in classifiers.* The fifth International Conference on Computer and Information Technology, 2005, pp. 28-33, IEEE.

[146] V. J. Hodge, and J. Austin, *A survey of outlier detection methodologies.* J. Artificial Intelligence, 2004, 22(2): pp. 85-126, Kluwer Academic Publishers Norwell, MA, USA.

[147] V. Podgorelec, M. Hericko, and I. Rozman, *Improving mining of medical data by outliers prediction.* in *Proceedings of the Eighteenth IEEE Symposium on Computer-Based Medical Systems.* 2005, pp. 91-96, IEEE.

[148] J. Thongkam, G. Xu, Y. Zhang, and F. Huang, *Support vector machines for outlier detection in cancer survivability prediction.* International Workshop on Health Data Management, 2008, pp. 99-109, APWeb'08.

[149] C. E. Brodley, and M. A. Friedl, *Identifying mislabeled training data.* J. Artificial Intelligence Research, 1999, 11: pp. 131-167, CiteSeerX.

[150] C. M. Teng, *Applying noise handling techniques to genomic data: a cases study.* Third IEEE International Conference on Data Mining, 2003, pp. 743-746, IEEE.

[151] Z.-H. Zhou, and Y. Jiang, *NeC 4.5: neural ensemble based C4.5.* IEEE Transactions on Knowledge and Data Engineering, 2004, 16 (6): pp. 770-773, IEEE.

[152] D. Q. Huynh, R. Hartley, and A. Heyden, *Outlier correction in image sequences for the affine camera.* in *Proceedings of the Ninth IEEE International Conference on Computer Vision.* 2003, 2: pp. 585-590, IEEE Computer Society Washington, D.C., USA.

[153] P. M. T. Broersen, *ARMAsel for detection and correction of outliers in univariate stochastic data.* IEEE Transactions on Instrumentation and Measurement, 2008, 57(3): pp. 446-453, IEEE.

[154] G. H. John, *Robust decision trees: removing outliers from databases.* in *Proceedings of the First International Conference on Knowledge Discovery and Data Mining.* 1995, pp. 174-179, AAAI Press.

[155] D. Gamberger, *A minimization approach to propositional inductive learning.* in *Proceedings of the Eighth European Conference on Machine Learning.* 1995, 912: pp. 151-160, Springer-Verlag London, UK.

[156] J. Xie, and Z. Qiu, *The effect of imbalanced data sets on LDA: A theoretical and empirical analysis.* J. Pattern Recognition Society 2007, 40(2): pp. 557-562, Elsevier B.V.

[157] G. He, H. Han, and W. Wang, *An over-sampling expert system for learing from imbalanced data sets.* International Conference on Neural Networks and Brain, 2005, 1: pp. 537-541, IEEE.

[158] R. Alejo, V. Garcia, J. M. Sotoca, R. A. Mollineda, and J. S. Sánchez, Improving the classification accuracy of RBF and MLP neural networks trained with imbalanced samples. in *Intelligent Data Engineering and Automated Learning,* Vol. 4224, E. Corchado, Ed., 2006, pp. 464-471, Springer Berlin / Heidelberg.

[159] A. Estabrooks, T. Jo, and N. Japkowicz, *A multiple resampling method for learning from imbalanced data sets.* J. Computational Intelligence, 2004, 20(1): pp. 18-36, Wiley InterScience.

[160] J. Wang, M. Xu, H. Wang, and J. Zhang, *Classification of imbalanced data by using the SMOTE algorithm and locally linear embedding.* in *Proceedings of the Eighth International Conference on Signal Processing.* 2006, 3: pp. 1-4, IEEE.

[161] L. Pelayo, and S. Dick, *Applying novel resampling strategies to software defect prediction.* Annual Meeting of the North American Fuzzy Information Processing Society, 2007, pp. 69-72, IEEE.

[162] J.-P. Renno, D. Makris, and G. A. Jones, *Object classification in visual surveillance using AdaBoost.* IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1-8, IEEE.

[163] M. Zhou, and H. Wei, *Face verification using GaborWavelets and AdaBoost.* Eighteenth International Conference on Pattern Recognition, 2006, 1: pp. 404-407, IEEE.

[164] S. Jinbo, X. Li, and L. Wenhuang, *The application ofAdaBoost in customer churn prediction.* International Conference on Service Systems and Service Management, 2007, pp. 1-6, IEEE.

[165] X. Mantao, and P. Franti, *A heuristic K-means clustering algorithm by kernel PCA.* in *Proceedings of the International Conference on Image.* 2004, 5: pp. 3503-3506, IEEE.

[166] P. Heum, K. Young-Gi, and K. Hyuk-Chul, *A feature selection for Korean Web document clustering.* Thirtieth Annual Conference of IEEE Industrial Electronics Society, 2004, 3: pp. 2650-2654, IEEE.

[167] C. Ordonez, *Clustering binary data streams with K-means.* in *Proceedings of the Eighth ACM SIGMOD Workshop on Research issues in Data Mining and Knowledge Discovery.* 2003, pp. 12-19, ACM New York, NY, USA.

[168] P. Sun, Y. Ma, Y. Wei, Z. Ma, L. Lu, Y. Cui, and P. Huang, *Application of improved K-mean clustering in predicting protein-protein interactions.* International Conference on BioMedical Engineering and Informatics, 2008, 1: pp. 83-86, IEEE.

[169] X.-L. Xia, M. R. Lyu, T.-M. Lok, and G.-B. Hyabg, Methods of decreasing the number of support vectors via κ-mean clustering. in *International Conference on Intelligent Computing,* Vol. 3644, D. S. Huang, X.-P. Zhang, and G.-B. Huang, Eds., 2005, pp. 717-726, Springer-Veriag Berlin Heidelberg.

[170] F. Liu, *An attribute selection approach and Its application.* International Conference on Neural Networks and Brain, 2005, 2: pp. 636-640, IEEE.

[171] A. Kalousis, J. Prados, and M. Hilario, *Stability of feature selection algorithms.* in *Proceedings of the Fifth IEEE International Conference on Data Mining.* 2005, pp. 1-8, IEEE.

[172] T. G. Dietterich, *An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization.* J. Machine Learning, 1999, 40(2): pp. 1-22, Kluwer Academic, Boston. Manufactured in The Netherlands.

[173] K. Kira, and L. A. Rendell, *A practical approach to feature selection.* in *Proceedings of the Ninth International Conference on Machine Learning.* 1992, pp. 249-256, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA.

[174] K. Krishna, and M. Narasimha Murty, *Genetic K-means algorithm.* J. IEEE Transactions on Systems, Man and Cybernetics, 1999, 29(3): pp. 433-439, IEEE.

[175] Q. Zhang, K. J. Lee, and T. K. Whangbo, *K-mean and double cross-validation algorithm for LS-SVM in Sasang Typology classification.* IEEE International Conference on Granular Computing Automation and Logistics, 2007, pp. 426-430, IEEE.

[176] W. Tang, and T. M. Khoshgoftaar, *Noise identification with the k-means algorithm.* Sixteenth International Conference on Tools with Artificial Intelligence, 2004, pp. 373-378, IEEE.

[177] Y. Sun, and J. Li, *Iterative RELIEF for feature weighting.* in *Proceedings of the Twenty-Third International Conference on Machine Learning.* 2006, 148: pp. 913-920, ACM New York, NY, USA.

[178] M. A. Hall, and G. Holmes, *Benchmarking attribute selection techniques for discrete class data mining.* IEEE Transactions on Knowledge and Data Engineering, 2003, 15 (6): pp. 1437-1447, IEEE.

[179] R. E. Schapire, and Y. Singer, *Improved boosting algorithms using confidence-rated predictions.* J. Machine Learning, 1999, 37(3): pp. 297-336, Springer Netherlands.

[180] R. E. Schapire, *A brief introduction to boosting.* in *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence* 1999, pp. 1401-1405, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA.

[181] E. Bauer, and R. Kohavi, *An empirical comparison of voting classification algorithms: bagging, boosting and variants.* J. Machine Learning, 1999, 36(1-2): pp. 105-139, Springer Netherlands.

[182] D. Mease, A. J. Wyner, and A. Buja, *Boosted classification trees and class probability/quantile estimation* J. Machine Learning Research, 2007, 8: pp. 409-439, MIT Press Cambridge, MA, USA.

[183] J. Friedman, T. Hastie, and R. Tibshirani, *Additive logistic regression: A statistical view of boosting.* J. Annals of Statistics, 2000, 38(2): pp. 337-374, Institute of Mathematical Statistics.

[184] E. Leon, G. Clarke, F. Sepulveda, and V. Callaghan, *Optimised attribute selection for emotion classification using physiological signals.* Twenty-Sixth Annual International Conference on Engineering in Medicine and Biology Society, 2004, 1: pp. 184-187, IEEE.

[185] W. Jin-Feng, W. Xi-Zhao, and H. Ming-Hu, *Attribute selection's impact on robustness of decision trees.* in *Proceedings of the International Conference on Machine Learning and Cybernetics.* 2002, 4: pp. 1829-1832, IEEE.

[186] J. Thongkam, G. Xu, Y. Zhang, and F. Huang, *Breast cancer survivability via AdaBoost algorithms.* Australiasian Workshop on Health Data and Knowledge Management, 2008, 80: pp. 1-10, Australian Computer Society, Inc. Darlinghurst, Australia.

[187] J. Thongkam, G. Xu, and Y. Zhang, *An analysis of data selection methods on classifiers accuracy measures.* J. KKU. Engineering Journal, 2008, 35(1): pp. 1-10, Khon Kaen University.

[188] A. Vezhnevets, *GML AdaBoost MATLAB toolbox.* http://research.graphicon.ru/ machine-learning/gml-adaboost-matlab-toolbox.html. 24 Feb 2007.

[189] H. B. Borges, and J. C. Nievola, *Attribute selection methods comparison for classification of diffuse large B-Cell Lymphoma.* in *Proceedings of the Fourth International Conference on Machine Learning and Applications.* 2005, pp. 1-6, IEEE.

[190] Z. Yin, P. Yin, F. Sun, and H. Wu, *A writer recognition approach based on SVM.* Multi Conference on Computational Engineering in Systems Applications, 2006, 1: pp. 581-586, IEEE.

[191] C. J. C. Burges, *A tutorial on support vector machines for pattern recognition.* J. Data Mining and Knowledge Discovery, 1998, 2: pp. 121-167, Kluwer Academic Publishers Hingham, MA, USA.

[192] D. Gamberger, T. Šmuc, and I. Marić, *Noise detection and elimination in data preprocessing experiments in medical domains.* J. Applied Artificial Intelligence, 2000, 14(2): pp. 205-223, Informaworld.

[193] S. Lallich, F. Muhlenbach, and D. A. Zighed, *Improving classification by removing or relabeling mislabeled instances.* in *Proceedings of the Thirteenth International Symposium on Foundations of Intelligent Systems.* 2002, 2366: pp. 5-15, Springer-Verlag London, UK.

[194] J.-w. Sun, F.-y. Zhao, C.-j. Wang, and S.-f. Chen, *Identifying and correcting mislabeled training instances.* Future Generation Communication and Networking, 2007, 1: pp. 244-250, IEEE.

[195] T. M. Khoshgoftaar, N. Seliya, and K. Gao, *Rule-based noise detection for software measurement data.* in *Proceedings of the IEEE International Conference on Information Reuse and Integration.* 2004, pp. 302-307, IEEE.

[196] D. Toth, and T. Aach, *Improved minimum distance classification with Gaussian outlier detection for industrial inspection.* in *Proceedings of the Eleventh International Conference on Image Analysis and Processing.* 2001, pp. 584-588, IEEE.

[197] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, *SMOTE: synthetic minority over-sampling technique.* J. Artificial Intelligence and Research, 2002, 16: pp. 321-357, AI Access Foundation and Morgan Kaufmann Publishers.

[198] R. Moffitt, J. Phan, S. Hemby, and M. Wang, *Effect of outlier removal on gene marker selection using support vector machines.* Twenty-Seventh Annual International Conference of the Engineering in Medicine and Biology Society, 2005, pp. 917-920, IEEE.

[199] D. A. Marquez, J. L. Paredes, and W. Garcia-Gabin, *Nonlinear filters based on support vector machines.* International Conference on Acoustics, Speech and Signal Processing, 2007, 2: pp. 581-584, IEEE.

[200] G. M. Weiss, and F. Provost, *Learning when training data are costly: the effect of class distribution on tree induction.* J. Artificial Intelligence Research, 2003, 19: pp. 315-354, AAAI.

[201] J. Thongkam, G. Xu, Y. Zhang, and F. Huang, *Toward breast cancer survivability prediction models through improving training space, (on-line doi:10.1016/j.physleth.2003.10.071).* J. Expert System with Application, 2009, pp. 1-14.

[202] P. L. Bartlett, and M. Traskin, *AdaBoost is consistent.* J. Machine Learning Research, 2007, 8: pp. 2347-2368, MIT Press Cambridge, MA, USA.

[203] T. T. Truyen, D. Q. Phung, S. Venkatesh, and H. H. Bui, *AdaBoost.MRF: Boosted Markov random forests and application to multilevel activity recognition.* Computer Society Conference on Computer Vision and Pattern Recognition, 2006, 2: pp. 1686-1693, IEEE.

[204] Y. Sun, Y. Wang, and A. K. C. Wong, *Boosting an associative classifier.* IEEE Transactions on Knowledge and Data Engineering, 2006, 18 (7): pp. 988-992, IEEE.

[205] G. Szarvas, R. Farkas, and A. Kocsor, A multilingual named entity recognition system using boosting and C4.5 decision tree learning algorithms in *Lecture Notes in Computer Science and Discovery Science,* Vol. 4265, N. Lavrac, L. Todorovski, and K. P. Jantke, Eds., 2006, pp. 267-278, Springer-Varlag Berlin Heidelderg.

[206] Y. Freund, and R. E. Schapire, *Experiments with a new boosting algorithm.* in *Proceedings of the Thirteenth International Conference on Machine Learning.* 1996, pp. 148-156, Center for Education and Research in Information Assurance and Security.

[207] W. T. Ho, and Y. H. Tay, *On detecting spatially similar and dissimilar objects using AdaBoost.* International Symposium on Information Technology, 2008, 2: pp. 1-5, IEEE.

[208] L. Breiman, *Random Forests.* J. Machine Learning 2001, 45: pp. 5–32, Kluwer Academic Publishers. Manufactured in The Netherlands.

[209] N. Meinshausen, *Quantile regression forests.* J. Machine Learning Research, 2006, 7: pp. 983–999, MIT Press Cambridge, MA, USA

[210] D. S. Kim, S. M. Lee, and J. S. Park, Building lightweight intrusion detection system based on random forest. in *Advances in Neural Networks,* Vol. 3973, J. Wang, Ed., 2006, pp. 224-230, Springer-Verlag Berlin Heidelberg.

[211] J. C.-W. Chan, and D. Paelinckx, *Evaluation of random forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery.* J. Remote Sensing of Environment, 2008, 112(6): pp. 2999-3011, Elsevier Inc.

[212] C. Nadeau, and Y. Bengio, *Inference for the generalization error.* J. Machine Learning, 2003, 52(3): pp. 239-281, Springer Netherlands.

[213] D. F. Nettleton, L. Calderon-Benavides, and R. Baeza-Yates, *Analysis of Web search engine clicked documents.* Fourth Latin American Web Congress, 2006, pp. 209-219, IEEE.

Appendices

Details of attributes in Table 3.1 are shown in the following appendices.

Appendix A.1: Marital status

No.	Value	Description
1	1	Single
2	2	Married
3	3	Non
4	9	Unknown

Appendix A.2: Occupation

No.	Value	Description (translated from the Thai description)
1	0	Other
2	100	Dressmaker
3	106	Doctor
4	108	Dentist
5	110	Lecturer
6	111	Teacher
7	112	Nurse
8	201	Officer
9	205	Police
10	207	Officer
11	208	Elderly with pension
12	209	Temporary Employee
13	210	Permanent Employee
14	217	Elderly with lumsum
15	302	Officer in a government corporation
16	303	Operator in a government corporation
17	401	Business Owner
18	402	Officer in a private company
19	403	Worker
20	502	General Framer
21	503	Grain farmer
22	505	Grain and truck farmer
23	606	Seller
24	719	Hairdresser
25	818	Building Worker
26	901	Priest
27	902	officer in a government sector
28	907	Elderly
29	909	Student
30	910	University Student
31	911	House wife
32	993	Unemployed
33	996	General worker

Appendix A.3: Basis of diagnosis

No.	Value	Description
1	1	History & Physical exam.
2	2	Endoscopy & Radiology
3	3	Surgery & Autopsy (no histol.)
4	4	Specific Biochem/Immuno tests
5	5	Cytology or Hematology
6	6	Histology of Metastasis
7	7	Histology of Primary

Appendix A.4: Topography

No.	Value	Description
1	500	C50.0 Nipple
2	501	C50.1 Central portion of breast
3	502	C50.2 Upper-inner quadrant of b
4	503	C50.3 Lower-inner quadrant of b
5	504	C50.4 Upper-outer quadrant of b
6	505	C50.5 Lower-outer quadrant of b
7	506	C50.6 Axillary tail of breast
8	508	C50.8 Overl. lesion of breast
9	509	C50.9 Breast, NOS

Appendix A.5: Morphology

No.	Value	Description
1	8000	Neoplasm
2	8001	Tumour cells
3	8010	Epithelial tumour
4	8012	Large cell carcinoma, OS
5	8020	Carcinoma, differentiated, OS
6	8021	Carcinoma, anaplastic, NOS
7	8033	Pseudosarcomatous carcinoma
8	8041	Small cell carcinoma, NOS
9	8050	Papillary carcinoma, NOS
10	8070	Squamous cell carcinoma, NOS
11	8140	Adenocarcinoma, NOS
12	8141	Scirrhous adenocarcinoma
13	8200	Adenoid cystic carcinoma, NOS
14	8260	Papillary adenocarcinoma, NOS
15	8310	Clear cell adenocarcinoma, NOS
16	8480	Mucinous adenocarcinoma
17	8481	Mucin-producing adenocarcinoma
18	8490	Signet ring cell carcinoma
19	8500	Infiltrating duct carcinoma
20	8501	Comedocarcinoma, NOS
21	8510	Medullary carcinoma, NOS
22	8512	Medullary carc. with lymph. stroma
23	8513	Atypical medullary carcinoma
24	8520	Lobular carcinoma,NOS
25	8521	Infiltrating ductular carcinoma
26	8522	Infiltrating duct and lobular carcinoma

Appendix A.5: Morphology (con't)

No.	Value	Description
27	8523	Infiltrating duct mixed with other types
28	8530	Inflammatory carcinoma
29	8540	Paget's disease, mammary
30	8541	P. dis. & infil. duct carc.,breast
31	8543	Paget's disease and intraduct. ca. of br
32	8800	Soft tissue tumour
33	8900	Rhabdomyosarcoma, NOS
34	8930	Endometrial stromal sarcoma
35	9020	phyllodes tumour
36	9120	Hemangiosarcoma
37	9591	Lmphoma,non-Hodgkin's, NOS

Appendix A.6: Extent

No.	Value	Description
1	1	In situ
2	2	Localized
3	3	Direct extension
4	4	Regional lymph nodes
5	5	Distant metastases
6	8	Not applicable
7	9	Not known

Appendix A.7: Stage

No.	Value	Description
1	0	Stage 0
2	1	Stage I, A
3	2	Stage II, B
4	3	Stage III, C
5	4	Stage IV, D
7	9	Unknown

Appendix A.8: Received treatments including surgery, radiation, chemotherapy, hormonal therapy, immunotherapy, other therapy and supportive therapy.

No.	Value	Description
1	1	Received treatment
2	2	Do not Received treatment
3	9	Unknown

FACULTY OF MEDICINE
KHON KAEN UNIVERSITY

Our ref: K.K.U.0514.7.13.1/6526

December 19, 2006

Professor Yanchun ZHANG
Principal Supervisor, Director, ITArl
School of Computer Science & Mathematics
Victoria University, PO Box 14428
Melbourne City MC VIC 8001
AUSTRALIA
Telephone: +61-03-9919-5060, Facsimile: +61-03-9919-4050
E-mail: Yanchun.Zhang@vu.edu.au

Dear Professor Yanchun Zhang:
Re: Permission letter for obtaining the database

Greetings from the Faculty of Medicine, Khon Kaen University.

I am writing to convey official permission to **Ms. Jaree THONGKAM**, a Master degree student at the School of Computer Science & Mathematics, Victoria University, to obtain the database, comprising breast cancer patients between January 1990 and December 2005 seen at Srinagarind Hospital, Faculty of Medicine, Khon Kaen University.

I trust that this permission for data support will allow for a deeper and richer research experience for Ms. THONGKAM. Thank you very much for your cooperation. Should you have any suggestions, please do not hesitate to contact us.

Sincerely,

Vinai TANTIYASAWASDIKUL, MD
Director, Srinagarind Hospital

Cc: 1. Professor Somporn POTHINAM, MD, Dean, Faculty of Medicine, Mahasarakam University
2. Head, Cancer Unit, Srinagarind Hospital