# BUILDING MINIMAL CLASSIFICATION RULES FOR WISCONSIN BREAST CANCER DATASET

PHONETHEP DOUANGNOULACK

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ENGINEERING IN COMPUTING IN ENGINEERING SYSTEMS

INTERNATIONAL COLLEGE

KING MONGKUTS INSTITUTE OF TECHNOLOGY LADKRABANG

ACADEMIC YEAR 2017

KMITL-2017-IC-M-11-07

# BUILDING MINIMAL CLASSIFICATION RULES FOR WISCONSIN BREAST CANCER DATASET

PHONETHEP DOUANGNOULACK

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ENGINEERING IN COMPUTING IN ENGINEERING SYSTEMS

INTERNATIONAL COLLEGE

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

ACADEMIC YEAR 2017

KMITL-2017-IC-M-11-07

**COPYRIGHT ACADEMIC YEAR 2017**

**INTERNATIONAL COLLEGE**

**KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG**

| | |
|---|---|
| Thesis Title | Building Minimal Classification Rules for Wisconsin Breast Cancer Dataset |
| Student | Phonethep Douangnoulack |
| Student ID | 59610023 |
| Degree | Master of Engineering |
| Program | Computing In Engineering Systems |
| Thesis Advisor | Assoc.Prof.Dr. Veera Boonjing |

# ABSTRACT

A rule-based classifier has been widely applied in the breast cancer diagnosis. The classifiers with a good performance of disease classification have been developed and highly required over the past decades. Since classification rules are derived from previous diagnosis with a large number of features, it is challenging to build a minimal number of rules with high performance while retaining all diagnosis information. In order to solve the problems, Principal Component Analysis (PCA) was used to reduce the number of features. It is known as a lossless data reduction technique with good classification performance. Therefore, this research aims at finding the best performance classifier giving minimal classification rules by employing PCA. Based on the overall experiment on three Wisconsin data sets (WBC, WDBC and WPBC), in term of accuracy, J48 decision tree classifier is found to be the best among the three classifiers: J48 decision tree, Reduced Error Pruning Tree, and Random Tree.

# ACKNOWLEDGEMENTS

The author would like to be grateful to International College, King Mongkuts Institute of Technology Lardkrabang for awarding the master studentship to conduct the research activities that produced the results disseminated in this thesis. First of all, the author wishes to particularly thank supervisor Assoc. Prof. Dr. Veera Boonjing, who supplied me with invaluable advice and guidance throughout my time during the research and thesis. Without him, the author will never be able to complete the work with this ease. In addition, the author also needs to give thanks to Asst. Prof. Dr. Chaiwat Nuthong for supporting the facilitated my involvement in the writing for this thesis and suggestion from Dr. Ukrit Watchareeruetai as well as Dr. William H.Wolberg at the University of Wisconsin for providing breast cancer data set which used in this thesis and thankful to all my colleagues for helping. Finally, the author would also like to express gratitude to my family back in Laos. This thesis would not have been possible without their big support, inspirational, and encouragement.

Bangkok, July 2018

Phonethep Douangnoulack

# TABLE OF CONTENTS

Page

**ABSTRACT** . . . .

**ACKNOWLEDGEMENTS** . . . .

**TABLE OF CONTENTS** . . . . i

**LIST OF FIGURES** . . . . iii

**LIST OF TABLES** . . . . v

**CHAPTER**

**1 Introduction** . . . . 1

1.1 Background and Problem Statement . . . . 1

1.2 Objectives . . . . 1

1.3 Scope . . . . 2

1.4 Thesis Organization . . . . 2

**2 Literature Review** . . . . 3

2.1 Feature Reduction . . . . 3

2.2 Classification . . . . 7

**3 Research Methodology** . . . . 10

3.1 Data Collection . . . . 10

3.2 Data Preprocessing . . . . 10

3.3 Principal Component Analysis . . . . 11

3.3.1 Data set . . . . 11

3.3.2 Subtract the means . . . . 12

3.3.3 Calculate the covariance matrix . . . . 13

3.3.4 Compute the Eigenvalues and Eigenvector . . . . 14

3.3.5 Selecting the principal components . . . . 14

3.3.6 Deriving the new data set . . . . 16

3.4 Applying Machine Learning Techniques . . . . 17

3.5 Decision Tree . . . . 18

3.5.1 J48 Decision Tree . . . 18
3.5.2 Reduce Error Pruning Tree . . . 22
3.5.3 Random Tree . . . 24
3.6 Cross Validation . . . 26
3.7 Evaluation Performance . . . 27
3.8 Software Development Tool . . . 28
**4 Discussion and Experimental Results** . . . 29
4.1 Experimental setup . . . 29
4.2 Experimental Results . . . 29
4.2.1 Performance results of WBC data set . . . 29
4.2.2 Performance results of WDBC data set . . . 33
4.2.3 performance results of WPBC data set . . . 35
4.3 Discussion . . . 39
**5 Conclusion and Recommendation** . . . 41
**REFERENCES** . . . 42
**APPENDICES** . . . 45
**APPENDIX A Data set** . . . 45
**APPENDIX B Eigenvalues** . . . 47
**APPENDIX C Classification Rules of Wisconsin data set** . . . 49
**APPENDIX D Publications** . . . 61
**AUTHOR BIOGRAPHY** . . . 67

# LIST OF FIGURES


| Figure | | Page |
|---|---|---|
| 1.1 | The study outline | 2 |
| 2.1 | Flowchart of the Breast Cancer prognosis model | 3 |
| 2.2 | The Basic flow of proposed method | 4 |
| 2.3 | The flowchart of PCA and SFS approach | 5 |
| 2.4 | The performance of two Linear and Quadratic classifiers | 6 |
| 2.5 | An example of classification task | 7 |
| 2.6 | The process of experiment protocol | 9 |
| 3.1 | The graphical representation of bivariate data | 12 |
| 3.2 | The graphical representation of adjusted bivariate data | 13 |
| 3.3 | The graphical of Scree test | 15 |
| 3.4 | The graphical representation of two Eigenvectors | 16 |
| 3.5 | An example of decision tree model | 18 |
| 3.6 | The partially learned decision tree from the first step of J48 | 22 |
| 3.7 | The representation of subtree in REP Tree | 23 |
| 3.8 | An example of Reduced Error Pruning Tree | 24 |
| 3.9 | An example of training model for Random tree | 25 |
| 3.10 | An example of testing model for Random tree | 25 |
| 3.11 | The example of three-fold cross-validation method | 26 |
| 4.1 | The correlation matrix of WBC data set | 30 |
| 4.2 | The comparison of classification accuracy of WBC data set | 31 |
| 4.3 | The comparison of classification F-measure of WBC data set | 31 |
| 4.4 | The comparison of classification rules of WBC | 32 |
| 4.5 | The correlation matrix of WDBC data set | 33 |
| 4.6 | The comparison of classification accuracy of WDBC | 34 |
| 4.7 | The comparison of classification F-measure of WDBC | 34 |
| 4.8 | The comparison of classification rules of WDBC | 35 |
| 4.9 | The eigenvalues of WDBC data set | 36 |
| 4.10 | The correlation matrix of WDBC data set | 36 |
| 4.11 | The comparison of classification accuracy of WPBC | 37 |

Figure Page

4.12 The comparison of classification F-measure of WPBC . . . . . 38
4.13 The comparison of classification rules of WPBC . . . . . 38
C.1 The rules of J48 decision tree without PCA in WBC data set . . . . . 50
C.2 The rules of REP tree without PCA in WBC data set . . . . . 50
C.3 The rules of Random tree without PCA in WBC data set . . . . . 51
C.4 The rules of Random tree with PCA in WBC data set . . . . . 52
C.5 The rules of J48 decision tree without PCA in WDBC data set . . . . . 53
C.6 The rules of REP tree without PCA in WDBC data set . . . . . 53
C.7 The rules of Random tree without PCA in WDBC data set . . . . . 54
C.8 The rules of J48 decision tree with PCA in WDBC data set . . . . . 55
C.9 The rules of REP tree with PCA in WDBC data set . . . . . 55
C.10 The rules of Random tree with PCA in WDBC data set . . . . . 56
C.11 The rules of J48 Decision tree without PCA in WPBC data set . . . . . 57
C.12 The rules of REP tree without PCA in WPBC data set . . . . . 57
C.13 The rules of Random tree without PCA in WPBC data set . . . . . 58
C.14 The rules of J48 decision tree with PCA in WPBC data set . . . . . 59
C.15 The rules of REP tree with PCA in WPBC data set . . . . . 59
C.16 The rules of Random tree with PCA in WPBC data set . . . . . 60

# LIST OF TABLES


| Table | | Page |
|---|---|---|
| 2.1 | The comparison performance of PCA-ANN model | 4 |
| 2.2 | The results of classification techniques | 5 |
| 2.3 | The Eigenvalue of PCA and its error in data reduction | 6 |
| 2.4 | The comparison of precision and recall for traditional and hybrid logistic regression | 7 |
| 2.5 | The performance of decision tree algorithms for Wisconsin data | 8 |
| 3.1 | Bivariate data samples | 11 |
| 3.2 | An adjusted bivariate data sample | 12 |
| 3.3 | The bivariate data results of Eigenvalues and Eigenvectors | 14 |
| 3.4 | The bivariate data results of two Eigenvectors | 17 |
| 3.5 | The bivariate data results of one Eigenvector | 17 |
| 3.6 | An example of data set for playing golf | 19 |
| 3.7 | Information Gain values of weather data | 21 |
| 3.8 | The values of Split Information | 21 |
| 3.9 | The values of Gain ratio | 21 |
| 3.10 | The validation set of weather forecast | 23 |
| 3.11 | The Confusion Matrix | 27 |
| 4.1 | The details of Wisconsin data set | 29 |
| 4.2 | The comparison of accuracy between WBC data set with and without PCA | 30 |
| 4.3 | The classification rules of J48 and REP tree with PCA for WBC data set | 32 |
| 4.4 | Comparison of accuracy results between WDBC data set of original and original+ PCA | 33 |
| 4.5 | Comparison of accuracy results between WPBC data set of original and original+ PCA | 37 |
| A.1 | Data set of Wisconsin (Original) or WBC | 46 |
| A.2 | Data set of Wisconsin (Diagnostic) or WDBC | 46 |
| A.3 | Data set of Wisconsin (prognostic) or WPBC | 46 |
| B.1 | The eigenvalues of WBC data set | 48 |
| B.2 | The eigenvalues of WDBC data set | 48 |
| B.3 | The eigenvalues of WPBC data set | 48 |

# CHAPTER 1

# INTRODUCTION

## 1.1 Background and Problem Statement

A rule-based classifier plays an important role in modern breast cancer diagnosis. The good classifier equips with high accurate classification rules which are obtained from historical diagnosis. Since each diagnosis consists of a huge amount of data features which can lead to high dimensional data. Therefore, it is challenging to build minimal high accurate classification rules from such data. Basically, feature reduction techniques could help to reduce the number of classification rules. But the trade-off is classification performance. However, if we could find a technique of feature reduction giving high classification accuracy, it would assist to obtain minimal high accurate classification rules.

Fortunately, the Principal Component Analysis (PCA) is a data reduction technique giving new features (less than original features) that strongly differ across the classes. Hence, rules obtained from these new features always give classification performance better than the rules of original features. Therefore, this research proposes to use the PCA as a data reduction technique to achieve the goal of obtaining high accurate minimal classification rules. Among decision tree classifiers, these three classifiers namely J48, REP Tree, and Random Tree are known of their ability of providing rules ready to use in a rule-based system. Therefore, the research aims at finding the best classifier, in terms of number of rules and classification accuracy, among them on PCA reduced data of Wisconsin Breast Cancer Data set (WBCD).

## 1.2 Objectives

This study aims to build the minimal classification rules from the breast cancer diagnosis data with J48, REP Tree, and Random Tree. It employs Principal Component Analysis (PCA) to increase classification accuracy and reduce the number of rules.

## 1.3 Scope

This study is outlined as shown in Figure 1.1. It uses three collections of Wisconsin data sets including Wisconsin Breast cancer (WBC), Wisconsin Diagnostic Breast cancer (WDBC), and Wisconsin Prognostic Breast Cancer (WPBC). It prepares the data with and without applying PCA. The preprocessed data is then served as an input to the three classifiers: J48, REP Tree, and Random Tree. Ten-fold cross validation is used to validate the classification results.

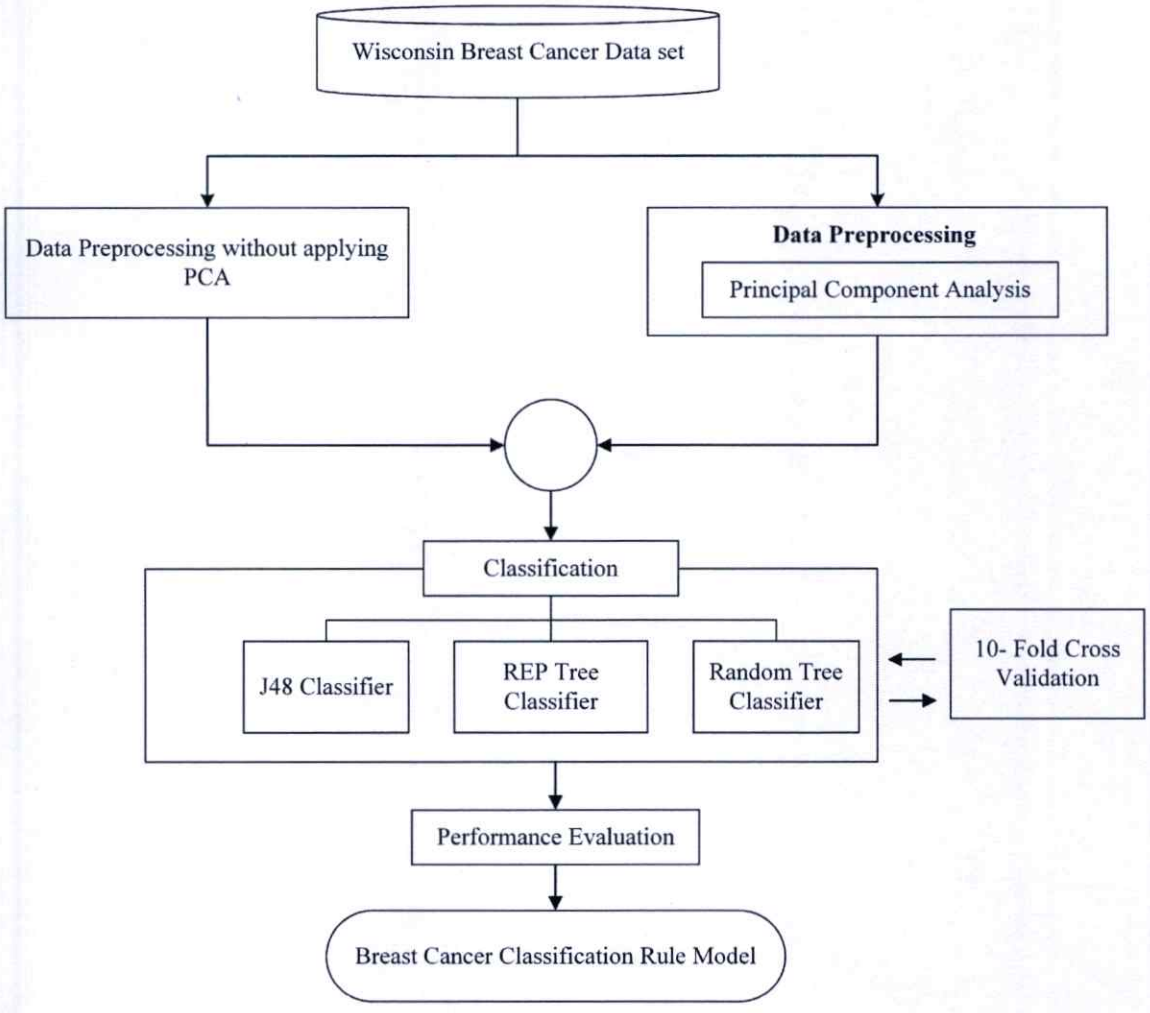



**Figure 1.1:** The study outline

## 1.4 Thesis Organization

The rest of the thesis is organized as follows. Chapter 2 provides a brief literature review and background knowledge on feature reduction and classification techniques. Chapter 3 gives details on the research methodology. Chapter 4 presents experimental results and discussion. Chapter 5 concludes the study and gives recommendation.

# CHAPTER 2

# LITERATURE REVIEW

This chapter provides the existing literature review of the previous research on breast cancer and presents a background study of feature reduction. It also introduces machine learning techniques and the main case study outlines that support to this study.

## 2.1 Feature Reduction

Due to the size of data features are increased so the computational consumption is required in the processing procedure. In order to minimize the features and computational time, feature reduction technique is applied. It required to extract the effective features form the whole data feature to a new smaller set for the classification process. Several methods have been implemented to extract the most effective features from the enormous data. Thus, the feature extraction technique has become a powerful approach. There are two categories for feature extraction algorithms: Supervised learning, i.e., LDA and Unsupervised learning, i.e., PCA and ICA. This study was used PCA feature reduction for reducing the number of features. In this section reviews the previous work about the classifiers which can be used with feature extraction.

Smita Jhajharia et al [2] implemented a hybrid prediction model which combined principal component analysis (PCA) technique with artificial intelligence based on machine learning techniques. The aim of the study was identified the breast cancer tumor and compared to other classifiers. Figure 2.1 is illustrated the proposed structure. The propose method used PCA-ANN hybrid model for classifying the instances as benign or malignant. The architecture of the ANN used feed forward neural network having an input layer, hidden layer, and output layer. The study has been conducted on WBC data set which divided into 2 groups by applied PCA. The

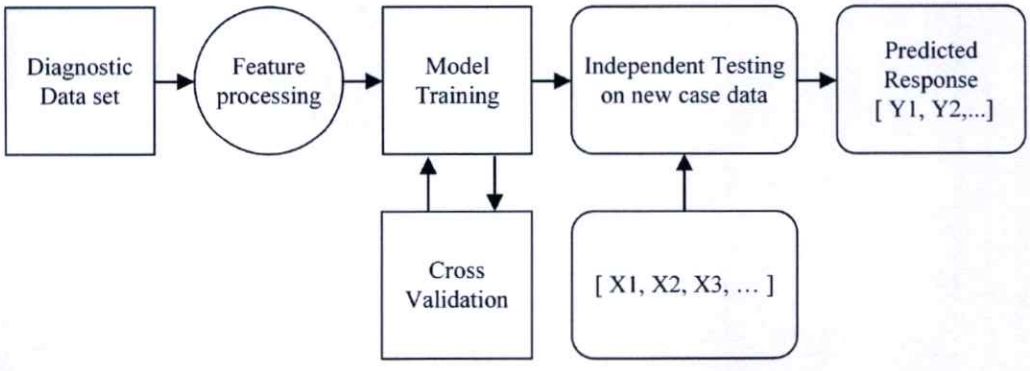



**Figure 2.1:** Flowchart of the Breast Cancer prognosis model

experimental results found that PCA-ANN model was given better classification result than other standard techniques as shown in Table 2.1. Nevertheless, these network problems failed to give the classification rules and take more time in computational complexity.

**Table 2.1:** The comparison performance of PCA-ANN model

| Prediction Model | Accuracy (%) |
|---|---|
| PCA-ANN | 98.39 |
| Support Vector Machine | 96.90 |
| Naive Bayes | 95.90 |
| IBK | 95.10 |
| Decision Tree | 94.50 |
| OneR | 92.70 |

N. Sharma and K. Saroha [3] proposed a novel method to reduce the dimensional data by using PCA and feature ranking. In order to reduce dimension, they applied PCA as a preprocessing step, the data dimension was reduced such that the computational efficiency of learning model and classification accuracies were improved. However, PCA did not provide a subset of real attributes and the correlation of attributes with the classes. Hence, feature evaluation and ranking algorithms are applied as shown in Figure 2.2. In order to classify the data, K-nearest neighbor

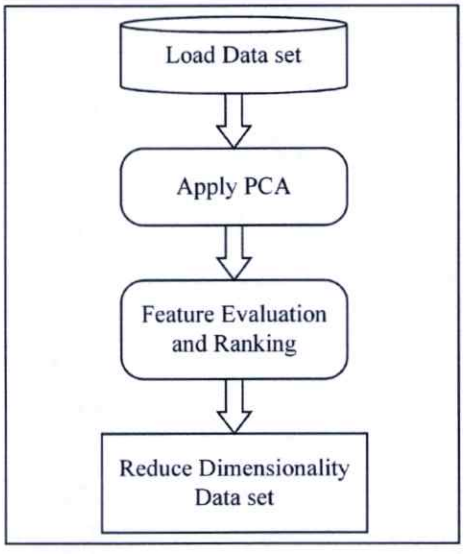



**Figure 2.2:** The Basic flow of proposed method

algorithm was used. This experiment was conducted on WDBC data set. The results found that the proposed method was achieved a better result than other methods which given 93.14% classification accuracy with five features as summarized in Table 2.2. Even though their work has better results from other works; however, in term of accuracy is required to be improved.

**Table 2.2:** The results of classification techniques

| Data set | Number of features | Accuracy (%) |
|---|---|---|
| Original | 30 | 92.97 |
| Using PCA | 10 | 92.26 |
| Using Feature Evaluation and Ranking | 8 | 92.97 |
| Using Proposed method | 5 | 93.14 |

T. M. Mohamed [4] presented an algorithm for decreasing the number of the features required for the classification task. To handle this problem, feature extraction and feature selection were used in preprocessing step. This research was carried out of using sequential forward selection (SFS) and PCA algorithms for comparing the performance of reduction techniques as shown in Figure 2.3. In this experiment, linear classification function (LCF) and quadratic classification

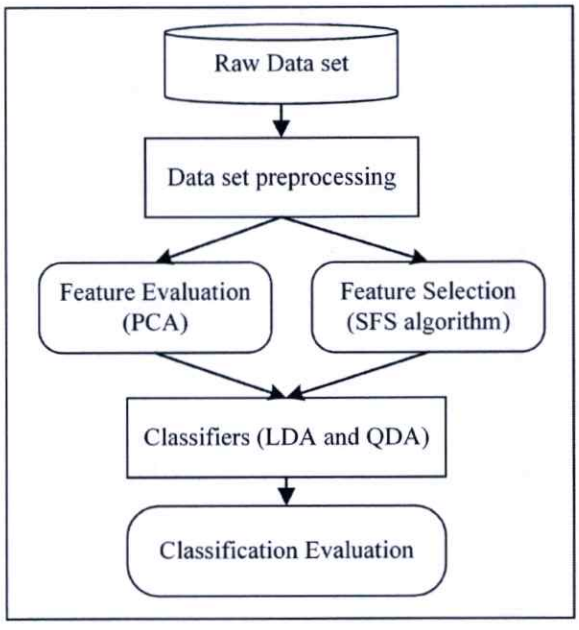



**Figure 2.3:** The flowchart of PCA and SFS approach

functions (QCF) were applied. Furthermore, they were also used $K$-means clustering technique to cluster the test cases and identified the possible abnormal (outliers) patterns in the data set. WBC data set was used as a training and testing the classification model. The results indicated that the data set were not significant outliers patterns. The highest accuracy achieved by 97.16 % with four principal components when using LCF technique. While the QCF was enhanced when combined with SFS algorithm which obtained 96.21% accuracy as shown Figure 2.4. However, SFS algorithm technique may lead to losing some information during the selecting feature process, i.e., removing some effective features. Additionally, the number of principal components could also be reduced less than 4 PCs which is able to save the computational time.

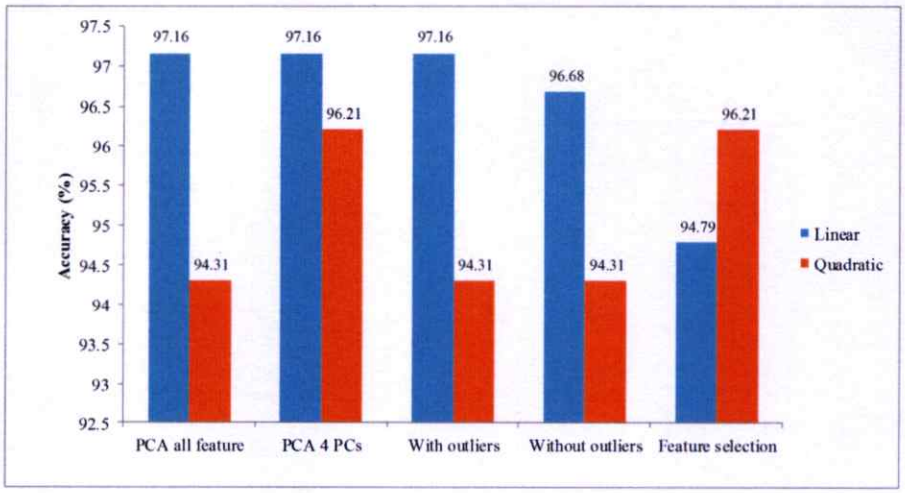



**Figure 2.4:** The performance of two Linear and Quadratic classifiers

C. L.l Sabharwal and B. Anjum [5] proposed a hybrid predictive model that adaptively used PCA to improve the linear and logistic regression algorithms. The purposes of the study were focused on both data reduction and prediction by improving Logistic Regression. This experiment collected the data set of California Hospital Rating from HealthData.gov to demonstrate in a learning process of the model. The data set was created in two versions, i.e., raw data and mean centered with unit standard deviation. In order to perform PCA, Eigenvalues and Eigenvector were required to calculate as summarized in Table 2.3. The principal component of normalized data was more realistic and distributable. Therefore, 40% of maximum variance data in PCA analysis with were used traditional logistic regression. The results of traditional and hybrid logistic

**Table 2.3:** The Eigenvalue of PCA and its error in data reduction

| | Raw Data | Normalized Data |
|---|---|---|
| Eigenvalues | -71.5188 | -108.94 |
| Eigenvectors | 0.0601 | 0.0856 |
| | -0.1819 | 0.0264 |
| | 0.1183 | 0.0166 |
| | 0.7228 | 0.0461 |
| | 0.6534 | 0.9948 |
| Errors | 0.0542 | 0.0347 |

regression indicated that the values of precision between Raw and PCA data was almost the same values whereas recall was given a better value. However, when using hybrid logistic regression algorithm, they found that the hybrid algorithm was outperformed the traditional algorithms as shown in Table 2.4.

**Table 2.4:** The comparison of precision and recall for traditional and hybrid logistic regression

| | Traditional logistic regression | | Hybrid logistic regression | |
|---|---|---|---|---|
| | Raw Data | PCA 40% Data | Raw Data | PCA 40% Data |
| Precision | 0.952 | 0.951 | 0.965 | 0.957 |
| Recall | 0.807 | 0.744 | 0.897 | 0.759 |

## 2.2 Classification

Classification is the process of learning a set of features (inputs) that assigns items in a collection to the target categories or class labels (output). Each record contains a set of attributes. It is represented in the form of (x, y) where x is the feature set and y is the class label. The aim of classification is to correctly predict the target class (unseen data) for each case in data, i.e., a classification model could be used to diagnose breast cancer as a benign or malignant tumor based on Decision Tree, Rule-based Methods, Support Vector Machines classifier, etc.

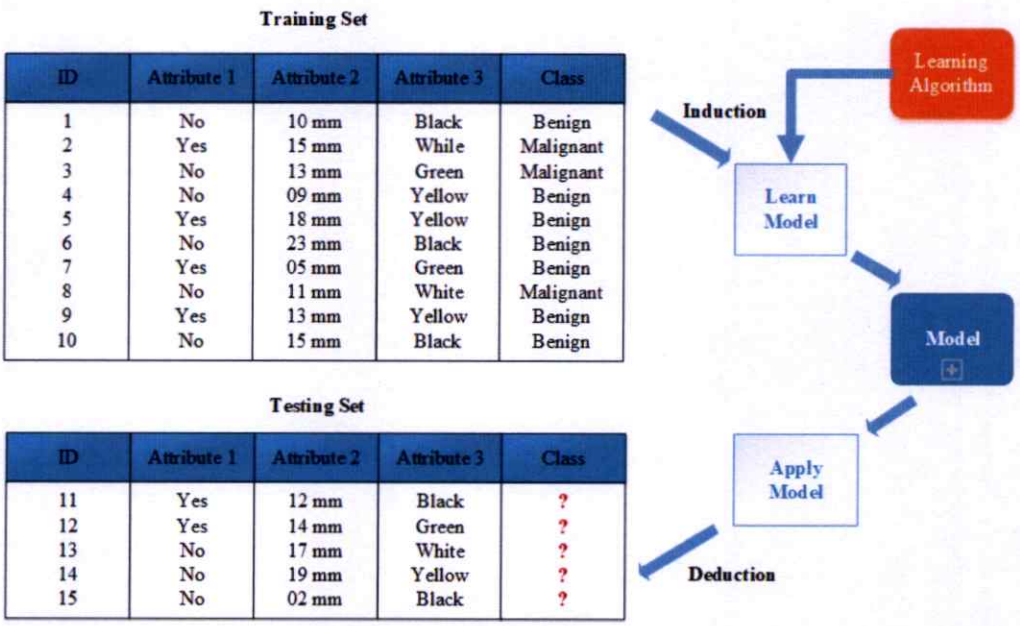



**Figure 2.5:** An example of classification task

The classification problems can be solved as illustrated in Figure 2.5. The training data consists of a set of training samples, each training sample represents a pair of an input features and class label. Similarly, each testing sample is a pair of input features with no class value. The classification model is created based on training data and produced an inferred function, also known as a classifier. During the testing phase, the inferred function is used to classify the correct class for any valid pair of input features. Thus, the inferred function predicts the class of testing samples. This requires the machine learning to generalize using the training data to classify unseen samples in an appropriate manner. During the past few years, various researches

have been conducted the efficiency classification model of breast cancer diagnosis and prognosis in tumor level.

P. Hamsagayathri and P. Sampath [6] proposed to find the best performance of four different decision tree algorithms including J48, REP Tree, Random Tree, and Priority based decision tree classifiers. Wisconsin data sets were used in the experiment which are WBC, WPBC, and WDBC. The evaluation of classification model was used ten-fold cross validation by using WEKA software. The experimental results indicated that the Priority based decision tree classifier proved to be the most accurate in WBC, WDBC, and WPBC data sets with 94.70 %, 96.48%, and 83.83% accuracies respectively as illustrated in Table 2.5. This work was given a minimum RMSE (Root Mean Squared Error) among of classifiers. By comparing classification results, they confirmed that the priority-based decision tree was better than other classification algorithms. Nevertheless, J48, REP Tree and Random tree classifiers could be performed better in term of accuracy.

**Table 2.5:** The performance of decision tree algorithms for Wisconsin data

| Classifiers | WBC | | WDBC | | WPBC | |
|---|---|---|---|---|---|---|
| | Accuracy (%) | RMSE | Accuracy(%) | RMSE | Accuracy(%) | RMSE |
| J48 | 93.56 | 0.239 | 95.43 | 0.208 | 83.33 | 0.4766 |
| Random Tree | 94.13 | 0.2422 | 94.20 | 0.2408 | 68.68 | 0.5596 |
| REP Tree | 94.13 | 0.2311 | 94.37 | 0.2175 | 76.7 | 0.4294 |
| Priority based | 94.70 | 0.2198 | 96.48 | 0.1823 | 83.83 | 0.3768 |

Ronak Sumbaly et al [8] presented an automatic diagnosis based on pattern recognition system for detecting breast cancer in its early stages. In this research used J48 decision tree as a classifier for creating the diagnosis model based on WBC data set. According to the performance of the J48 classifier, the experiment result shown that it achieved the classification accuracy of 94.56 %. Furthermore, this work demonstrated the efficiency of J48 classifier and the ability to generate the simple rules with flexibility for breast cancer diagnosis problems. Although their accuracy and performance were acceptable, reducing the data features could improve the classification model.

Hind Elouedi et al [7] proposed a hybrid diagnosis approach of breast cancer based on decision trees and clustering as given in Figure 2.6, the proposed diagram. The aim of this classification model was to improve the classification of malignant instances by using K-means algorithm for clustering and C4.5 for classification. The experimental results found that the splitting up of malignant instances into two clusters and feeding them into the decision tree algorithm were given better results up to 95.14%. This classification application could be useful for breast cancer diagnosis. However, in term of accuracy the classification model needs to be improved.

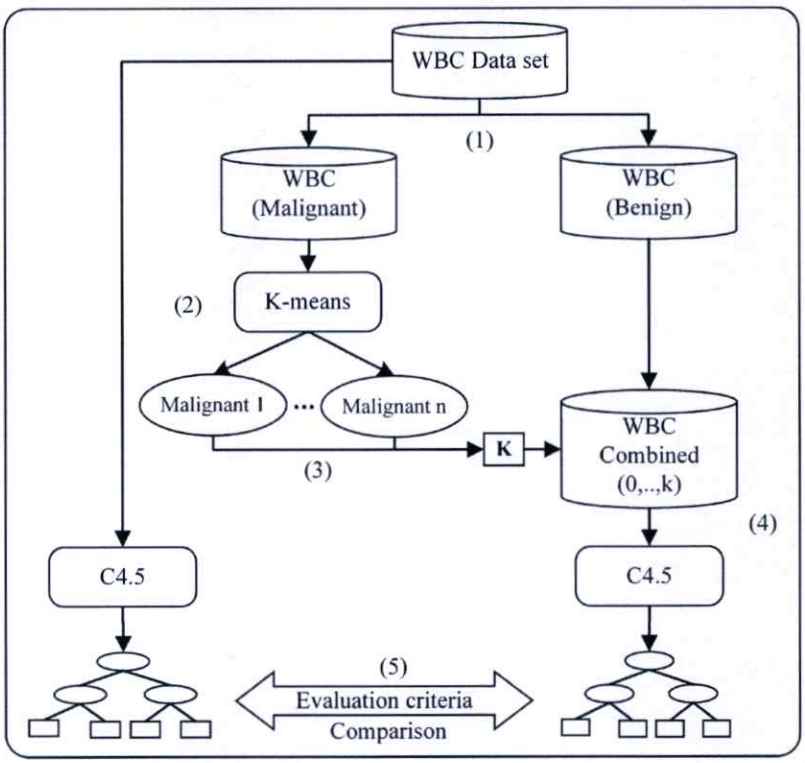



**Figure 2.6:** The process of experiment protocol

# CHAPTER 3

# RESEARCH METHODOLOGY

This chapter presents the details of research methodology and data source that applied in this study. The main idea of the thesis is to build the minimal classification rules on breast cancer diagnosis. In order to develop this system, the breast cancer data need to be analyzed. Additionally, the classification tasks are applied and evaluated the performance of the model. This study is also conducted based on a principal component analysis technique.

## 3.1 Data Collection

High-quality data is very important to acquire from the public resources. One of that resource is an online database, it is collected from clinical environment and have approval processes from the committee. It also freely for research purposes. The advantage of using online databases is the comparison efficacy between existing solutions and our solutions by using the same data set.

UCI is a collection of databases, domain theories, and data generators that are used by machine learning for training and testing algorithms. In 1987, the repository database was created by Avid Aha and fellow graduate students at UC Irvine [9]. Since that time, there were many students, educators, and researchers are widely used as a primary source of machine learning databases. In this study has used different breast cancer data sets. The samples of clinical were collected by University of Wisconsin Hospitals. There are three mains collection that available including WDBC (consists of 569 instances with 32 attributes), WBC (consists of 699 instances with 11 attributes), and WPBC (consists of 198 instances with 35 attributes) are considered which are publicly available on UCI Machine Learning Repository [10]. The description of data set is summarized in Table A.1, A.2, and A.3 of Appendix A.

## 3.2 Data Preprocessing

Data collection phase was contained incomplete data, outlier data, and inconsistent data. Incomplete data can occur for numerous reasons. When learning from these data, the models may confuse in the process. It can be leading to data over fitting and the accuracy of the model can be poor, i.e., some values of the attributes are missing, and some attributes value are not effective for learning the model [11]. Outliers are often discarded as noise that do not comply with general behavior or model of the data. It may be detected using statistical or distance measurement that

assume a distribution or probability model for the data [11]. Inconsistencies data occurs when there are records containing discrepancies in codes or names on the data set.

The elements of quality data are defined as completeness, accuracy, and consistent data. Data preprocessing technique is an important phase of learning model to produce the high quality of data elements. It can also improve the accuracy and efficiency of the classification model as possible. Therefore, this study was applied data preprocessing task to ensure that the data set are ready to use for machine learning process. The study had proposed PCA feature reduction technique as a preprocessing phase. To reach the high quality of data, PCA is created a new set of data from the original that contains high variances data. Then, it finds a linear combination of variables by selecting the subset of principal component which maintained the most important data features. The best significant principal component had been collected. At the end of current task, the data had been ready to utilize in the next step for the experiment.

### 3.3 Principal Component Analysis

Principal component analysis (PCA) is one of the dimensional reduction techniques based on a mathematical method that defines an orthogonal linear transformation. The purpose of PCA is to decrease the multidimensional data that contains a large number of correlated attributes by transforming original data space to a new space in which attributes are uncorrelated. These uncorrelated attributes are called as principal components (PCs). The number of PCs are less than or equal to number of original attributes [12]. Thus, the dimension of the multivariate data is reduced due to most of the largest possible variances will be retained in the PCs. In addition, PCA does not need class labeled to train because it is an unsupervised learning method [13]. The feature reduction algorithm of PCA can be explained as follows.

#### 3.3.1 Data set

In $X$ data matrix of order $n \times p$, that $n > p$ and $n$ is the number of elements in the set $X$, $p$ is the number of variables. Supposedly, we made-up bivariate data for this example as shown in Table 3.1 and Figure 3.1.

**Table 3.1:** Bivariate data samples

| $x$ | 2.5 | 0.5 | 2.2 | 1.9 | 3.1 | 2.3 | 2 | 1 | 1.5 | 1.1 |
|---|---|---|---|---|---|---|---|---|---|---|
| $y$ | 2.4 | 0.7 | 2.9 | 2.2 | 3 | 2.7 | 1.6 | 1.1 | 1.6 | 0.9 |

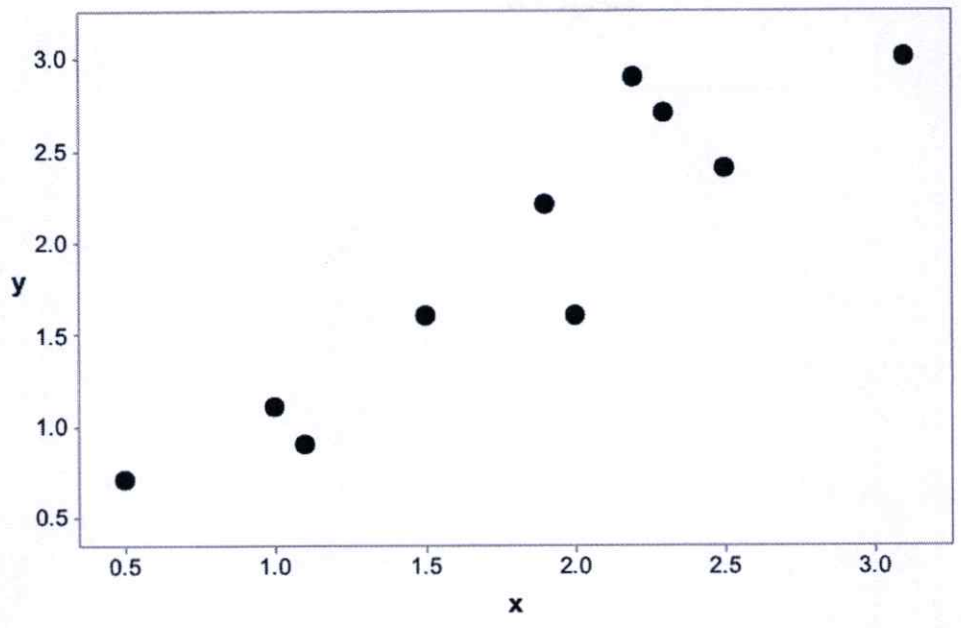



**Figure 3.1:** The graphical representation of bivariate data

### 3.3.2 Subtract the means

For scaling features, we calculate the means ($\bar{X}$ ) of the sample size in the equation (3.1). Then subtract the mean of each column features from the corresponding data component to re-center the data. Assumed that, $S$ is a scaled version of matrix $X$. These adjusted or shifting data consisted of mean zeros, as given in Table 3.2 and Figure 3.2 .

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{3.1}$$

**Table 3.2:** An adjusted bivariate data sample

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $x$ | 0.69 | -1.31 | 0.39 | 0.09 | 1.29 | 0.49 | 0.19 | -0.81 | -0.31 | -0.71 |
| $y$ | 0.49 | -1.21 | 0.99 | 0.29 | 1.09 | 0.79 | -0.31 | -0.81 | -0.31 | -1.01 |

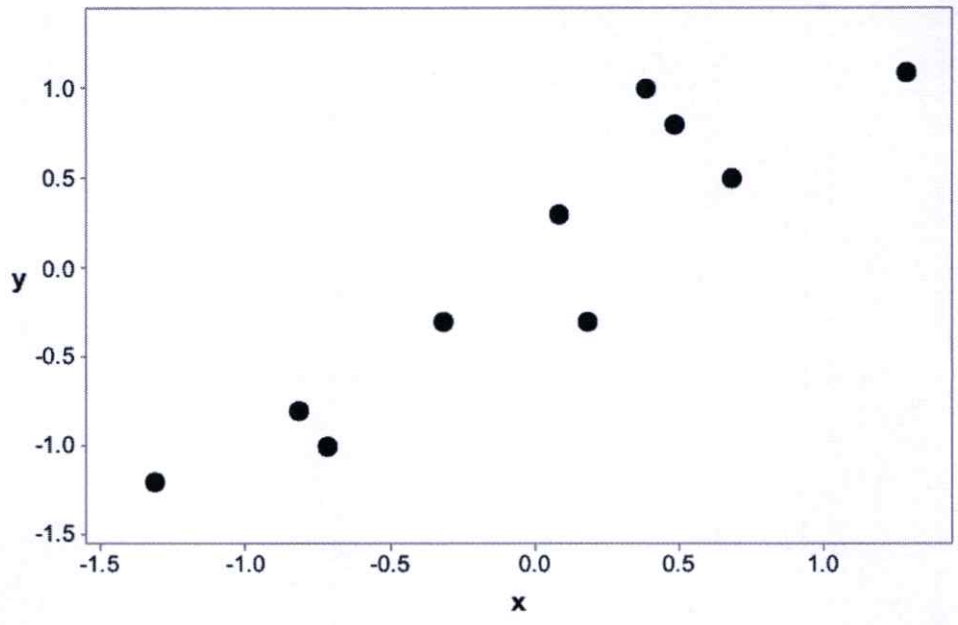



**Figure 3.2:** The graphical representation of adjusted bivariate data

### 3.3.3 Calculate the covariance matrix

In this section, there were two approaches to calculate. Firstly, compute the sample variance-covariance matrix $A$ by equation (3.2). Since we re-centered the data at mean zero. So, vector $\bar{X} = (0, 0)$. Secondly, if the data was standardized, then matrix $A$ was correlation matrix by equation (3.3). Where $n$ was number of elements in the set $X$ and transposed of matrix $S$ was $\bar{S}$. Since our data was two dimension, the matrix was $2 \times 2$ matrix.

$$A = \frac{1}{n-1}(X - 1\bar{X}')'(X - 1\bar{X}') = \frac{1}{n-1}X'X \tag{3.2}$$

$$A = \frac{1}{n-1}\bar{S}S \tag{3.3}$$

Note that we used covariance matrix to perform PCA. This was possible because the data set seemed to have the same scale. It meant that we used covariance matrix for PCA, if the variables were the same scale or unit for both variables. On the other hand, the correlation matrix was used for PCA, if the variables were different scale or unit. After calculation we obtained the covariance matrix results as following:

$$A = \frac{1}{10-1}\begin{pmatrix} 0.69 & -1.31 & ... & -0.71 \\ 0.49 & -1.21 & ... & -1.01 \end{pmatrix}_{2\times 10} \times \begin{pmatrix} 0.69 & 0.49 \\ -1.31 & -1.21 \\ ... & ... \\ -0.71 & -1.01 \end{pmatrix}_{10\times 2}$$

$$A = \frac{1}{10-1}\begin{pmatrix} 5.549 & 5.539 \\ 5.539 & 6.449 \end{pmatrix}_{2\times 2}$$

$$A = \begin{pmatrix} 0.616556 & 0.615444 \\ 0.615444 & 0.716556 \end{pmatrix}_{2\times 2}$$

### 3.3.4 Compute the Eigenvalues and Eigenvector

This section we analyzed the eigenvalues and eigenvector, then computed the percentage of variability captured by the principal components. Based on this inequality $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq ... \geq \lambda_p \geq 0$, $\lambda_p$ is the eigenvalues of matrix $A$ (in decreasing order) with corresponding eigenvector $\vec{e_1}, ..., \vec{e_p}$. The eigenvectors are represented as the direction of the line (horizontal, vertical). In addition, it is also called perpendicular or orthogonal, i.e., no matter how many dimensions you have, and eigenvalue is a number of variances in the datas direction of eigenvector.

Notice that: the trace of matrix $A$ is the sum of the diagonal entries of $A$, which is the sum of the variances of all $p$ variables. Let $T_v$ is the total sample variance of data. On the other hand, the trace of a matrix is equal the sum of its eigenvalues. So $T_v = \lambda_1 + \lambda_2 + ... + \lambda_p$. For the fraction of the total variance is calculated by $\frac{\lambda_p}{T_v}$ as demonstrated in Table 3.3.

**Table 3.3:** The bivariate data results of Eigenvalues and Eigenvectors

| Variables | Eigenvector 1 | Eigenvector 2 |
|---|---|---|
| $x$ | 0.677837 | 0.735179 |
| $y$ | 0.735179 | -0.677873 |
| Eigenvalues ($\lambda$) | 1.28403 | 0.04908 |
| Percentage of Total variance (%) | 96.3 | 3.7 |

### 3.3.5 Selecting the principal components

We choose the principal components based on the eigenvector matrix $T$ which was a transformation matrix. The sequence of components is arranged by eigenvalues from the highest to the lowest. Hence, we decided to ignore the components of lesser significance. However, it did not lose too much of variance information.

In this study, determining the dimensionality reduction of the data was applied by two rules of thumb which are Kaisers Rule (KG-rule) and Scree test. They were commonly used criteria for selecting the number of components and effective methods.

For the first method, KG-rule was applied. It has retained only the PCs whose eigenvalues are greater than 1 (See Table 3.3) [14] . However, for large variable spaces, KG-rule has retained too many PCs. In order to perform Scree test, it is a graphical method of defining the number of PCs to retain. Scree test finds a break between the components with relatively large and small eigenvalues on the Scree plot [14]. The components that appear before the break are retained and after the break are not retained (see Figure 3.3). Therefore, as to reduce the dimensions of data, we kept only the first principal component (Eigenvector 1 or PC1).

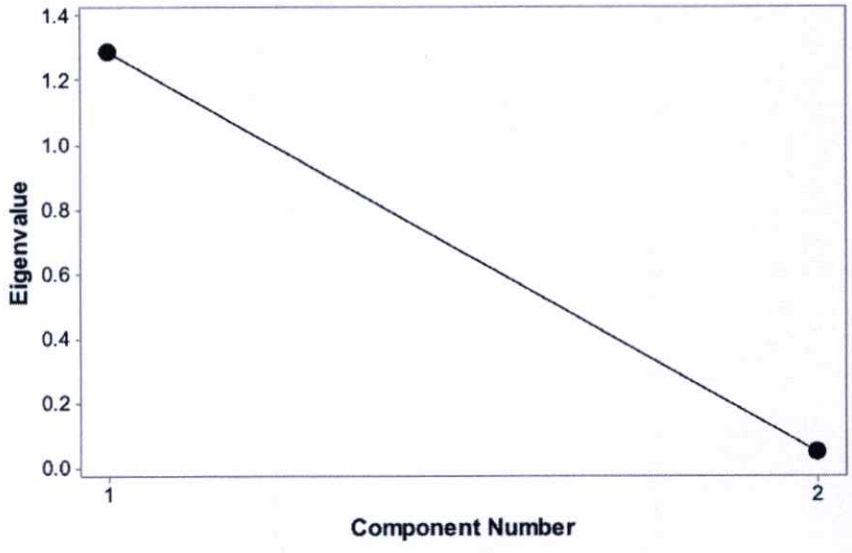



**Figure 3.3:** The graphical of Scree test

Supposedly, we decided to retain both principal components, then

$$T = \begin{pmatrix} 0.677873 & 0.735179 \\ 0.735179 & -0.677873 \end{pmatrix}$$

If we discarded the less significant principal component, then

$$T = \begin{pmatrix} 0.677873 \\ 0.735179 \end{pmatrix}$$

### 3.3.6 Deriving the new data set

This is the last step in PCA. Derive the new data set by taking $F = XT$, where $F$ is transformed data matrix of principal components (final data set with data items in columns), $X$ is original data matrix, and $T$ is (transformation matrix) matrix with the eigenvectors in the columns transposed. Thus, the eigenvectors are now in the rows with the most significant eigenvector at the top. Basically, the data has been transformed. Then, it is expressed in terms of patterns between them, where the patterns are the lines that most closely describe the relationships between the data.

In order to transform data by using both eigenvectors, we obtained the new data and plot on the graph as shown in Figure 3.4 and Table 3.4. This plot was basically the original data rotated. So that the eigenvectors were the axes. This is understandable since we have no lost information in this decomposition.

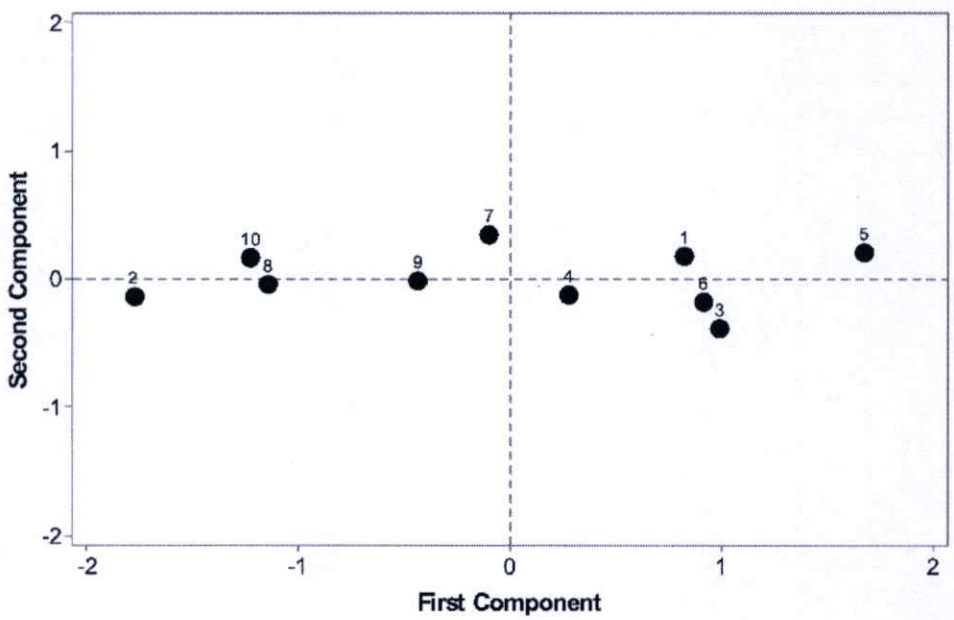



**Figure 3.4:** The graphical representation of two Eigenvectors

**Table 3.4:** The bivariate data results of two Eigenvectors

| $x$ | $y$ |
|---|---|
| 0.82797 | 0.17512 |
| -1.77758 | -0.14286 |
| 0.99220 | -0.38437 |
| 0.27421 | -0.13042 |
| 1.67580 | 0.20950 |
| 0.91295 | -0.17528 |
| -0.09911 | 0.34982 |
| -1.14457 | -0.04642 |
| -0.43805 | -0.01776 |
| -1.22382 | 0.16268 |

The other transformation used only one eigenvector with the largest eigenvalue to represent the data as given in Table 3.5. As expected, it only had a single dimension. When we compare this data with the one resulting from using both eigenvectors, we noticed that this data was exactly the first column of the other. So, when this data was plot, it was 1 dimensional data and was a line in exactly the positions of the points in the plot graph as shown in Figure 3.4. We have effectively thrown away the whole other axis, which was the other eigenvector.

**Table 3.5:** The bivariate data results of one Eigenvector

| $x$ |
|---|
| 0.82797 |
| -1.77758 |
| 0.99220 |
| 0.27421 |
| 1.67580 |
| 0.91295 |
| -0.09911 |
| -1.14457 |
| -0.43805 |
| -1.22382 |

## 3.4 Applying Machine Learning Techniques

The study has focused on feature reduction techniques as a method to achieve the high-quality attributes and reduce the number of rules from large data features. In our study, a comparison between the classifiers with and without applying PCA technique based on data set from previous step and three different decision tree classifications such J48, REP Tree, and Random Tree classifiers are applied.

### 3.5 Decision Tree

Decision Tree is a predictive machine learning model utilized to decide the outcome of new samples based on the available historical data. It is also called a supervised machine learning algorithm. Many problems have been solved by decision tree classification approach based on dividing and conquering strategy. The decision tree is represented by a rule based (if-then rules) which described by nominal and numeric properties. Generally, each path from the root node to the leaf node is represented as a classification rule. The tree consists of nodes, branches, and leaves as shown in Figure 3.5. The construction of decision tree is created from a root node at the top of the tree to any leaf node that defined the features. Each node in the tree is connected with one or more nodes using branches, the last node in the tree without branches is called leaf node. The leaf node is represented a classification category as a target class.

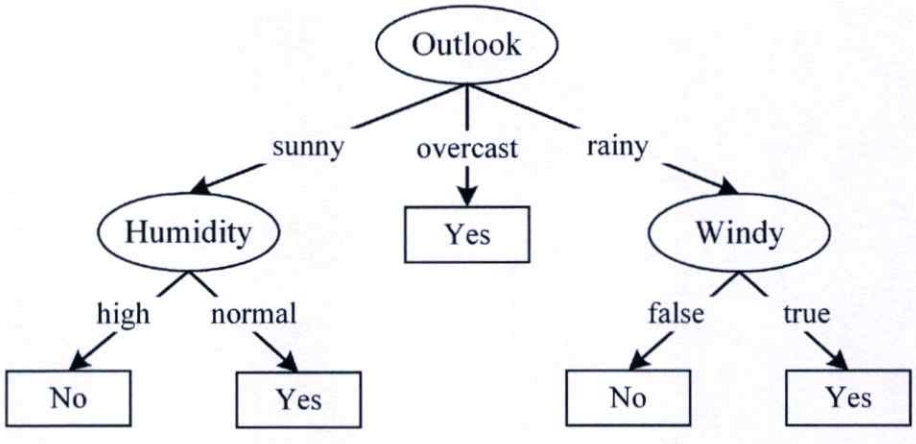



**Figure 3.5:** An example of decision tree model

In order to classify a new instance, a decision tree classification model was constructed as given in Figure 3.5. It is started from the root to a leaf node by applying the test criteria to the records and followed a matching path base on the outcome testing. According to the proposed method, three decision tree classification models including J48, REP Tree, and Random Tree classifiers were applied.

#### 3.5.1 J48 Decision Tree

J48 is a well-known algorithm used to generate a decision tree for decision making. It is an implementation of C4.5 algorithm by using Java application in the Weka data mining tool which presented by Quinlan [15]. The additional features of C4.5 algorithm overcome the limitation of ID3 algorithm due to the limitation of information gain approach such as the attributes indicating as distinct values and the biased towards tests with many outcomes, i.e., attribute $Day$ as illus-

**Table 3.6:** An example of data set for playing golf

| Day | Outlook | Humidity | Windy | Decision |
|---|---|---|---|---|
| 1 | Sunny | High | False | No |
| 2 | Sunny | High | True | No |
| 3 | Overcast | High | False | Yes |
| 4 | Rainy | High | False | Yes |
| 5 | Rainy | Normal | False | Yes |
| 6 | Rainy | Normal | True | No |
| 7 | Overcast | Normal | True | Yes |
| 8 | Sunny | High | False | No |
| 9 | Sunny | Normal | False | Yes |
| 10 | Rainy | Normal | False | Yes |
| 11 | Sunny | Normal | True | Yes |
| 12 | Overcast | High | True | Yes |
| 13 | Overcast | Normal | False | Yes |
| 14 | Rainy | High | True | No |

trated in Table 3.6. Since attribute $Day$, the number of every records are unique. Testing on its value will always yield low entropy values (entropy equal 0). This computation achieves high information gain and has a chance to be a root node. However, this node is not useful for learning the classification model. To overcome these problems, C4.5 was improved this bias by using split information of each attribute for calculating the Gain Ratio criterion. The highest value of information gain ratio is selected as a root node and then splitting process is continued until reaching to the leaf node. The decision tree model construction is based on Entropy, Information Gain, and Information Gain Ratio.

Shanon Entropy [16] or degree of impurity is defined as the quantitative measure of the randomness or disorder in the information being processed. To identify a perfectly classified set, the value of Entropy must be zero. In other words, a set with high entropy is hard to draw any conclusions from that information. The entropy values can range from 0 to 1. Supposed $S$ is a random variable with possible values $\{S_1, S_2, S_3, ..., S_n\}$, the Entropy is specified as follows:

$$Entropy(S) = \sum_{i=1}^{n} -p_i \times log_2(p_i) \tag{3.4}$$

where $n$ is referred to the number of class levels and $p_i$ is a probability of class in $S$.

Information Gain [16] is used to determine which attributes should be split on at each stage in the decision tree construction. Gain $(S, A)$ of an attribute $A$, compared to a set of examples $S$, Information Gain can be defined by using equation (3.5).

$$Gain(S, A) = Entropy(S) - \sum_{j \in Values(A)} \frac{|S_j|}{|S|} Entropy(S_j) \tag{3.5}$$

where values $A$ is a set of all possible for attribute $A$ and $S_j$ are subsets of $S$ which attribute $A$ has value $j$ ($X_A = j$). There are two sections of equation (3.5). For the first section is entropy of original collection $S$ and the second section is the expected value entropy which is calculated the sum of the entropies of each subset weighted by the fraction of examples. The equations of Split Information and Gain Ratio are defined as follows:

$$SplitInfo(S, A) = -\sum_{i=1}^{n} \frac{|S_j|}{|S|} \times \log_2 \frac{|S_j|}{|S|} \tag{3.6}$$

and

$$GainRatio(S, A) = \frac{Gain(S, A)}{SplitInfo(S, A)} \tag{3.7}$$

Supposedly, we applied a J48 algorithm to decide whether play or not play tennis based on the weather data as given in Table 3.6. The set of $S$ training data consists of 14 instances and includes two labels: yes and no. There are 9 decisions labeled yes and 5 for no labeled. We determined the information to classify such Entropy as following:

$$Entropy(S_1) = -[p(Yes) \times log_2 p(Yes) + p(No) \times log_2 p(No)]$$

$$Entropy(S_1) = -[(9/14) \times log_2(9/14) + (5/14) \times log_2(5/14)] = \mathbf{0.94}$$

Now, we found the most dominant factor for a decision and then Information Gain is calculated. Assume that $Outlook$ attribute is the first feature in the calculation processes. It is calculated as follows.

$$Entropy(S_2\{Outlook = sunny\}) = -[(2/5) \times log_2(2/5) + (3/5) \times log_2(3/5)] = 0.97$$

$$Entropy(S_2\{Outlook = overcast\}) = -[(4/4) \times log_2(4/4) + (0/4) \times log_2(0/4)] = 0$$

$$Entropy(S_2\{Outlook = rainy\}) = -[(3/5) \times log_2(3/5) + (2/5) \times log_2(2/5)] = 0.97$$

$$Gain(S_1, Outlook) = Entropy(S_1) - Entropy(S_2 = \{Outlook\})$$

$$= Entropy(S_1) - p(Outlook = sunny) \times Entropy(S_2 = \{Outlook = sunny\}$$

$$-p(Outlook = overcast) \times Entropy(S_2 = \{Outlook = overcast\}$$

$$-p(Outlook = rainy) \times Entropy(S_2 = \{Outlook = rainy\}$$

$$= 0.94 - [5/14 \times 0.97 + 4/14 \times 0 + 5/14 \times 0.97]$$

$$= 0.94 - 0.69 = \mathbf{0.247}$$

The remaining attributes that can be chosen as a root node of the tree for training data such as humidity and windy. Furthermore, to obtain their values, we applied the same calculation to find the most dominant factor which attributes should be selected as given in Table 3.7.

**Table 3.7:** Information Gain values of weather data

| Gain | Values |
|---|---|
| Gain ($S_1$, Outlook) | 0.247 |
| Gain ($S_1$, Humidity) | 0.152 |
| Gain ($S_1$, Windy) | 0.048 |

Splitting information can be calculated for each attribute, e.g., the split information of $Outlook$ attribute is computed as belows:

$$SplitInfo(S1, Outlook) = -5/14 \times log_2(5/14) - 4/14 \times log_2(4/14) - 5/14 \times log_2(5/14)$$

$$SplitInfo(S1, Outlook) = 1.58$$

**Table 3.8:** The values of Split Information

| Split Information | Values |
|---|---|
| Split($S_1$, Outlook) | 1.577 |
| Split($S_1$, Humidity) | 1.00 |
| Split($S_1$, Windy) | 0.985 |

Gain Ratio is used for attributes selection. To compute a Gain Ratio, Split Information must be calculated. This is an example of calculating Gain Ratio of $Outlook$ attribute.

$$GainRatio(S1, Outlook) = 0.25/1.58 = 0.156$$

**Table 3.9:** The values of Gain ratio

| Gain Ratio | Values |
|---|---|
| Gain Ratio ($S_1$, Outlook) | 0.156 |
| Gain Ratio ($S_1$, Humidity) | 0.152 |
| Gain Ratio ($S_1$, Windy) | 0.049 |

The Gain Ratio is shown in Table 3.9. Since Outlook attribute proves to be the best Gain Ratio. Therefore, this attribute is used as a root node. In Figure 3.6 is shown the partially learned decision tree from the first step of J48 classifier. The final decision tree was learned by J48 classifier from 14 training samples are given in Figure 3.5.

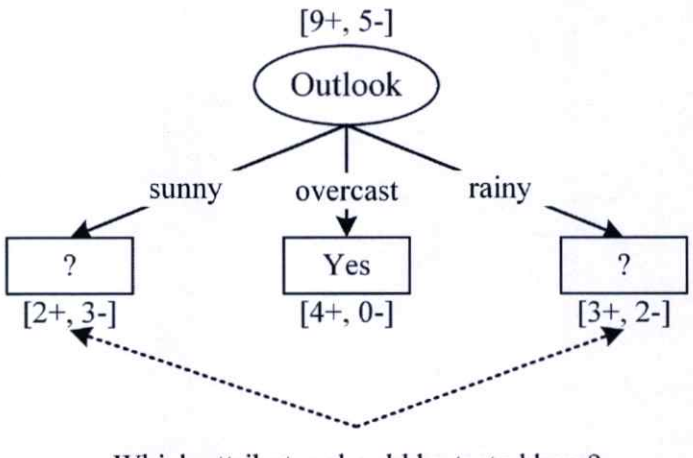



**Figure 3.6:** The partially learned decision tree from the first step of J48

The process of selecting a new attribute and dividing the training sample is now repeated for each non-terminal descendant node. It continues the process for each new leaf node until either of two criteria: every attribute has already been included along with this path through the tree and entropy is zero or splitting stops when data cannot be split any further.

### 3.5.2 Reduce Error Pruning Tree

REP Tree builds a decision tree based on information gain or variance of attributes and uses the pruning set in order to evaluate the goodness of the subtree [17]. It is used Reduced Error Pruning technique to minimize the error rate from the variance [18]. The idea is to hold out some of the available instances, i.e., validation set or pruning set that separated from the training data. It means that the instances in the validation set are not used for constructing the decision tree. To prune the tree, the decision tree must be constructed. The algorithm is explored by traversing from bottom-up strategy. REP Tree is the simplest idea by using a pruning set to evaluate the performance accuracy of the sub-tree and individual nodes. In addition, it is also provided a less bias estimate of its error rate on unseen data and used less time for learning a model. Suppose $B$ is a subtree root at node $A$ as shown in Figure 3.7 and $P_A$ is a gain from pruning at node $A$. The gain from a node pruning is specified as following:

$$P_A = B_n - A \tag{3.8}$$

where $A$ is a number of misclassification at node $A$ and $B_n$ is the total error of misclassification at leaves $B$.

To prune the node of decision tree, it performs replacing a subtree by a leaf. The process is repeated further at a node with the largest gain until the gain is negative value then a node is maintained and stopped pruning at that node.

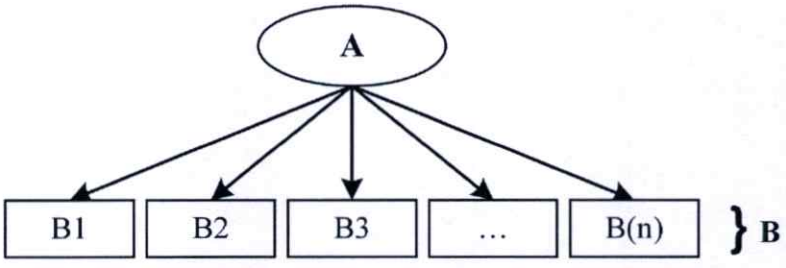



**Figure 3.7:** The representation of subtree in REP Tree

For example, Figure 3.5 shows an un-pruned decision tree. In order to prune the tree, validation set must be applied. Assume that the validation set 11 examples were made up as given in Table 3.10 . After applying validation set the classification results are given in Figure 3.8a. In each tree, the number of instances in the pruning data misclassified by the individual nodes

**Table 3.10:** The validation set of weather forecast

| Day | Outlook | Humidity | Windy | Decision |
|---|---|---|---|---|
| 17 | Sunny | High | True | Yes |
| 18 | Sunny | High | False | Yes |
| 19 | Sunny | Normal | False | Yes |
| 20 | Rainy | High | False | Yes |
| 21 | Overcast | Normal | True | Yes |
| 22 | Sunny | High | True | No |
| 23 | Sunny | Normal | False | No |
| 24 | Sunny | Normal | True | No |
| 25 | Sunny | Normal | False | No |
| 26 | Sunny | Normal | True | No |
| 27 | Rainy | High | True | No |

are given in parentheses. Suppose that the tree is traversed from left to right, Humidity node is considered for pruning the subtree (Figure 3.8 (b)). Because the subtree's error on the pruning data (6 errors) exceeds the error of node Humidity itself (3 errors) or $P_{humidity}$ is provided positive value, node Humidity is converted to a leaf (Figure 3.8 (c)). Next, the subtree extending from

node Windy is considered for pruning. Because the subtree attached to node Windy has 0 errors than node Windy itself (1 error) or $P_{windy}$ is given negative value , the subtree remains in place. Having processed all leaf of its successors, node Outlook is considered for the next pruning. From the pruning procedure $P_{outlook}$ is obtained negative value. So, the tree is unchanged as depicted in Figure 3.8 (c).

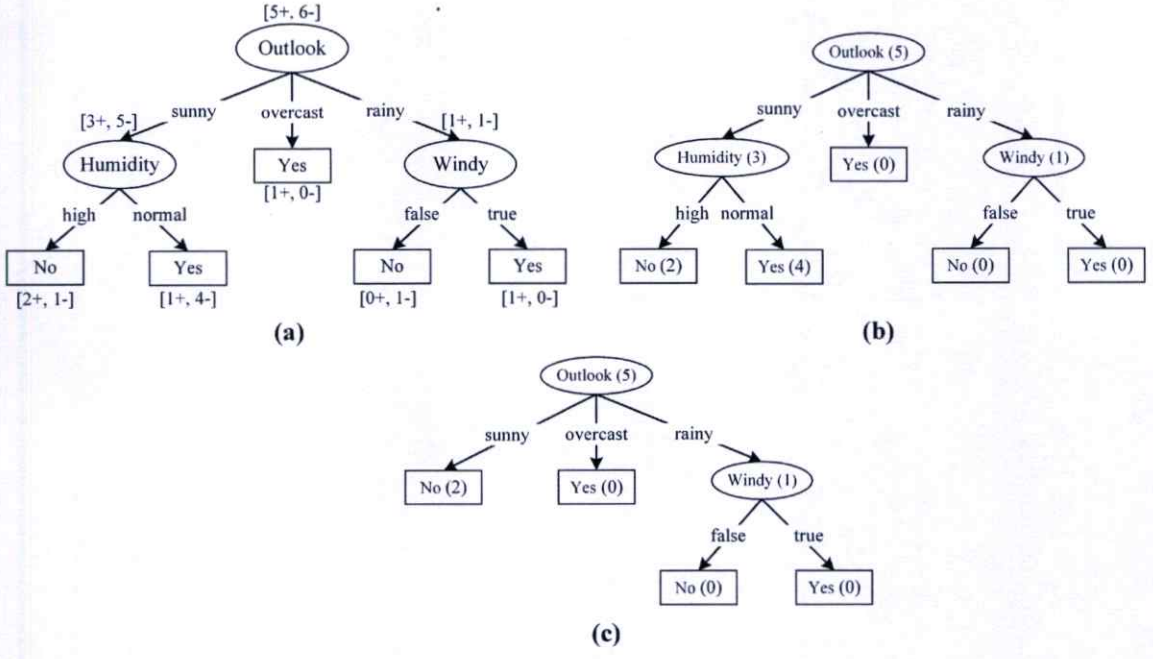



**Figure 3.8:** An example of Reduced Error Pruning Tree

### 3.5.3 Random Tree

Random tree is a supervised learning classifier which presented by Leo Breiman and Adele Cutler [19]. It is an ensemble learning method that generates several individual learners. In addition, Random tree is used a Bagging idea which provides a random set of data for creating a decision tree [18]. In a random tree, each node is split using the best among the subset of predictors randomly selected features at that node [17]. In order to classify, every tree in the forest must be evaluated and given the outcome of the tree with the majority vote. Random Tree can deal with both classification and regression problems. It is a powerful technique that combined the large set of the random tree which leads to the accurate classification models. Random Tree performs no pruning in each node of decision tree.

For example, this experiment is used weather data as shown in Table 3.6. Assume that three decision tree models were constructed based on the same training data as displayed in Figure 3.9. In order to test the model with unseen data, the decision tree classification must be created as shown in Figure 3.10. The experimental results show that the decision tree models 1 and 2 are

correctly classified the target class as $Yes$ and $No$ in the day 15 and 16. However, the decision tree model 3 is given different result in the Day 16 which is provided the outcome $Yes$. To perform the best class label, Random tree is used a majority vote technique. The results found that two-thirds of the models are responded the target class $No$ in the Day 16.

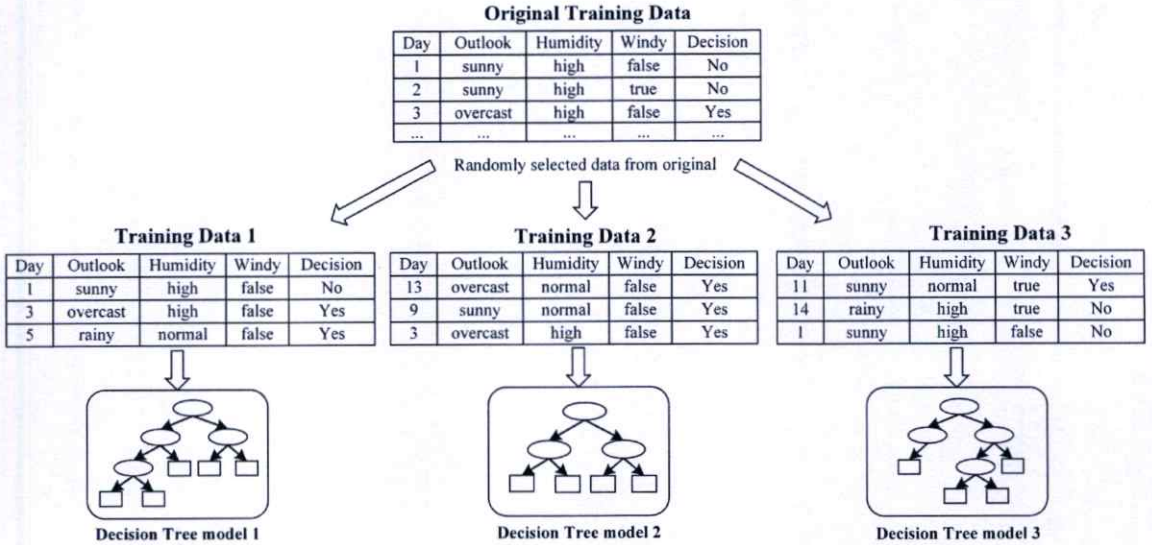



**Figure 3.9:** An example of training model for Random tree

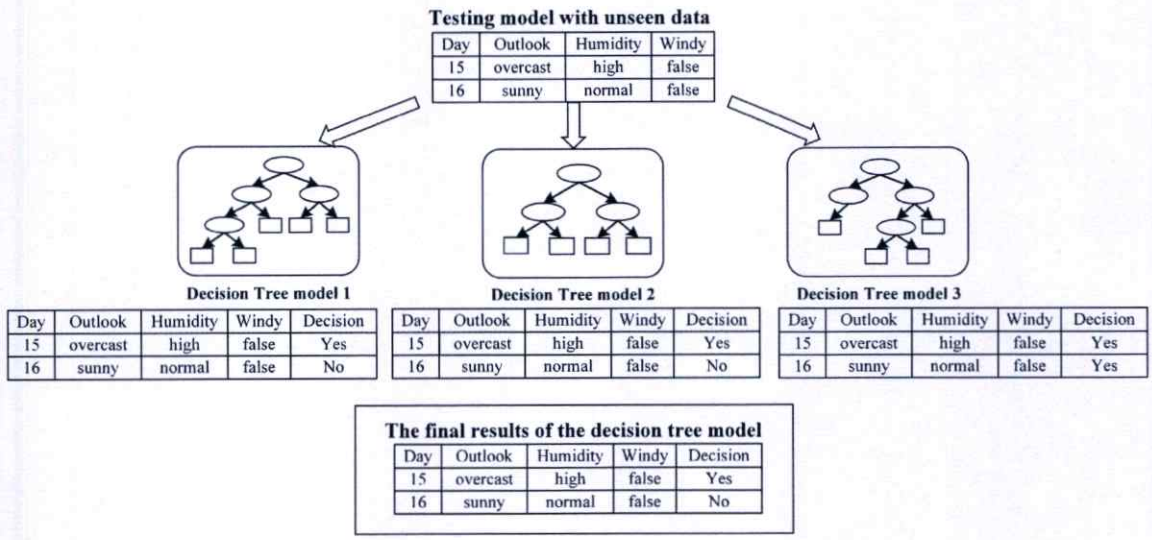



**Figure 3.10:** An example of testing model for Random tree

### 3.6 Cross Validation

In data classification, cross validation is a statistic method for evaluating the model by dividing data into two sections, i.e., learn or train and validate the model. Generally, the training and validation set of cross validation must be cross-over in sequent rounds [20]. Because of each data point can have an occasion to validate. The $k-$fold cross validation is well known as a basic form of cross-validation. In order to train and validated, the data set is randomly split into $k$ equally size segment or folds. Training and validation are done in $k$ iterations such that in each iteration, we leave one-fold for validation while the remaining $k-1$ folds are used for learning. The process of cross-validation is then repeated $k$ times, with each of the $k$ segments used once as the validation data. The $k$ results (true error) from the folds can be average error on the test patterns to produce a single estimation [21]. The average error rate for $k-$fold cross-validation can be calculated by following expression.

$$CV_{avg} = \frac{1}{k}\sum_{i=1}^{k} CV_i \tag{3.9}$$

where $k$ represents the number of folds and $CV_i$ represents the true error rate for each of the $k-$folds.

The major advantage of this technique is that all the patterns in the observation are used for training and validation. For example, Figure 3.11 illustrates an example with $k = 3$. The data of blue part are used for training phase while the data of orange are used for validation (test).

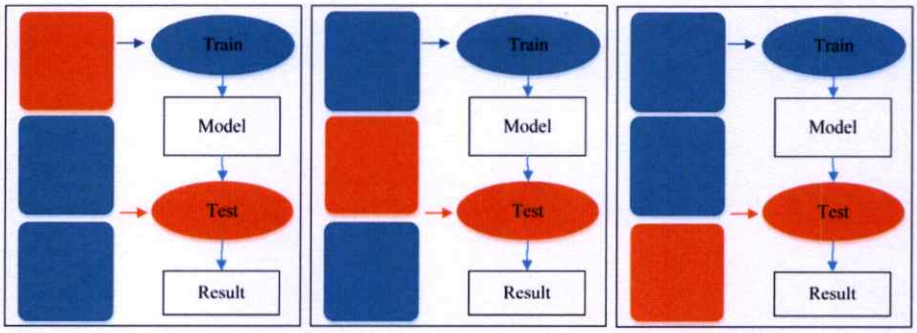



**Figure 3.11:** The example of three-fold cross-validation method

There are two considerations concerning the way to choose the correct number of folds [21]. The bias of true error estimate is large when the number of folds $k$ is small then computational consumption is not heavy. On the other hand, when the number of fold is large, the bias of the true error estimate is small. Unfortunately, due to a large number of iteration then computational consumption are large. 10-fold cross-validation is the most common for model validation.

### 3.7 Evaluation Performance

The aim of this phase is to test and assess the proposed model, i.e., three classifiers have been compared the accuracy results between data preprocessing with and without applying PCA. Among these two approaches, there were a technique which provided the minimal number of rules for breast cancer diagnosis. When the classification accuracy did not achieve the expectations, we rebuild the model by changing some parameters or performed preprocessing phase until the desired results were accomplished.

The most effective approach to evaluate the performance of the model was based on the confusion matrix as shown in Table 3.11. The confusion matrix is a specific table layout that shows the information about actual and predicted value by a classification model. There are four possible classification methods for each instance: a true positive (TP), a true negative (TN), a false positive (FP), and false negative (FN).

**Table 3.11:** The Confusion Matrix

| | Predicted Positive | Predicted Negative |
|---|---|---|
| Actual Positive | TP | FP |
| Actual Negative | FN | TN |

According to confusion matrix, the classification accuracy and F-measure were used to perform in this experiment. In order to test the performance of the model, an accuracy was the most intuitive measurement, i.e., when the model has high accuracy then the model is a good performance. Accuracy can calculate as equation (3.10). It is a ratio of correctly predicted sample from the total sample. However, when the model was performed with imbalanced data, the classification results could give incorrectly classify, e.g., assumed that, there are 10,000 patients that need to be classified as benign or malignant. 9,990 patients can classify as benign and 10 patients were malignant. The classification accuracy was given 99.9 %. Therefore, the only accuracy was not enough for the performance measure. This work had used another approach to evaluate the performance of the models.

The F-measure is an approach that testing the classification accuracy. It is used to balance precision and recall when the values were fluctuated. Supposed that, there were two classifiers that provided the different performance, i.e., one classifier had higher precision but lower recall than other. In order to make them comparable, it would be computed one single result for each of them. Fortunately, F-measure allows comparisons at different levels between recall and precision. It is known as the weighted harmonic mean of the precision and recall of the test. The formula of

F-measure is calculated as equation 3.11.

$$Accuarcy = \frac{TP + TN}{TP + FN + FP + TN} \tag{3.10}$$

$$F - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3.11}$$

### 3.8 Software Development Tool

This study had used a well-known standard machine learning tool such as WEKA. WEKA is an acronym stands for Waikato Environment for Knowledge Analysis. It is an opens source of machine learning software for data analysis which issues under the GNU General Public License. Java language is used for creating the software which developed by University of Waikato, New Zealand. WEKA provides many techniques of machine learning and data mining for data preprocessing, classification, regression, clustering, association rules, and visualization. It also plays an important part to apply Big data [17].

# CHAPTER 4

# DISCUSSION AND EXPERIMENTAL RESULTS

This chapter indicates the experimental results with a new data features which were reduced by applying PCA. It also discusses the benefit and limitation of feature reduction and classifiers to obtain the best classification model for breast cancer diagnosis.

## 4.1 Experimental setup

The experiments were taken in three original data sets derived UCI machine learning repository. The information of each data set was given in Table 4.1. The experiment was used ten-fold cross validation for training and testing in 10 iterations. We compared the accuracy, F-measure, and the number of rules by employing decision tree model including J48, REP Tree and Random Tree classifier through PCA and without PCA. In this work, the performance comparison was relied on statistical tests with confidence interval at 95% confidence level. Weka tool was used to construct and evaluate the models. The Scree Test and KG-Rule were used for selecting the optimal principal component in PCA technique; the Scree Test was applied for WBC and WDBC data sets, and KG-rule was used for WPBC data set.

**Table 4.1:** The details of Wisconsin data set

| No. | Data Set | Malignant | Benign | Total of instances | Attributes |
|---|---|---|---|---|---|
| 1 | WBC | 239 | 444 | 683 | 10 |
| 2 | WDBC | 212 | 357 | 569 | 31 |
| 3 | WPBC | 46 | 148 | 194 | 34 |

## 4.2 Experimental Results

### 4.2.1 Performance results of WBC data set

To increase the classification efficiency, standardized data were applied to use in the PCA process. In order to perform PCA procedure, Eigenvalue and Eigenvector were computed from the correlation matrix. To retain the number of principal components (PCs) from the Eigenvalue, KG-Rule was applied. According to Table B.1 in Appendix B shows the first PC achieved eigenvalues that are greater than 1 by using only 65 % of variance data. Consequently, only one principal component was used in the experiment of WBC data for the classifiers with PCA.

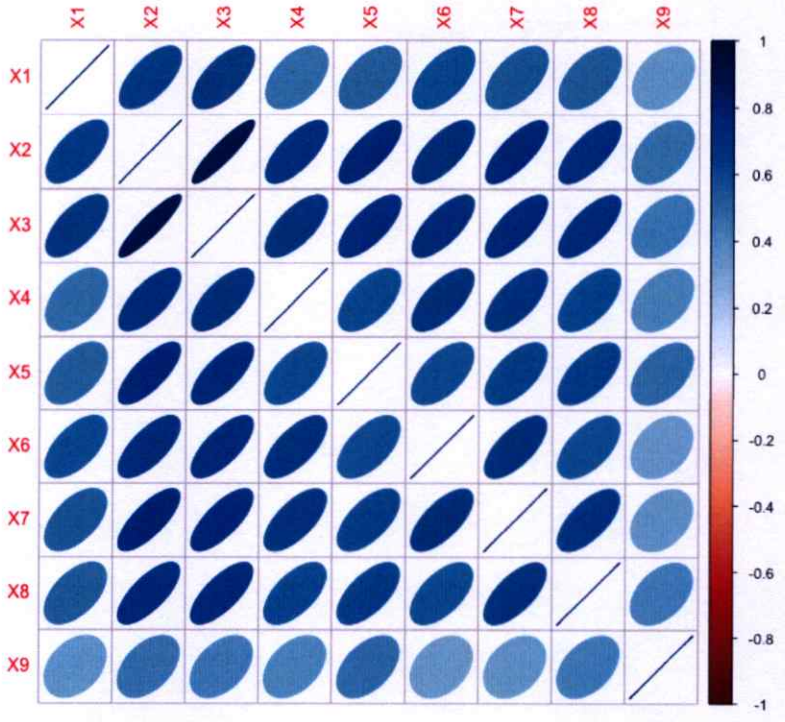



**Figure 4.1:** The correlation matrix of WBC data set

Table 4.2 represents the comparison of rule-based classification accuracy in WBC data set. Without using PCA, 9 features from the original data set are used. In term of PCA, one principal component is applied. In case of with and without applying PCA, there are significantly different values in the group of J48 and REP Tree classifiers individually. However, Random Tree is not significant difference. Additionally, the results also found that J48 classifiers with PCA obtained a better result than other classifiers in term of accuracy and F-measure with the achievement of 97.19 % and 0.9721, respectively as shown in Figure 4.2 and Figure 4.3.

**Table 4.2:** The comparison of accuracy between WBC data set with and without PCA

| Classifiers | Original | | Original+PCA | |
|---|---|---|---|---|
| | Accuracy | F-meausre | Accuracy | F-measure |
| J48 | 95.44±0.52 | 0.9546±0.0051 | 97.19±0.37 | 0.9721±0.0036 |
| REPTree | 95.11±0.46 | 0.9511±0.0045 | 97.13±0.40 | 0.9715±0.0039 |
| Random Tree | 95.71±0.42 | 0.9570±0.0042 | 95.29±0.45 | 0.9527± 0.0046 |

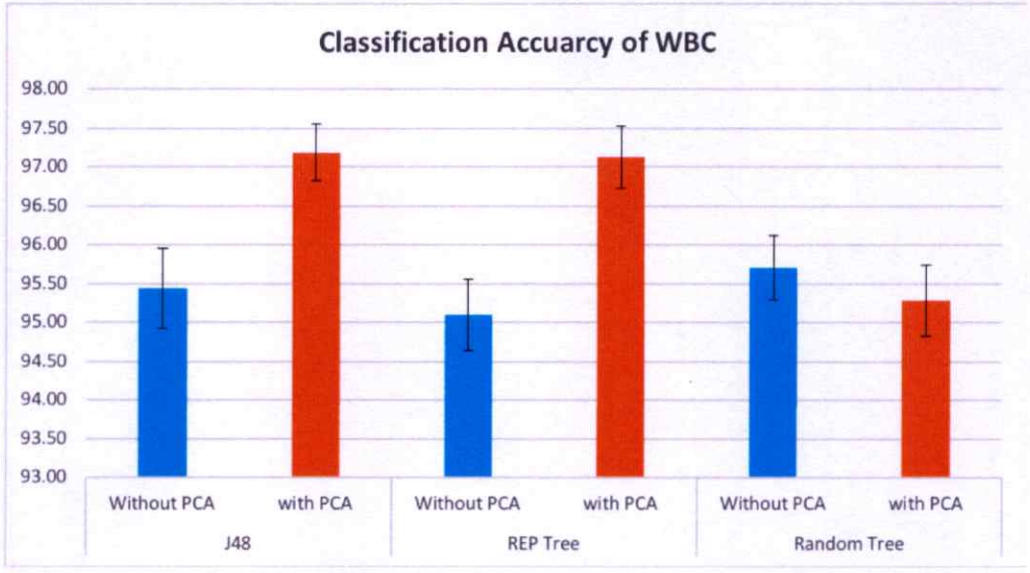



**Figure 4.2:** The comparison of classification accuracy of WBC data set

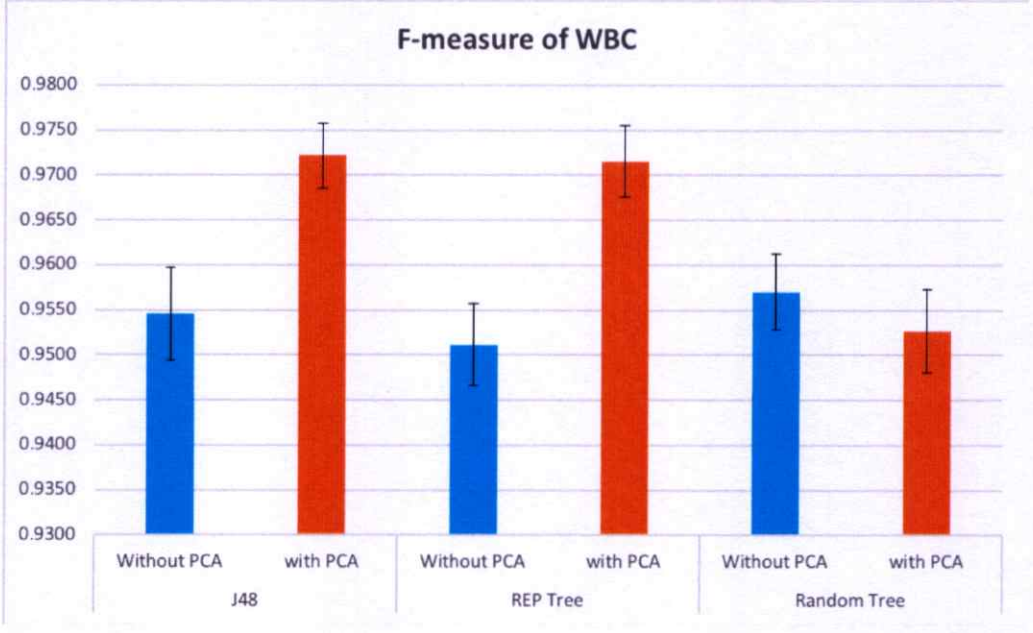



**Figure 4.3:** The comparison of classification F-measure of WBC data set

Figure 4.4 shows the comparison results of classification rules with and without PCA approaches. According to the chart, the number of rules is decreased in case of the classifiers with PCA. Especially, J48 and REP Tree classifier perform significantly in producing the rules better than Random Tree classifiers. They provide only 2 rules while Random Tree classifier needs 28 rules. However, the number of rules still high when PCA is not used, i.e., REP Tree, J48, and Random Tree classifiers need 7, 11, 32 rules, respectively.

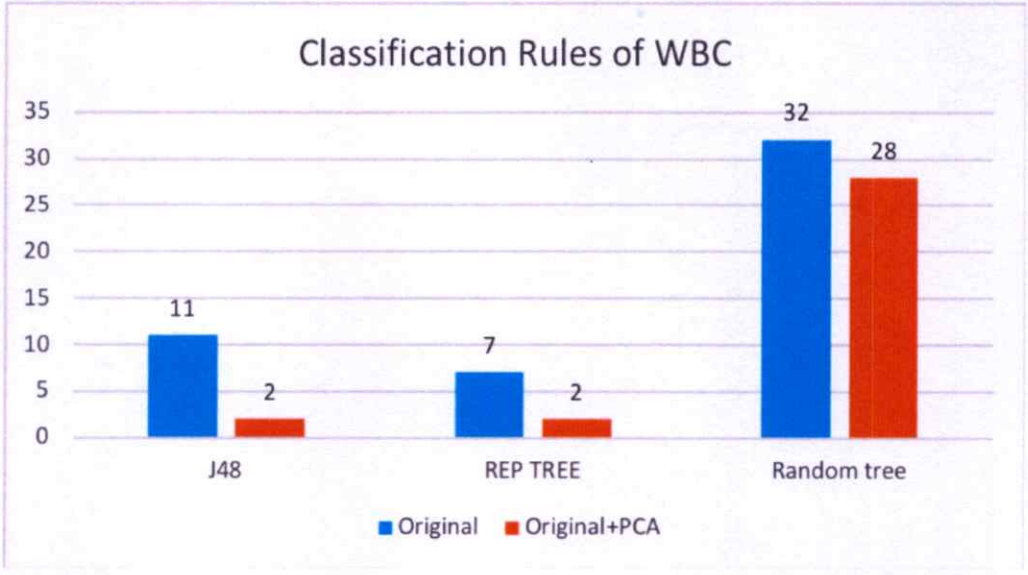



**Figure 4.4:** The comparison of classification rules of WBC

Table 4.3 indicates the results of two classification rules by applying PCA in WBC data set. There is only one principal component used for the classification rules. For example, the first classification rules of J48 classifier is "IF PC1 <= 0.418953 Then malignant ". In this case, one principal component (PC) contains all features in the data set including clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitosis. The value of 0.418953 is implied the Information Gain Ratio value for splitting data. This means that the value of PC1 is less than or equal 0.418953 then we can define as Malignant. For the second rule, the value of PC1 is greater than 0.418953 then we can define as Benign. Additionally, the classification rules of REP Tree and Random Tree can interpret the meaning in the same way as a J48 classifier. But the values of REP and Random Tree classifiers are calculated from Information Gain. e.g., the first classification rule of REP Tree is "IF PC1 < 0.51 Then Malignant". The value of 0.51 is computed from Information Gain. Furthermore, the classification rules of each classifier in WDBC and WPBC data sets are described in the same way with regard to the above. According to our classification accuracy, a J48 classifier with PCA proved to be the most accurate classification rules which produced only 2 rules for this data set. For more details of all classification rules with and without applying PCA in the WBC data set can be seen in Figure C.1 to Figure C.4 in Appendix C.

**Table 4.3:** The classification rules of J48 and REP tree with PCA for WBC data set

| No. | J48 with PCA | REP Tree with PCA |
|---|---|---|
| 1 | IF PC1 <= 0.418953 Then Malignant | IF PC1 <0.51 Then Malignant |
| 2 | IF PC1 >0.418953 Then Benign | IF PC1 >= 0.51 Then Benign |

### 4.2.2 Performance results of WDBC data set

In order to process the original data with PCA, standardized data were applied. The correlation matrix, eigenvalue and eigenvector must be calculated. Figure 4.5 shows the correlation matrix of WDBC data. According to Table B.2 in Appendix B, KG rule was applied to select the principal component for a new data. The results found that, the first sixth PC attains eigenvalues that are greater than 1. Thus, 6 PCs were retained to use in this experiment.

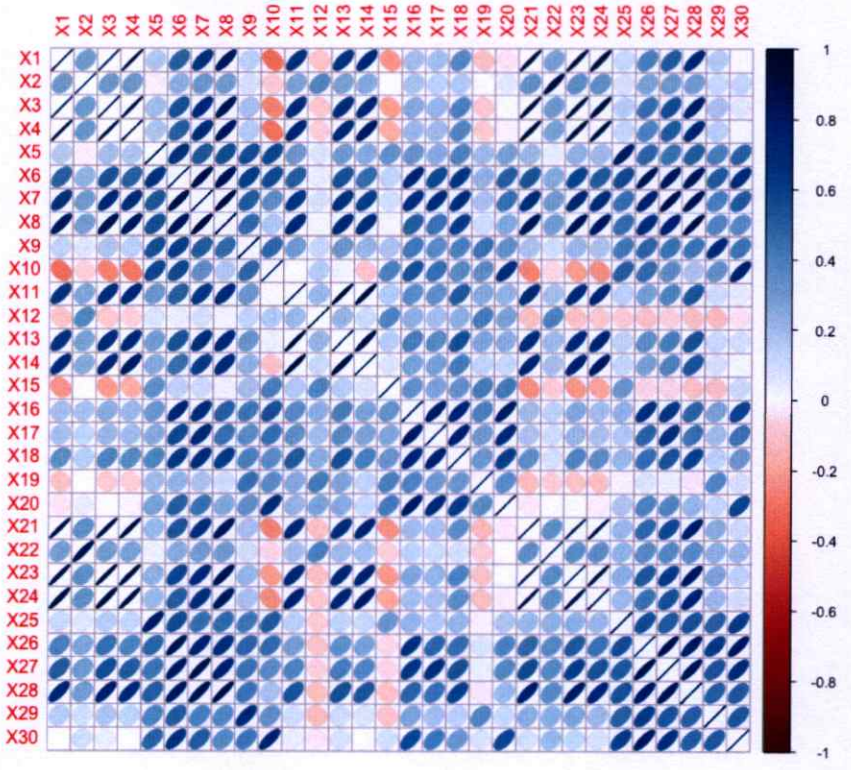



**Figure 4.5:** The correlation matrix of WDBC data set

**Table 4.4:** Comparison of accuracy results between WDBC data set of original and original+ PCA

| Classifiers | Original | | Original+PCA | |
|---|---|---|---|---|
| | Accuracy | F-meausre | Accuracy | F-measure |
| J48 | 93.27±0.70 | 0.9326±0.0070 | 94.84±0.64 | 0.9482±0.0064 |
| REPTree | 93.08±0.70 | 0.9306±0.0070 | 93.52±0.68 | 0.9351±0.0069 |
| Random Tree | 92.78±0.63 | 0.9277±0.0063 | 93.91±0.65 | 0.9390±0.0066 |

Table 4.4 describes the comparison of rule-based classification accuracy in WDBC data set. For this experiment, 6 principal components are applied. In contrast, 30 features are utilized. There are significantly difference in only the group of J48. However, the group of REP tree

and Random are not significant difference. The results indicated that J48 classifier with PCA performed the best among the other classifiers by achieving 94.84 % accuracy and 0.9492 F-measure as shown in Figure 4.6 and Figure 4.7.

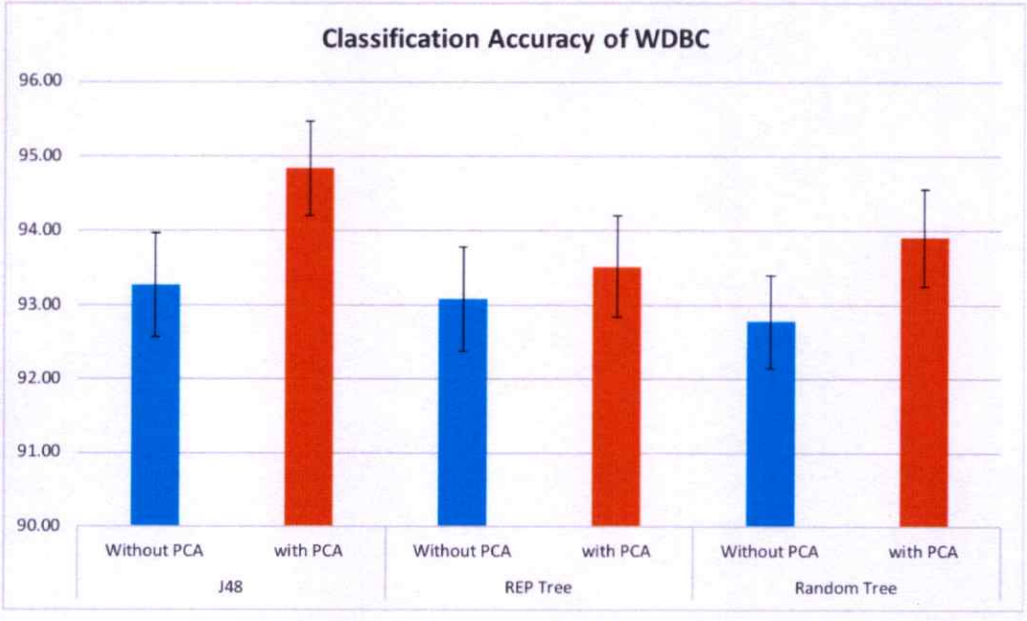



**Figure 4.6:** The comparison of classification accuracy of WDBC

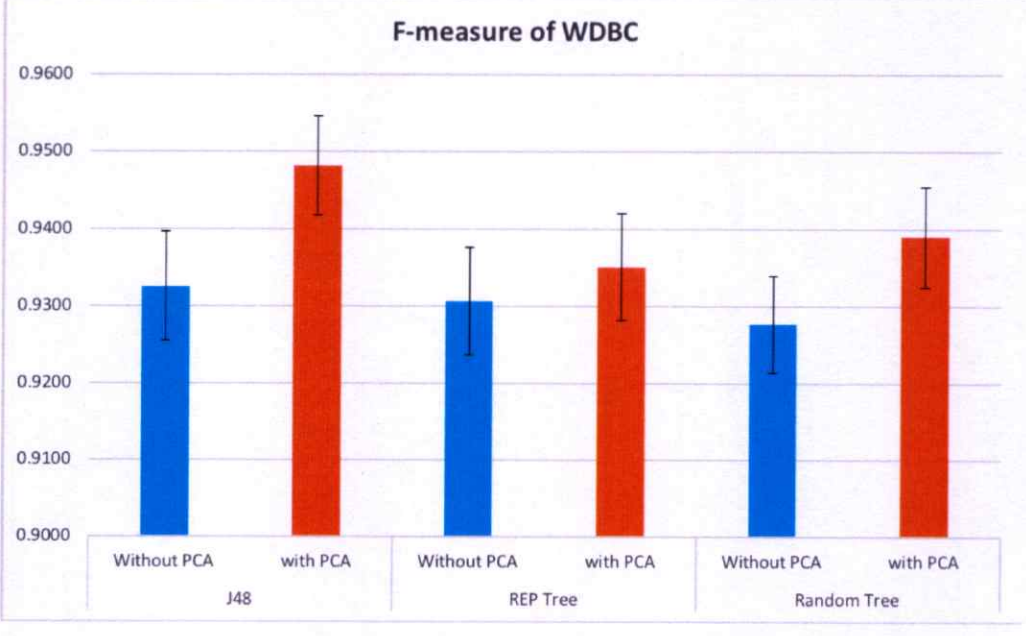



**Figure 4.7:** The comparison of classification F-measure of WDBC

Figure 4.8 illustrates the comparison results of classification rules with and without PCA approaches. According to the chart, the classification rules after applying PCA are fluctuated. Furthermore, REP Tree classifier provides the minimal number of rules among them with and without PCA. However, their accuracy of classification performance was less than J48 classifier when using PCA. Hence, the classification rules of J48 classifier with PCA was the best rules for

this classification model. The details of all the classification rules can be seen in Figure C.5 to Figure C.10 in Appendix C.

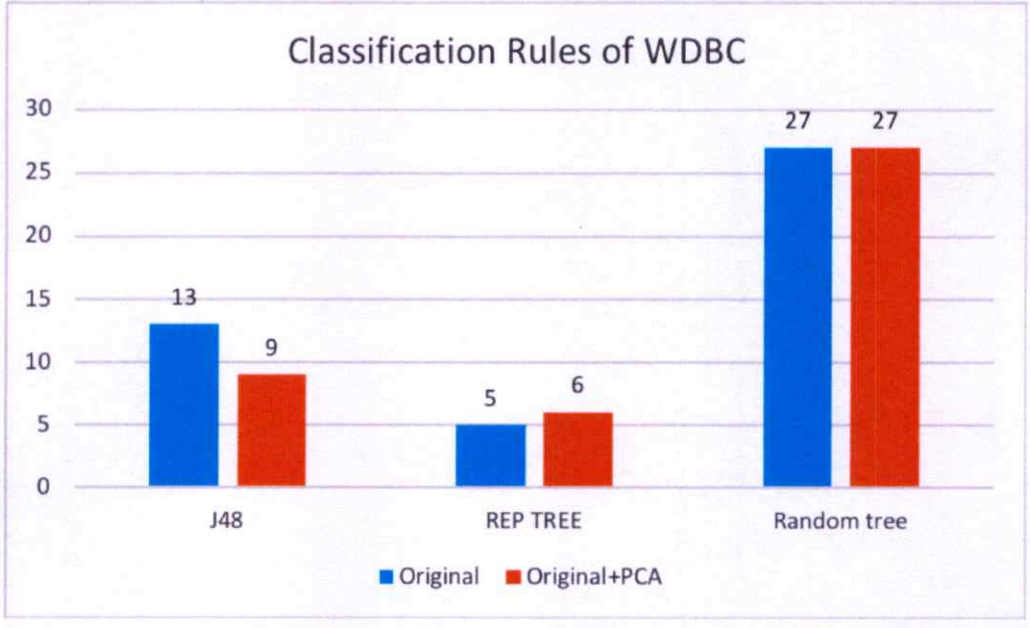



**Figure 4.8:** The comparison of classification rules of WDBC

### 4.2.3 performance results of WPBC data set

In order to analyze PCA, the correlation matrix, eigenvalue and eigenvector must be calculated. The correlation coefficient matrix is shown in Figure 4.10 and the eigenvalue is given in Table B.3 of Appendix B. To select the best principal component is not only based on KG Rule but also on Scree Test. Sine there were unbalance data in our experiment. Thus, this experiment is carried out by Scree test to define PCs. Figure 4.9 shows the graph that should be maintained for 9 PCs.

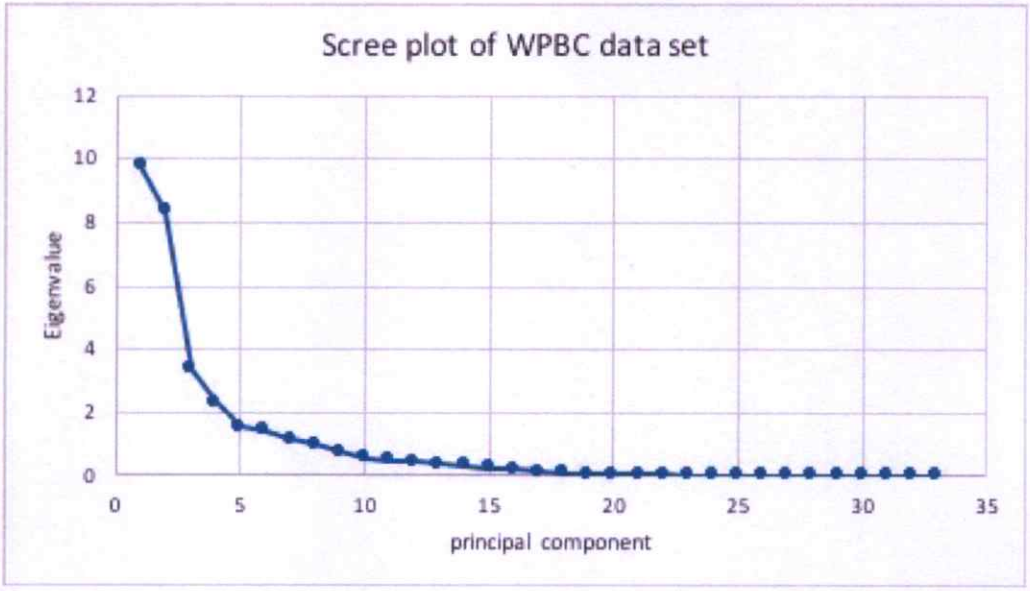



**Figure 4.9:** The eigenvalues of WDBC data set

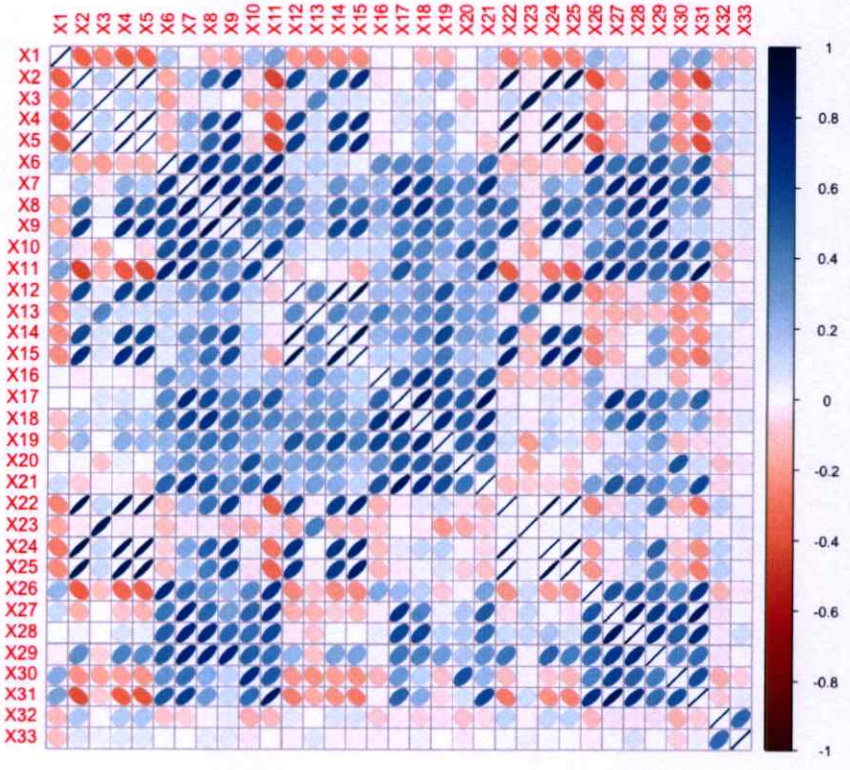



**Figure 4.10:** The correlation matrix of WDBC data set

Table 4.5 indicates the comparison of rule-based classification accuracy in WPBC data set. There are 9 principal components in case of PCA, otherwise 33 features are applied. According to the data set is imbalanced, this experiment focuses only the F-measure to evaluate the performance. The results revealed that J48, REP Tree and Random Tree classifiers with and without PAC were not significant difference like illustrated in Figure 4.12.

**Table 4.5:** Comparison of accuracy results between WPBC data set of original and original+ PCA

| Classifiers | Original | | Original+PCA | |
|---|---|---|---|---|
| | Accuracy | F-meausre | Accuracy | F-measure |
| J48 | 73.24±1.87 | 0.7203±0.0179 | 75.77±1.54 | 0.7384±0.0169 |
| REPTree | 76.25±1.15 | 0.6851±0.0148 | 76.70±1.10 | 0.7032±0.0148 |
| Random Tree | 67.57±2.06 | 0.6754±0.0189 | 68.36±2.07 | 0.6811±0.0204 |

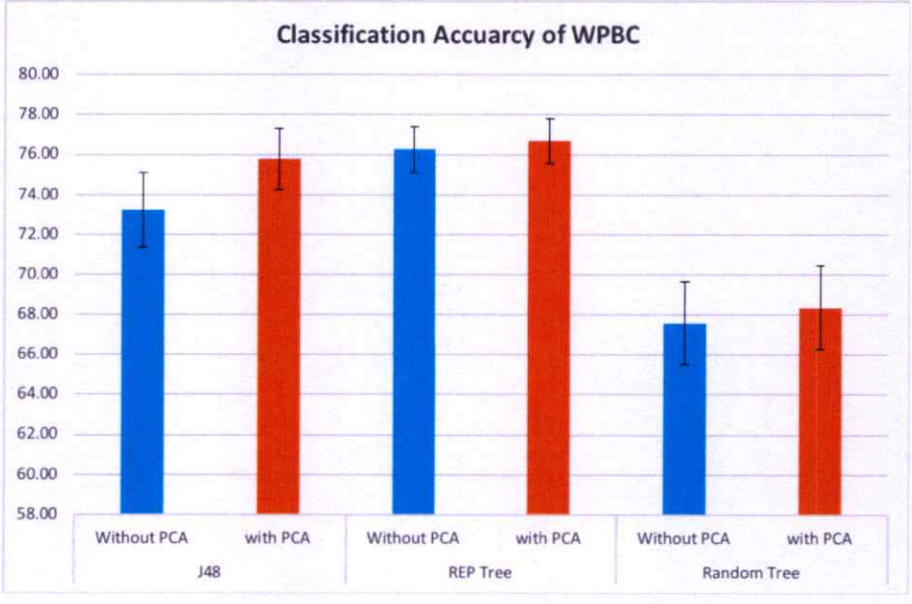



**Figure 4.11:** The comparison of classification accuracy of WPBC

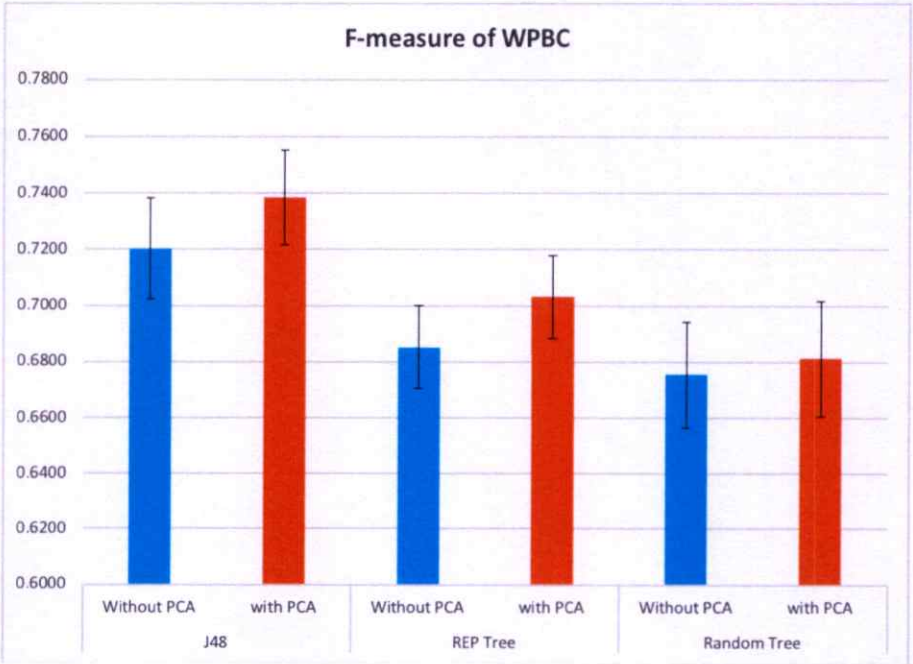



**Figure 4.12:** The comparison of classification F-measure of WPBC

Figure 4.13 indicates the comparison results of classification rules with and without PCA approaches. According to the chart, the numbers of rules are increased in case of the classifiers with PCA; specifically, J48 and Random Tree classifiers. In contrast, REP Tree performs significantly in producing the rules better than J48 and Random Tree classifiers. The details of all the classification rules can be shown in Figure C.11 to Figure C.16 in Appendix C.

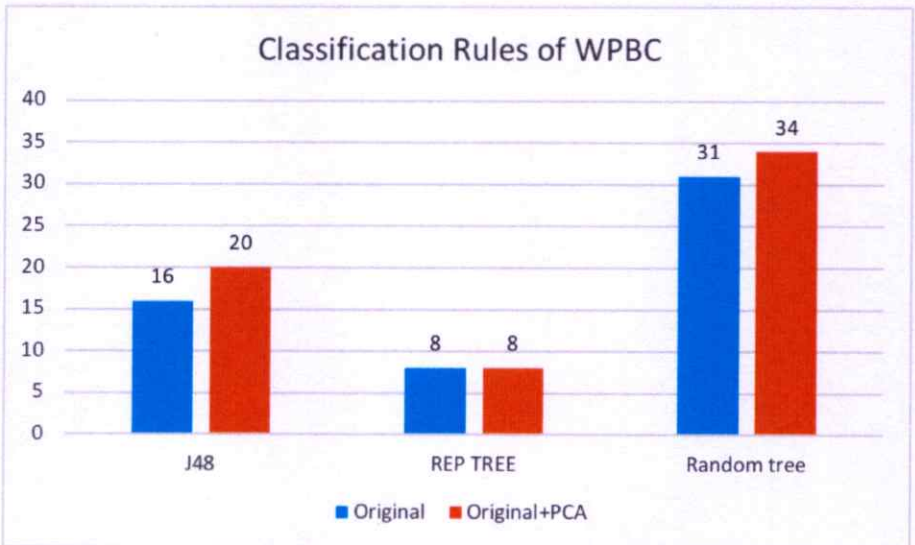



**Figure 4.13:** The comparison of classification rules of WPBC

## 4.3 Discussion

This study focused on the best performance of classifiers with minimal classification rules. According to our experimental results in section 4.2, the accuracy of classification performance and the number of rules in each technique were given different results. Among the previous works in the WBC data set, the J48 classifier showed the best performance in classification accuracy [8] . To compare with them, this study obviously shown that J48 classifiers with PCA performed better than the previous works since the accuracy has been improved by 2.63 % from 94.56% to 97.19%. Regard to our experiment, J48 with PCA also provided only 2 rules for the classification, while J48 without applying PCA needed 11 rules with an achievement of 95.44% accuracy. Additionally, N. Sharma and K. Saroha [3] found the most effective features for learning the classification model in WDBC dataset by using KNN classifier with PCA and Feature Evaluation and Ranking. Their work could classify 93.14 % correctly with minimum 5 features. In contrast, our results found that the J48 classifier with PCA (applying 6 principal components) achieved a higher classification accuracy of 1.7 % increased. In addition, For WPBC data set, P. Hamsagayathri and P. Sampath [6] indicated that the classification model obtained 83.83 % of accuracy by using Priority Based Tree. However, REP Tree with PCA in our experiment achieved 76.70 % accuracy. According to our experimental results, the decision tree model by applying PCA are significantly better performance compared to the previous decision tree model with original data. The reason why those three models did not improve a better result when applied PCA was that the data features had weak correlation in the WPBC data set.

Further consideration there are two main alternative parts which supported our great classification model. The first main point is on data preprocessing step. In this procedure, PCA technique was applied as mentioned in the previous paragraph. Due to each data structure has its own different features and sizes as shown in Table 4.1. For WBC data set, the relationship between each feature has very strong positive correlation as given in Figure 4.1. Hence, an application of the principal components analysis could give better performance in this data set. However, the other two data sets of WDBC and WPBC had a quite weak correlation as illustrated in Figure 4.5 and 4.10 . Especially, the data set of WPBC is very small sample size and imbalanced data. Thus, the traditional classification algorithms could not perform well. This indicated that the sample size and imbalanced data of training samples which have an impact on the classification accuracy for the classifiers. The second reason relates to the variation in classification accuracy and the number of rules across different classifiers in different data sets. In each decision tree classifier, they also have the advantages and disadvantages to classify the different data type such as numeri-

cal, continuous, categorical, etc. For example, J48 classifier creates the node by using information gain ratio to obtain the best node. REP tree is built the decision tree by using information gain and prunes it using reduce error pruning technique. Random tree classifier is an ensemble method by creating multiple models of decision tree, random sampling in each node and it performs no pruning. Therefore, Random Tree provided the largest number of rules for all data set in this experiment.

# CHAPTER 5

# CONCLUSION AND RECOMMENDATION

A rule-based system for breast cancer diagnosis has been a powerful tool supporting doctor diagnosis. Such a system requires classification rules derived from historical diagnosis. The desirable rules should be minimal in their number and give a good performance. This study is to obtain such rules from the Wisconsin Breast cancer (WBC, WDBC, and WPBC) data set. It performed experiments on the data set with PCA feature reduction to determine the best classifier among J48 decision tree, REP Tree, and Random Tree.

Experimental results on WBC and WDBC data sets with PCA show that J48 classifier giving the best accuracy and smallest number of rules. However, this conclusion is not valid for WPBC data set with and without PCA. This is because the WPBC data set is an imbalance data set. The imbalance techniques such as the SMOTE (synthetic minority oversampling technique) algorithm is recommended to be used to improve the classification performance of such a data set.

# REFERENCES


[1] G. Sujatha and K. U. Rani, "Evaluation of decision tree classifiers on tumor data sets," *International Journal of Emerging Trends and Technology in Computer Science (IJETTCS)*, vol. 2, pp. 418 – 423, 2013.

[2] S. Jhajharia, H. K. Varshney, S. Verma, and R. Kumar, "A neural network based breast cancer prognosis model with pca processed features," *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 1896 –1901, 2016.

[3] N. Sharma and K. Saroha, "A novel dimensionality reduction method for cancer dataset using pca and feature ranking," *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 2261 – 2264, 2015.

[4] T. M. Mohamed, "Efficient breast cancer detection using sequential feature selection techniques," *7th International Conference on Intelligent Computing and Information Systems (ICICIS)*, pp. 458 – 463, 2015.

[5] C. L. Sabharwal and B. Anjum, "Principal component analysis as an integral part of data mining in health informatics," *Proceedings of 31th International Society Conference on Computers And Their Application (CATA)*, pp. 251 – 256, 2016.

[6] P. Hamsagayathri and P. Sampath, "Performance analysis of breast cancer classification using decision tree classifier," *International Journal Of Current Pharmaceutical Research (IRCPR)*, vol. 9, no. 2, 2017.

[7] H. Elouedi, W. Meliani, Z. Elouedi, and N. Ben Amor, "A hybrid approach based on decision trees and clustering for breast cancer classification," *6th International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, 2014.

[8] R. Sumbaly, N. Vishnusri, and S. Jeyalatha, "Diagnosis of breast cancer using decision tree data mining technique," *International Journal of Computer Applications*, vol. 98, no. 10, pp. 0975 – 8887, 2014.

[9] M. Lichman, "UCI machine learning repository, university of california, irvine, school of information and computer sciences," accessed February 3, 2018. [Online]. Available: {http://archive.ics.uci.edu/ml}

[10] "UCI machine learning repository," accessed February 3, 2018. [Online]. Available: {https://archive.ics.uci.edu/ml/datasets.html}

[11] J. Han and M. Kamber, "Data mining concepts and techniques," 2006. [Online]. Available: {http://ccs1.hnue.edu.vn/hungtd/DM2012/DataMining_BOOK.pdf}

[12] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, pp. 37–52, 1987.

[13] H. Hasan and N. M. Tahir, "Feature selection of breast cancer based on principal component analysis," *6th International Colloquium on Signal Processing and Its Application(CSPA)*, pp. 242–245, 2010.

[14] I. Jolliffe, *Principal Component Analysis*. Springer Series, 2001, pp. 111–137.

[15] T. R. Patil and S. S. Sherekar, "Performance analysis of naive bayes and j48 classification algorithm for data classification," *International Journal Of Computer Science And Applications*, vol. 6, no. 2, pp. 256–261, 2013.

[16] B. Hssina, A. Merbouha, H. Ezzikouri, and M. Erritali, "A comparative study of decision tree id3 and c4.5," *International Journal of Advanced computer Science and Applications*, 2014.

[17] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005.

[18] S. Kalmegh, "Analysis of weka data mining algorithm reptree, simple cart and randomtree for classification of indian news," *International Journal of Innovative Science Engineering and Technology (IJISET)*, vol. 2, pp. 438–1037, 2015.

[19] S. R. Kalmegh, "Comparative analysis of weka data mining algorithm random forest, random tree and ladtree for classification of indigenous news data," *International Journal of Emerging Technology and Advanced Engineering*, vol. 5, 2015.

[20] P. Refaeilzadeh, L. Tang, and H. Liu, *Cross-Validation*. Boston and MA: Springer US, 2009, pp. 532–538. [Online]. Available: {https://doi.org/10.1007/978-0-387-39940-9_565}

[21] R. kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," *International Joint Conference on Artificial intelligence (IJCAI)*, 1995.

# APPENDICES

# APPENDIX A Data set

**Table A.1:** Data set of Wisconsin (Original) or WBC

| Attribute | Values |
|---|---|
| Sample code number | 1-10 |
| Clump Thickness | 1-10 |
| Uniformity of Cell Size | 1-10 |
| Uniformity of Cell Shape | 1-10 |
| Marginal Adgesion | 1-10 |
| Single Epithelial Cell Size | 1-10 |
| Bare Nuclei | 1-10 |
| Bland Chromatin | 1-10 |
| Normal Nucleoli | 1-10 |
| Mitoses | 1-10 |
| Class | 2 for benign, 4 for malignant |

**Table A.2:** Data set of Wisconsin (Diagnostic) or WDBC

| **Attribute** | **Values** |
|---|---|
| ID Number | Numeric |
| Diagnosis | M= malignant, B=benign |
| Radius(mean, standard error and worst) | Numeric |
| texture (mean, standard error and worst) | Numeric |
| perimeter (mean, standard error and worst) | Numeric |
| area (mean, standard error and worst) | Numeric |
| smoothness (mean, standard error and worst) | Numeric |
| compactness (mean, standard error and worst) | Numeric |
| concavity (mean, standard error and worst) | Numeric |
| concave points (mean, standard error and worst) | Numeric |
| symmetry (mean, standard error and worst) | Numeric |
| fractal dimension (mean, standard error and worst) | Numeric |

**Table A.3:** Data set of Wisconsin (prognostic) or WPBC

| **Attribute** | **Values** |
|---|---|
| ID Number | Numeric |
| Outcome | R= recur, N= no recur |
| Time | recurrence time if field 2 ='R', disease-free time if field 2 = 'N' |
| Radius(mean, standard error and worst) | Numeric |
| texture (mean, standard error and worst) | Numeric |
| perimeter (mean, standard error and worst) | Numeric |
| area (mean, standard error and worst) | Numeric |
| smoothness (mean, standard error and worst) | Numeric |
| compactness (mean, standard error and worst) | Numeric |
| concavity (mean, standard error and worst) | Numeric |
| concave points (mean, standard error and worst) | Numeric |
| symmetry (mean, standard error and worst) | Numeric |
| fractal dimension (mean, standard error and worst) | Numeric |

# APPENDIX B Eigenvalues

**Table B.1:** The eigenvalues of WBC data set

| No. | Eigenvalue | Proportion | Cumulative |
|---|---|---|---|
| 1 | 5.89950 | 0.65550 | 0.65550 |
| 2 | 0.77595 | 0.08622 | 0.74172 |
| 3 | 0.53925 | 0.05992 | 0.80163 |
| 4 | 0.45963 | 0.05107 | 0.85270 |
| 5 | 0.38028 | 0.04225 | 0.89496 |
| 6 | 0.30188 | 0.03354 | 0.92850 |
| 7 | 0.29440 | 0.03271 | 0.96121 |
| 8 | 0.26074 | 0.02897 | 0.99018 |
| 9 | 0.08838 | 0.00982 | 1 |

**Table B.2:** The eigenvalues of WDBC data set

| No. | Eigenvalue | Proportion | Cumulative |
|---|---|---|---|
| 1 | 13.28161 | 0.44272 | 0.44272 |
| 2 | 5.69135 | 0.18971 | 0.63243 |
| 3 | 2.81795 | 0.09393 | 0.72636 |
| 4 | 1.98064 | 0.06602 | 0.79239 |
| 5 | 1.64873 | 0.05496 | 0.84734 |
| 6 | 1.20736 | 0.04025 | 0.88759 |
| 7 | 0.67522 | 0.02251 | 0.91010 |
| ... | ... | ... | ... |
| 30 | 0.00013 | 0 | 1 |

**Table B.3:** The eigenvalues of WPBC data set

| No. | Eigenvalue | Proportion | Cumulative |
|---|---|---|---|
| 1 | 9.78359 | 0.29647 | 0.29647 |
| 2 | 8.35713 | 0.25325 | 0.54972 |
| 3 | 3.35825 | 0.10177 | 0.65148 |
| 4 | 2.33398 | 0.07073 | 0.72221 |
| 5 | 1.54946 | 0.04695 | 0.76916 |
| 6 | 1.41587 | 0.04291 | 0.81207 |
| 7 | 1.16427 | 0.03528 | 0.84735 |
| 8 | 1.00399 | 0.03042 | 0.87777 |
| 9 | 0.7742 | 0.02346 | 0.90123 |
| 10 | 0.55275 | 0.01675 | 0.91798 |
| ... | ... | ... | ... |
| 33 | 0.00036 | 0.00001 | 1 |

# APPENDIX C Classification Rules of Wisconsin data set

| No. | Classification rules of J48 Decision Tree |
|---|---|
| 1 | IF uniformity cell size <=2 and Bare Nuclei <=3 Then Benign |
| 2 | IF uniformity cell size <=2 and Bare Nuclei > 3 and Clump Thickness <= 3 Then Benign |
| 3 | IF uniformity cell size <=2 and Bare Nuclei > 3 and Clump Thickness > 3 and Bland Chromatin <=2 and Marginal Adhesion <= 3 Then Malignant |
| 4 | IF uniformity cell size <=2 and Bare Nuclei > 3 and Clump Thickness > 3 and Bland Chromatin <=2 and Marginal Adhesion > 3 Then Benign |
| 5 | IF uniformity cell size <=2 and Bare Nuclei > 3 and Clump Thickness > 3 and Bland Chromatin > 2 Then Malignant |
| 6 | IF Uniformity Cell Size > 2 and Uniformity Cell Shape <= 2 and Clump Thickness <= 5 Then Benign |
| 7 | IF Uniformity Cell Size > 2 and Uniformity Cell Shape <= 2 and Clump Thickness > 5 Then Malignant |
| 8 | IF Uniformity Cell Size > 2 and Uniformity Cell Shape > 2 and Uniformity Cell Size <= 4 and Bare Nuclei <= 2 and Marginal Adhesion <= 3 Then Benign |
| 9 | IF Uniformity Cell Size > 2 and Uniformity Cell Shape > 2 and Uniformity Cell Size <= 4 and Bare Nuclei <= 2 and Marginal Adhesion <= 3 Then Benign |
| 10 | IF Uniformity Cell Size > 2 and Uniformity Cell Shape > 2 and Uniformity Cell Size <= 4 and Bare Nuclei <= 2 and Marginal Adhesion > 3 Then Malignant |
| 11 | IF Uniformity Cell Size > 2 and Uniformity Cell Shape > 2 and Uniformity Cell Size > 4 Then Malignant |

**Figure C.1:** The rules of J48 decision tree without PCA in WBC data set

| No. | Classification rules of REP Tree |
|---|---|
| 1 | IF Uniformity Cell Shape < 3.5 and Bare Nuclei < 1.5 Then Benign |
| 2 | IF Uniformity Cell Shape < 3.5 and Bare Nuclei >=1.5 and Clump Thickness < 5.5 and Uniformity Cell Shape < 2.5 Then Benign |
| 3 | IF Uniformity Cell Shape < 3.5 and Bare Nuclei >=1.5 and Clump Thickness < 5.5 and Uniformity Cell Shape >= 2.5 Then Malignant |
| 4 | IF Uniformity Cell Shape < 3.5 and Bare Nuclei >=1.5 and Clump Thickness >= 5.5 Then Malignant |
| 5 | IF Uniformity Cell Shape >= 3.5 and Uniformity Cell Size < 4.5 and Bare Nuclei < 1.5 Then Benign |
| 6 | IF Uniformity Cell Shape >= 3.5 and Uniformity Cell Size < 4.5 and Bare Nuclei >= 1.5 Then Malignant |
| 7 | IF Uniformity Cell Shape >= 3.5 and Uniformity Cell Size >= 4.5 Then Malignant |

**Figure C.2:** The rules of REP tree without PCA in WBC data set

| No | Classification rules of Random Tree |
|---|---|
| 1 | IF Bare Nuclei < 2.5 and Uniformity Cell Shape < 3.5 and Bland Chromatin >= 3.5 and Mitoses < 6 Then Benign |
| 2 | IF Bare Nuclei < 2.5 and Uniformity Cell Shape < 3.5 and Bland Chromatin >= 3.5 and Mitoses < 6 Then Benign |
| 3 | IF Bare Nuclei < 2.5 and Uniformity Cell Shape < 3.5 and Bland Chromatin >= 3.5 and Mitoses >= 6 Then Malignant |
| 4 | IF Bare Nuclei < 2.5 and Uniformity Cell Shape >= 3.5 and Marginal Adhesion < 3.5 and Uniformity Cell Size < 4.5 and Clump Thickness < 5.5 Then Benign |
| 5 | IF Bare Nuclei < 2.5 and Uniformity Cell Shape >= 3.5 and Marginal Adhesion < 3.5 and Uniformity Cell Size < 4.5 and Clump Thickness >= 5.5 and Uniformity Cell Size < 3.5 Then Malignant |
| 6 | IF Bare Nuclei < 2.5 and Uniformity Cell Shape >= 3.5 and Marginal Adhesion < 3.5 and Uniformity Cell Size < 4.5 and Clump Thickness >= 5.5 and Uniformity Cell Size >= 3.5 Then Benign |
| 7 | IF Bare Nuclei < 2.5 and Uniformity Cell Shape >= 3.5 and Marginal Adhesion < 3.5 and Uniformity Cell Size >= 4.5 Then Malignant |
| 8 | IF Bare Nuclei < 2.5 and Uniformity Cell Shape >= 3.5 and Marginal Adhesion >= 3.5 Then Malignant |
| 9 | IF Bare Nuclei >= 2.5 and Uniformity Cell Size < 1.5 and Clump Thickness < 5.5 and Single Epithelial Cell Size < 1.5 and Bland Chromatin < 1.5 Then Benign |
| 10 | IF Bare Nuclei >= 2.5 and Uniformity Cell Size < 1.5 and Clump Thickness < 5.5 and Single Epithelial Cell Size < 1.5 and Bland Chromatin >= 1.5 Then Malignant |
| 11 | IF Bare Nuclei >= 2.5 and Uniformity Cell Size < 1.5 and Clump Thickness < 5.5 and Single Epithelial Cell Size >= 1.5 Then Benign |
| 12 | IF Bare Nuclei >= 2.5 and Uniformity Cell Size < 1.5 and Clump Thickness >= 5.5 Then Malignant |
| 13 | IF Bare Nuclei >= 2.5 and Uniformity Cell Size >= 1.5 and Uniformity Cell Shape < 2.5 and Bland Chromatin < 3.5 and Clump Thickness < 7.5 Then Benign |
| 14 | IF Bare Nuclei >= 2.5 and Uniformity Cell Size >= 1.5 and Uniformity Cell Shape < 2.5 and Bland Chromatin < 3.5 and Clump Thickness >= 7.5 Then Malignant |
| 15 | IF Bare Nuclei >= 2.5 and Uniformity Cell Size >= 1.5 and Uniformity Cell Shape < 2.5 and Bland Chromatin >= 3.5 Then Malignant |
| 16 | IF Bare Nuclei >= 2.5 and Uniformity Cell Size >= 1.5 and Uniformity Cell Shape >= 2.5 and Marginal Adhesion < 5.5 and Bare Nuclei < 8.5 and Normal Nucleoli < 8.5 and Clump Thickness < 8.5 and Uniformity Cell Shape < 3.5 Then Malignant |
| 17 | IF Bare Nuclei >= 2.5 and Uniformity Cell Size >= 1.5 and Uniformity Cell Shape >= 2.5 and Marginal Adhesion < 5.5 and Bare Nuclei < 8.5 and Normal Nucleoli < 8.5 and Clump Thickness < 8.5 and Uniformity Cell Shape >= 3.5 and Marginal Adhesion < 3.5 and Bare Nuclei < 6.5 and Clump Thickness < 6.5 Then Benign |
| 18 | IF Bare Nuclei >= 2.5 and Uniformity Cell Size >= 1.5 and Uniformity Cell Shape >= 2.5 and Marginal Adhesion < 5.5 and Bare Nuclei < 8.5 and Normal Nucleoli < 8.5 and Clump Thickness < 8.5 and Uniformity Cell Shape >= 3.5 and Marginal Adhesion < 3.5 and Bare Nuclei < 6.5 and Clump Thickness >= 6.5 Then Malignant |
| 19 | IF Bare Nuclei >= 2.5 and Uniformity Cell Size >= 1.5 and Uniformity Cell Shape >= 2.5 and Marginal Adhesion < 5.5 and Bare Nuclei < 8.5 and Normal Nucleoli < 8.5 and Clump Thickness < 8.5 and Uniformity Cell Shape >= 3.5 and Marginal Adhesion < 3.5 and Bare Nuclei >= 6.5 and Single Epithelial Cell Size < 4.5 Then Malignant |
| 20 | IF Bare Nuclei >= 2.5 and Uniformity Cell Size >= 1.5 and Uniformity Cell Shape >= 2.5 and Marginal Adhesion < 5.5 and Bare Nuclei < 8.5 and Normal Nucleoli < 8.5 and Clump Thickness < 8.5 and Uniformity Cell Shape >= 3.5 and Marginal Adhesion < 3.5 and Bare Nuclei >= 6.5 and Single Epithelial Cell Size >= 4.5 and Clump Thickness < 4 Then Malignant |
| 21 | IF Bare Nuclei >= 2.5 and Uniformity Cell Size >= 1.5 and Uniformity Cell Shape >= 2.5 and Marginal Adhesion < 5.5 and Bare Nuclei < 8.5 and Normal Nucleoli < 8.5 and Clump Thickness < 8.5 and Uniformity Cell Shape >= 3.5 and Marginal Adhesion < 3.5 and Bare Nuclei >= 6.5 and Single Epithelial Cell Size >= 4.5 and Clump Thickness >= 4 Then Benign |
| 22 | IF Bare Nuclei >= 2.5 and Uniformity Cell Size >= 1.5 and Uniformity Cell Shape >= 2.5 and Marginal Adhesion < 5.5 and Bare Nuclei < 8.5 and Normal Nucleoli < 8.5 and Clump Thickness < 8.5 and Uniformity Cell Shape >= 3.5 and Marginal Adhesion >= 3.5 and Single Epithelial Cell Size < 3.5 Then Malignant |
| 23 | IF Bare Nuclei >= 2.5 and Uniformity Cell Size >= 1.5 and Uniformity Cell Shape >= 2.5 and Marginal Adhesion < 5.5 and Bare Nuclei < 8.5 and Normal Nucleoli < 8.5 and Clump Thickness < 8.5 and Uniformity Cell Shape >= 3.5 and Marginal Adhesion >= 3.5 and Single Epithelial Cell Size >= 3.5 and Uniformity Cell Shape < 5 Then Benign |
| 24 | IF Bare Nuclei >= 2.5 and Uniformity Cell Size >= 1.5 and Uniformity Cell Shape >= 2.5 and Marginal Adhesion < 5.5 and Bare Nuclei < 8.5 and Normal Nucleoli < 8.5 and Clump Thickness < 8.5 and Uniformity Cell Shape >= 3.5 and Marginal Adhesion >= 3.5 and Single Epithelial Cell Size >= 3.5 and Uniformity Cell Shape >= 5 and Normal Nucleoli < 3 Then Benign |
| 25 | IF Bare Nuclei >= 2.5 and Uniformity Cell Size >= 1.5 and Uniformity Cell Shape >= 2.5 and Marginal Adhesion < 5.5 and Bare Nuclei < 8.5 and Normal Nucleoli < 8.5 and Clump Thickness < 8.5 and Uniformity Cell Shape >= 3.5 and Single Epithelial Cell Size >= 3.5 and Uniformity Cell Shape >= 5 and Normal Nucleoli >= 3 Then Malignant |
| 26 | IF Bare Nuclei >= 2.5 and Uniformity Cell Size >= 1.5 and Uniformity Cell Shape >= 2.5 and Marginal Adhesion < 5.5 and Bare Nuclei < 8.5 and Normal Nucleoli < 8.5 and Clump Thickness >= 8.5 Then Malignant |
| 27 | IF Bare Nuclei >= 2.5 and Uniformity Cell Size >= 1.5 and Uniformity Cell Shape >= 2.5 and Marginal Adhesion < 5.5 and Bare Nuclei < 8.5 and Normal Nucleoli >= 8.5 Then Malignant |
| 28 | IF Bare Nuclei >= 2.5 and Uniformity Cell Size >= 1.5 and Uniformity Cell Shape >= 2.5 and Marginal Adhesion < 5.5 and Bare Nuclei >= 8.5 and Marginal Adhesion < 4.5 Then Malignant |
| 29 | IF Bare Nuclei >= 2.5 and Uniformity Cell Size >= 1.5 and Uniformity Cell Shape >= 2.5 and Marginal Adhesion < 5.5 and Bare Nuclei >= 8.5 and Marginal Adhesion >= 4.5 and Clump Thickness < 6.5 and Uniformity Cell Size < 5 Then Benign |
| 30 | IF Bare Nuclei >= 2.5 and Uniformity Cell Size >= 1.5 and Uniformity Cell Shape >= 2.5 and Marginal Adhesion < 5.5 and Bare Nuclei >= 8.5 and Marginal Adhesion >= 4.5 and Clump Thickness < 6.5 and Uniformity Cell Size >= 5 Then Malignant |
| 31 | IF Bare Nuclei >= 2.5 and Uniformity Cell Size >= 1.5 and Uniformity Cell Shape >= 2.5 and Marginal Adhesion < 5.5 and Bare Nuclei >= 8.5 and Marginal Adhesion >= 4.5 and Clump Thickness >= 6.5 Then Malignant |
| 32 | IF Bare Nuclei >= 2.5 and Uniformity Cell Size >= 1.5 and Uniformity Cell Shape >= 2.5 and Marginal Adhesion >= 5.5 Then Malignant |

**Figure C.3:** The rules of Random tree without PCA in WBC data set

| No. | Classification rules of Random Tree |
|---|---|
| 1 | IF PC1< 0.43 and PC1< -1.49 and PC1< -2.39 Then Malignant |
| 2 | IF PC1< 0.43 and PC1< -1.49 and PC1>= -2.39 and PC1< -2.37 Then Benign |
| 3 | IF PC1< 0.43 and PC1< -1.49 and PC1>= -2.39 and PC1>= -2.37 and PC1< -2.19 and PC1< -2.21 Then Malignant |
| 4 | IF PC1< 0.43 and PC1< -1.49 and PC1>= -2.39 and PC1>= -2.37 and PC1< -2.19 and PC1>= -2.21 Then Benign |
| 5 | IF PC1< 0.43 and PC1< -1.49 and PC1>= -2.39 and PC1>= -2.37 and PC1>= -2.19 Then Malignant |
| 6 | IF PC1< 0.43 and PC1>= -1.49 and PC1< -1.43 and PC1< -1.47 Then Benign |
| 7 | IF PC1< 0.43 and PC1>= -1.49vPC1< -1.43 and PC1>= -1.47 and PC1< -1.46 Then Malignant |
| 8 | IF PC1< 0.43 and PC1>= -1.49vPC1< -1.43 and PC1>= -1.47 and PC1>= -1.46 and PC1< -1.45 Then Benign |
| 9 | IF PC1< 0.43 and PC1>= -1.49 and PC1< -1.43 and PC1>= -1.47 and PC1>= -1.46 and PC1>= -1.45 and PC1< -1.44 Then Malignant |
| 10 | IF PC1< 0.43 and PC1>= -1.49 and PC1< -1.43 and PC1>= -1.47 and PC1>= -1.46 and PC1>= -1.45 and PC1>= -1.44 Then Benign |
| 11 | IF PC1< 0.43 and PC1>= -1.49 and PC1>= -1.43 and PC1< -1.29 Then Malignant |
| 12 | IF PC1< 0.43 and PC1>= -1.49 and PC1>= -1.43 and PC1>= -1.29 and PC1< -1.28 Then Benign |
| 13 | IF PC1< 0.43 and PC1>= -1.49 and PC1>= -1.43 and PC1>= -1.29 and PC1>= -1.28 and PC1< -1.06 Then Malignant |
| 14 | IF PC1< 0.43 and PC1>= -1.49 and PC1>= -1.43 and PC1>= -1.29 and PC1>= -1.28 and PC1>= -1.06 and PC1< -1.04 Then Benign |
| 15 | IF PC1< 0.43 and PC1>= -1.49 and PC1>= -1.43 and PC1>= -1.29 and PC1>= -1.28 and PC1>= -1.06 and PC1>= -1.04 and PC1< -0.87 Then Malignant |
| 16 | IF PC1< 0.43 and PC1>= -1.49 and PC1>= -1.43 and PC1>= -1.29 and PC1>= -1.28 and PC1>= -1.06 and PC1>= -1.04 and PC1>= -0.87 and PC1< -0.85 Then Benign |
| 17 | IF PC1< 0.43 and PC1>= -1.49 and PC1>= -1.43 and PC1>= -1.29 and PC1>= -1.28 and PC1>= -1.06 and PC1>= -1.04 and PC1>= -0.87 and PC1>= -0.85 and PC1< -0.61 Then Malignant |
| 18 | IF PC1< 0.43 and PC1>= -1.49 and PC1>= -1.43 and PC1>= -1.29 and PC1>= -1.28 and PC1>= -1.06 and PC1>= -1.04 and PC1>= -0.85 and PC1>= -0.61 and PC1< -0.01 and PC1< -0.29 and PC1< -0.59 Then Benign |
| 19 | IF PC1< 0.43 and PC1>= -1.49 and PC1>= -1.43 and PC1>= -1.29 and PC1>= -1.28 and PC1>= -1.06 and PC1>= -1.04 and PC1>= -0.87 and PC1>= -0.85 and PC1>= -0.61 and PC1< -0.01 and PC1< -0.29 and PC1>= -0.59 and PC1< -0.51 and PC1< -0.53 Then Malignant |
| 20 | IF PC1< 0.43 and PC1>= -1.49 and PC1>= -1.43 and PC1>= -1.29 and PC1>= -1.28 and PC1>= -1.06 and PC1>= -1.04 and PC1>= -0.87 and PC1>= -0.85 and PC1>= -0.61 and PC1< -0.01 and PC1< -0.29 and PC1>= -0.59 and PC1< -0.51 and PC1>= -0.53 Then Benign |
| 21 | IF PC1< 0.43 and PC1>= -1.49 and PC1>= -1.43 and PC1>= -1.29 and PC1>= -1.28 and PC1>= -1.06 and PC1>= -1.04 and PC1>= -0.87 and PC1>= -0.85 and PC1>= -0.61 and PC1< -0.01 and PC1< -0.29 and PC1>= -0.59 and PC1< -0.51 and PC1>= -0.51 Then Malignant |
| 22 | IF PC1< 0.43 and PC1>= -1.49 and PC1>= -1.43 and PC1>= -1.29 and PC1>= -1.28 and PC1>= -1.06 and PC1>= -1.04 and PC1>= -0.87 and PC1>= -0.85 and PC1>= -0.61 and PC1< -0.01 and PC1>= -0.29 Then Benign |
| 23 | IF PC1< 0.43 and PC1>= -1.49 and PC1>= -1.29 and PC1>= -1.28 and PC1>= -1.06 and PC1>= -1.04 and PC1>= -0.87 and PC1>= -0.61 and PC1< -0.01 and PC1>= -0.01 and PC1< 0.39 Then Malignant |
| 24 | IF PC1< 0.43 and PC1>= -1.49 and PC1>= -1.43 and PC1>= -1.29 and PC1>= -1.28 and PC1>= -1.06 and PC1>= -1.04 and PC1>= -0.87 and PC1>= -0.85 and PC1>= -0.61 and PC1< -0.01 and PC1>= -0.01 and PC1>= 0.39 and PC1< 0.42 Then Benign |
| 25 | IF PC1< 0.43 and PC1>= -1.49 and PC1>= -1.43 and PC1>= -1.29 and PC1>= -1.28 and PC1>= -1.06 and PC1>= -1.04 and PC1>= -0.85 and PC1>= -0.61 and PC1< -0.01 and PC1>= -0.01 and PC1>= 0.39PC1>= 0.42 Then Malignant |
| 26 | IF PC1>= 0.43 and PC1< 1.27 and PC1< 1.27 Then Benign |
| 27 | IF PC1>= 0.43 and PC1< 1.27 and PC1>= 1.27 Then Malignant |
| 28 | IF PC1>= 0.43 and PC1>= 1.27 Then Benign |

**Figure C.4:** The rules of Random tree with PCA in WBC data set

| No. | Classification rules of J48 Decision Tree |
|---|---|
| 1 | IF area worst <= 880.8 and concave points worst <= 0.1357 and area se <= 36.46 Then benign |
| 2 | IF area worst <= 880.8 and concave points worst <= 0.1357 and area se > 36.46 and radius mean <= 14.97 and texture se <= 1.978 Then benign |
| 3 | IF area worst <= 880.8 and concave points worst <= 0.1357 and area se > 36.46 and radius mean <= 14.97 and texture se > 1.978 and texture se <= 2.239 Then malignant |
| 4 | IF area worst <= 880.8 and concave points worst <= 0.1357 and area se > 36.46 and radius mean <= 14.97 and texture se > 1.978 and texture se > 2.239 Then benign |
| 5 | IF area worst <= 880.8 and concave points worst <= 0.1357 and area se > 36.46 and radius mean > 14.97 Then malignant |
| 6 | IF area worst <= 880.8 and concave points worst > 0.1357 and texture worst <= 27.37 and concave points worst <= 0.1789 and area se <= 21.91 Then benign |
| 7 | IF area worst <= 880.8 and concave points worst > 0.1357 and texture worst <= 27.37 and concave points worst <= 0.1789 and area se > 21.91 and perimeter se <= 2.615 Then malignant |
| 8 | IF area worst <= 880.8 and concave points worst > 0.1357 and texture worst <= 27.37 and concave points worst <= 0.1789 and area se > 21.91 and perimeter se > 2.615 Then benign |
| 9 | IF area worst <= 880.8 and concave points worst > 0.1357 and texture worst <= 27.37 and concave points worst > 0.1789 Then malignant |
| 10 | IF area worst <= 880.8 and concave points worst > 0.1357 and texture worst > 27.37 Then malignant |
| 11 | IF area worst > 880.8 and concavity mean <= 0.0716 and texture mean <= 19.54 Then benign |
| 12 | IF area worst > 880.8 and concavity mean <= 0.0716 and texture mean > 19.54 Then malignant |
| 13 | IF area worst > 880.8 and concavity mean <= 0.0716 and concavity mean > 0.0716 Then malignant |

**Figure C.5:** The rules of J48 decision tree without PCA in WDBC data set

| No. | Classification rules of REP Tree |
|---|---|
| 1 | IF perimeter worst < 114.45 and concave points worst < 0.11 Then Benign |
| 2 | IF perimeter worst < 114.45 and concave points worst >= 0.11 and texture worst < 25.73 Then Benign |
| 3 | IF perimeter worst < 114.45 and concave points worst >= 0.11 and texture worst >= 25.73 and concave points mean < 0.05 Then Benign |
| 4 | IF perimeter worst < 114.45 and concave points worst >= 0.11 and texture worst >= 25.73 and concave points mean >= 0.05 Then Malignant |
| 5 | IF perimeter worst >= 114.45 Then Malignant |

**Figure C.6:** The rules of REP tree without PCA in WDBC data set

| No. | Classification rules of Random Tree |
|---|---|
| 1 | IF concave points mean < 0.05 and concave points worst < 0.11 and area se < 40.22 and fractal dimension mean < 0.05 and radius mean < 15.29 Then Benign |
| 2 | IF concave points mean < 0.05 and concave points worst < 0.11 and area se < 40.22 and fractal dimension mean < 0.05 and radius mean >= 15.29 Then Malignant |
| 3 | IF concave points mean < 0.05 and concave points worst < 0.11 and area se < 40.22 and fractal dimension mean >= 0.05 and radius worst < 14.48 Then Benign |
| 4 | IF concave points mean < 0.05 and concave points worst < 0.11 and area se < 40.22 and fractal dimension mean >= 0.05 and radius worst >= 14.48 and area mean < 513.15 and texture se < 1.21 Then Benign |
| 5 | IF concave points mean < 0.05 and concave points worst < 0.11 and area se < 40.22 and fractal dimension mean >= 0.05 and radius worst >= 14.48 and area mean < 513.15 and texture se >= 1.21 Then Malignant |
| 6 | IF concave points mean < 0.05 and concave points worst < 0.11 and area se < 40.22 and fractal dimension mean >= 0.05 and radius worst >= 14.48 and area mean >= 513.15 Then Benign |
| 7 | IF concave points mean < 0.05 and concave points worst < 0.11 and area se >= 40.22 and fractal dimension se < 0 and radius se < 0.47 Then Malignant |
| 8 | IF concave points mean < 0.05 and concave points worst < 0.11 and area se >= 40.22 and fractal dimension se < 0 and radius se >= 0.47 and compactness mean < 0.06 Then Malignant |
| 9 | IF concave points mean < 0.05 and concave points worst < 0.11 and fractal dimension se < 0 and radius se >= 0.47 and compactness mean >= 0.06 Then Benign |
| 10 | IF concave points mean < 0.05 and area se >= 40.22 and fractal dimension se >= 0 Then Benign |
| 11 | IF concave points mean < 0.05 and fractal dimension worst < 0.08 and texture worst < 27.66 Then Benign |
| 12 | IF concave points mean < 0.05 and concave points worst >= 0.11 and fractal dimension worst < 0.08 and texture worst >= 27.66 and radius worst < 15.54 Then Benign |
| 13 | IF concave points mean < 0.05 and concave points worst >= 0.11 and fractal dimension worst < 0.08 and texture worst >= 27.66 and radius worst >= 15.54 Then Malignant |
| 14 | IF concave points mean < 0.05 and concave points worst >= 0.11 and fractal dimension worst >= 0.08 and radius worst < 15.54 Then Benign |
| 15 | IF concave points mean < 0.05 and concave points worst >= 0.11 and fractal dimension worst >= 0.08 and radius worst >= 15.54 and smoothness worst < 0.14 and symmetry se < 0.01 and radius mean < 15.1 Then Benign |
| 16 | IF concave points mean < 0.05 and concave points worst >= 0.11 and fractal dimension worst >= 0.08 and radius worst >= 15.54 and smoothness worst < 0.14 and symmetry se < 0.01 and radius mean >= 15.1 Then Malignant |
| 17 | IF concave points mean < 0.05 and concave points worst >= 0.11 and fractal dimension worst >= 0.08 and radius worst >= 15.54 and smoothness worst < 0.14 and symmetry se >= 0.01 Then Benign |
| 18 | IF concave points mean < 0.05 and concave points worst >= 0.11 and fractal dimension worst >= 0.08 and radius worst >= 15.54 and smoothness worst >= 0.14 Then Malignant |
| 19 | IF concave points mean >= 0.05 and perimeter worst < 114.45 and concave points se < 0.02 and fractal dimension worst < 0.09 and texture mean < 20.38 and concave points mean < 0.05 Then Malignant |
| 20 | IF concave points mean >= 0.05 and perimeter worst < 114.45 and concave points se < 0.02 and fractal dimension worst < 0.09 and texture mean < 20.38 and concave points mean >= 0.05 Then Benign |
| 21 | IF concave points mean >= 0.05 and perimeter worst < 114.45 and concave points se < 0.02 and fractal dimension worst < 0.09 and texture mean < 20.38 and texture mean >= 20.38 Then Malignant |
| 22 | IF concave points mean >= 0.05 and perimeter worst < 114.45 and concave points se < 0.02 and fractal dimension worst >= 0.09 and area worst < 656.3 and symmetry mean < 0.21 and concavity se < 0.03 Then Malignant |
| 23 | IF concave points mean >= 0.05 and perimeter worst < 114.45 and concave points se < 0.02 and fractal dimension worst >= 0.09 and area worst < 656.3 and symmetry mean < 0.21 and concavity se >= 0.03 Then Benign |
| 24 | IF concave points mean >= 0.05 and perimeter worst < 114.45 and concave points se < 0.02 and fractal dimension worst >= 0.09 and area worst < 656.3 and symmetry mean >= 0.21 Then Malignant |
| 25 | IF concave points mean >= 0.05 and perimeter worst < 114.45 and concave points se < 0.02 and fractal dimension worst >= 0.09 and area worst >= 656.3 Then Malignant |
| 26 | IF concave points mean >= 0.05 and perimeter worst < 114.45 and concave points se >= 0.02 Then Benign |
| 27 | IF concave points mean >= 0.05 and perimeter worst >= 114.45 Then Malignant |

**Figure C.7:** The rules of Random tree without PCA in WDBC data set

| No. | Classification rules of J48 Decision Tree |
|---|---|
| 1 | IF PC1 <= -0.446152 and PC5 <= 1.981667 and PC1 <= -2.105948 Then Malignant |
| 2 | IF PC1 <= -0.446152 and PC5 <= 1.981667 and PC1 > -2.105948 and PC2 <= -0.971261 and PC3 <= 0.80637 Then Benign |
| 3 | IF PC1 <= -0.446152 and PC5 <= 1.981667 and PC1 > -2.105948 and PC2 <= -0.971261 and PC3 > 0.80637 Then Malignant |
| 4 | IF PC1 <= -0.446152 and PC5 <= 1.981667 and PC1 > -2.105948 and PC2 > -0.971261 Then Malignant |
| 5 | IF PC1 <= -0.446152 and PC5 > 1.981667 Then Benign |
| 6 | IF PC1 > -0.446152 and PC1 <= 0.997191 and PC2 <= 1.281368 and PC3 <= 1.742709 Then Benign |
| 7 | IF PC1 > -0.446152 and PC1 <= 0.997191 and PC2 <= 1.281368 and PC3 > 1.742709 Then Malignant |
| 8 | IF PC1 > -0.446152 and PC1 <= 0.997191 and PC2 > 1.281368 Then Malignant |
| 9 | IF PC1 > -0.446152 and PC1 > 0.997191 Then Benign |

**Figure C.8:** The rules of J48 decision tree with PCA in WDBC data set

| No. | Classification rules of REP Tree |
|---|---|
| 1 | PC1 < -0.42 and PC2 < -1 and PC1 < -2.07 Then Malignant |
| 2 | PC1 < -0.42 and PC2 < -1 and PC1 >= -2.07 Then Benign |
| 3 | PC1 < -0.42 and PC2 >= -1 Then Malignant |
| 4 | PC1 >= -0.42 and PC1 < 0.96 and PC2 < 1.28 Then Benign |
| 5 | PC1 >= -0.42 and PC1 < 0.96 and PC2 >= 1.28 Then Malignant |
| 6 | PC1 >= -0.42 and PC1 >= 0.96 Then Benign |

**Figure C.9:** The rules of REP tree with PCA in WDBC data set

| No. | Classification rules of Random Tree |
|---|---|
| 1 | IF PC3< 1.5 and PC1< -0.43 and PC5< 2.89 and PC2< -1 and PC2< -4.14 and PC1< -4.25 Then Malignant |
| 2 | IF PC3< 1.5 and PC1< -0.43 and PC5< 2.89 and PC2< -1 and PC2< -4.14 and PC1>= -4.25 Then Malignant |
| 3 | IF PC3< 1.5 and PC1< -0.43 and PC5< 2.89 and PC2< -1 and PC2>= -4.14 and PC1< -1.52 Then Malignant |
| 4 | IF PC3< 1.5 and PC1< -0.43 and PC5< 2.89 and PC2< -1 and PC2>= -4.14 and PC1>= -1.52 Then Benign |
| 5 | IF PC3< 1.5 and PC1< -0.43 and PC5< 2.89 and PC2>= -1 Then Malignant |
| 6 | IF PC3< 1.5 and PC1< -0.43 and PC5>= 2.89 Then Benign |
| 7 | IF PC3< 1.5 and PC1>= -0.43 and PC4 < -1.03 and PC4 < -1.83 Then Benign |
| 8 | IF PC3< 1.5 and PC1>= -0.43 and PC4 < -1.03 and PC4 >= -1.83 and PC4 < -1.09 and PC2< 1.23 and PC1< 1.99 and PC2< -0.28 Then Benign |
| 9 | IF PC3< 1.5 and PC1>= -0.43 and PC4 < -1.03 and PC4 >= -1.83 and PC4 < -1.09 and PC2< 1.23 and PC1< 1.99 and PC2>= -0.28 and PC1< 1.95 Then Benign |
| 10 | IF PC3< 1.5 and PC1>= -0.43 and PC4 < -1.03 and PC4 >= -1.83 and PC4 < -1.09 and PC2< 1.23 and PC1< 1.99 and PC2>= -0.28 and PC1>= 1.95 Then Malignant |
| 11 | IF PC3< 1.5 and PC1>= -0.43 and PC4 < -1.03 and PC4 >= -1.83 and PC4 < -1.09 and PC2< 1.23 and PC1>= 1.99 Then Benign |
| 12 | IF PC3< 1.5 and PC1>= -0.43 and PC4 < -1.03 and PC4 >= -1.83 and PC4 < -1.09 and PC2>= 1.23 and PC5< 0.74 and PC1< 2.24 Then Malignant |
| 13 | IF PC3< 1.5 and PC1>= -0.43 and PC4 < -1.03 and PC4 >= -1.83 and PC4 < -1.09 and PC2>= 1.23 and PC5< 0.74 and PC1>= 2.24 Then Malignant |
| 14 | IF PC3< 1.5 and PC1>= -0.43 and PC4 < -1.03 and PC4 >= -1.83 and PC4 < -1.09 and PC2>= 1.23 and PC5>= 0.74 Then Malignant |
| 15 | IF PC3< 1.5 and PC1>= -0.43 and PC4 < -1.03 and PC4 >= -1.83 and PC4 >= -1.09 Then Benign |
| 16 | IF PC3< 1.5 and PC1>= -0.43 and PC4 >= -1.03 and PC2< 0.82 Then Benign |
| 17 | IF PC3< 1.5 and PC1>= -0.43 and PC4 >= -1.03 and PC2>= 0.82 and PC1< 0.74 and PC6< -0.1 Then Malignant |
| 18 | IF PC3< 1.5 and PC1>= -0.43 and PC4 >= -1.03 and PC2>= 0.82 and PC1< 0.74 and PC6>= -0.1 Then Malignant |
| 19 | IF PC3< 1.5 and PC1>= -0.43 and PC4 >= -1.03 and PC2>= 0.82 and PC1>= 0.74 and PC3< -1.86 Then Malignant |
| 20 | IF PC3< 1.5 and PC1>= -0.43 and PC4 >= -1.03 and PC2>= 0.82 and PC1>= 0.74 and PC3>= -1.86 Then Benign |
| 21 | IF PC3>= 1.5 and PC2< 1.55 and PC1< 0.75 and PC1< -1.21 Then Malignant |
| 22 | IF PC3>= 1.5 and PC2< 1.55 and PC1< 0.75 and PC1>= -1.21 and PC1< 0.07 and PC1< -0.49 and PC2< -1.75 Then Benign |
| 23 | IF PC3>= 1.5 and PC2< 1.55 and PC1< 0.75 and PC1>= -1.21 and PC1< 0.07 and PC1< -0.49 and PC2>= -1.75 Then Malignant |
| 24 | IF PC3>= 1.5 and PC2< 1.55 and PC1< 0.75 and PC1>= -1.21 and PC1< 0.07 and PC1>= -0.49 Then Malignant |
| 25 | IF PC3>= 1.5 and PC2< 1.55 and PC1< 0.75 and PC1>= -1.21 and PC1>= 0.07 Then Benign |
| 26 | IF PC3>= 1.5 and PC2< 1.55 and PC1>= 0.75 Then Benign |
| 27 | IF PC3>= 1.5 and PC2< 1.55 and PC2>= 1.55 Then Malignant |

**Figure C.10:** The rules of Random tree with PCA in WDBC data set

| No. | Classification rules of J48 Decision Tree |
|---|---|
| 1 | IF time <= 49 and Texture se <= 1.667 and Lump node status <= 0 and Smoothness worst <= 0.1419 Then No Recur |
| 2 | IF time <= 49 and Texture se <= 1.667 and Lump node status <= 0 and Smoothness worst > 0.1419 and Area mean <= 610.7 Then No Recur |
| 3 | IF time <= 49 and Texture se <= 1.667 and Lump node status <= 0 and Smoothness worst > 0.1419 and Area mean > 610.7 Then Recur |
| 4 | IF time <= 49 and Texture se <= 1.667 and Lump node status > 0 and Area worst <= 1260 and Texture se <= 1.187 and Radius mean <= 13.98 Then Recur |
| 5 | IF time <= 49 and Texture se <= 1.667 and Lump node status > 0 and Area worst <= 1260 and Texture se <= 1.187 and Radius mean > 13.98 and Texture worst <= 35.59 Then No Recur |
| 6 | IF time <= 49 and Texture se <= 1.667 and Lump node status > 0 and Area worst <= 1260 and Texture se <= 1.187 and Radius mean > 13.98 and Texture worst > 35.59 Then Recur |
| 7 | IF time <= 49 and Texture se <= 1.667 and Lump node status > 0 and Area worst <= 1260 and Texture se > 1.187 Then No Recur |
| 8 | IF time <= 49 and Texture se <= 1.667 and Lump node status > 0 and Area worst > 1260 and Tumor size <= 1.3 Then No Recur |
| 9 | IF time <= 49 and Texture se <= 1.667 and Lump node status > 0 and Area worst > 1260 and Tumor size > 1.3 and Concave point se <= 0.02149 and Fractal dimension worst <= 0.08574 Then Recur |
| 10 | IF time <= 49 and Texture se <= 1.667 and Lump node status > 0 and Area worst > 1260 and Tumor size > 1.3 and Concave point se <= 0.02149 and Fractal dimension worst > 0.08574 and Concavity mean <= 0.1697 Then No Recur |
| 11 | IF time <= 49 and Texture se <= 1.667 and Lump node status > 0 and Area worst > 1260 and Tumor size > 1.3 and Concave point se <= 0.02149 and Fractal dimension worst > 0.08574 and Concavity mean > 0.1697 Then Recur |
| 12 | IF time <= 49 and Texture se <= 1.667 and Lump node status > 0 and Area worst > 1260 and Tumor size > 1.3 and Concave point se > 0.02149 Then No Recur |
| 13 | IF time <= 49 and Texture se > 1.667 and Symmetry se <= 0.02899 Then No Recur |
| 14 | IF time <= 49 and Texture se > 1.667 and Symmetry se > 0.02899 and Lump node status <= 0 Then No Recur |
| 15 | IF time <= 49 and Texture se > 1.667 and Symmetry se > 0.02899 and Lump node status > 0 Then Recur |
| 16 | IF time > 49 Then No Recur |

**Figure C.11:** The rules of J48 Decision tree without PCA in WPBC data set

| No. | Classification rules of REP Tree |
|---|---|
| 1 | IF time < 50 and Radius worst < 17.43 Then No Recur |
| 2 | IF time < 50 and Radius worst >= 17.43 and Concavity se < 0.02 Then Recur |
| 3 | IF time < 50 and Radius worst >= 17.43 and Concavity se >= 0.02 and Texture se < 1.67 and Tumor size < 3.85 and Radius worst < 20.86 Then No Recur |
| 4 | IF time < 50 and Radius worst >= 17.43 and Concavity se >= 0.02 and Texture se < 1.67 and Tumor size < 3.85 and Radius worst >= 20.86 and time < 16.5 Then Recur |
| 5 | IF time < 50 and Radius worst >= 17.43 and Concavity se >= 0.02 and Texture se < 1.67 and Tumor size < 3.85 and Radius worst >= 20.86 and time >= 16.5 Then No Recur |
| 6 | IF time < 50 and Radius worst >= 17.43 and Concavity se >= 0.02 and Texture se < 1.67 and Tumor size >= 3.85 Then Recur |
| 7 | IF time < 50 and Radius worst >= 17.43 and Concavity se >= 0.02 and Texture se >= 1.67 Then No Recur |
| 8 | IF time >= 50 Then No Recur |

**Figure C.12:** The rules of REP tree without PCA in WPBC data set

| No. | Classification rules of Random Tree |
|---|---|
| 1 | IF time < 50 and Concave point worst < 0.12 Then No Recur |
| 2 | IF time < 50 and Concave point worst >= 0.12 and Compactness se < 0.02 and Symmetry se < 0.01 Then Recur |
| 3 | IF time < 50 and Concave point worst >= 0.12 and Compactness se < 0.02 and Symmetry se >= 0.01 and Compactness se < 0.02 and Fractal dimension worst < 0.07 Then Recur |
| 4 | IF time < 50 and Concave point worst >= 0.12 and Compactness se < 0.02 and Symmetry se >= 0.01 and Compactness se < 0.02 and Fractal dimension worst >= 0.07 Then No Recur |
| 5 | IF time < 50 and Concave point worst >= 0.12 and Compactness se < 0.02 and Symmetry se >= 0.01 and Compactness se >= 0.02 Then Recur |
| 6 | IF time < 50 and Concave point worst >= 0.12 and Compactness se >= 0.02 and Compactness mean < 0.12 and Radius mean < 15.26 and Perimeter mean < 95.81 Then No Recur |
| 7 | IF time < 50 and Concave point worst >= 0.12 and Compactness se >= 0.02 and Compactness mean < 0.12 and Radius mean < 15.26 and Perimeter mean >= 95.81 Then Recur |
| 8 | IF time < 50 and Concave point worst >= 0.12 and Compactness se >= 0.02 and Compactness mean < 0.12 and Radius mean >= 15.26 Then No Recur |
| 9 | IF time < 50 and Concave point worst >= 0.12 and Compactness se >= 0.02 and Compactness mean >= 0.12 and Fractal dimension mean < 0.06 and Area worst < 2077 and Concavity se < 0.06 and Concavity se < 0.04 Then No Recur |
| 10 | IF time < 50 and Concave point worst >= 0.12 and Compactness se >= 0.02 and Compactness mean >= 0.12 and Fractal dimension mean < 0.06 and Area worst < 2077 and Concavity se < 0.06 and Concavity se >= 0.04 and Lump node status < 0.5 and Texture mean < 19.52 Then Recur |
| 11 | IF time < 50 and Concave point worst >= 0.12 and Compactness se >= 0.02 and Compactness mean >= 0.12 and Fractal dimension mean < 0.06 and Area worst < 2077 and Concavity se < 0.06 and Concavity se >= 0.04 and Lump node status < 0.5 and Texture mean >= 19.52 Then No Recur |
| 12 | IF time < 50 and Concave point worst >= 0.12 and Compactness se >= 0.02 and Compactness mean >= 0.12 and Fractal dimension mean < 0.06 and Area worst < 2077 and Concavity se < 0.06 and Concavity se >= 0.04 and Lump node status >= 0.5 Then Recur |
| 13 | IF time < 50 and Concave point worst >= 0.12 and Compactness se >= 0.02 and Compactness mean >= 0.12 and Fractal dimension mean < 0.06 and Area worst < 2077 and Concavity se >= 0.06 Then No Recur |
| 14 | IF time < 50 and Concave point worst >= 0.12 and Compactness se >= 0.02 and Compactness mean >= 0.12 and Fractal dimension mean < 0.06 and Area worst >= 2077 Then Recur |
| 15 | IF time < 50 and Concave point worst >= 0.12 and Compactness se >= 0.02 and Compactness mean >= 0.12 and Fractal dimension mean >= 0.06 and Concavity mean < 0.21 and Compactness se < 0.06 and Smoothness mean < 0.11 Then No Recur |
| 16 | IF time < 50 and Concave point worst >= 0.12 and Compactness se >= 0.02 and Compactness mean >= 0.12 and Fractal dimension mean >= 0.06 and Concavity mean < 0.21 and Compactness se < 0.06 and Smoothness mean >= 0.11 and Tumor size < 2.75 and time < 46.5 and Perimeter mean < 87.13 and Perimeter mean < 78 Then No Recur |
| 17 | IF time < 50 and Concave point worst >= 0.12 and Compactness se >= 0.02 and Compactness mean >= 0.12 and Fractal dimension mean >= 0.06 and Concavity mean < 0.21 and Compactness se < 0.06 and Smoothness mean >= 0.11 and Tumor size < 2.75 and time < 46.5 and Perimeter mean < 87.13 and Perimeter mean >= 78 Then Recur |
| 18 | IF time < 50 and Concave point worst >= 0.12 and Compactness se >= 0.02 and Compactness mean >= 0.12 and Fractal dimension mean >= 0.06 and Concavity mean < 0.21 and Compactness se < 0.06 and Smoothness mean >= 0.11 and Tumor size < 2.75 and time < 46.5 and Perimeter mean >= 87.13 Then No Recur |
| 19 | IF time < 50 and Concave point worst >= 0.12 and Compactness se >= 0.02 and Compactness mean >= 0.12 and Fractal dimension mean >= 0.06 and Concavity mean < 0.21 and Compactness se < 0.06 and Smoothness mean >= 0.11 and Tumor size < 2.75 and time >= 46.5 Then Recur |
| 20 | IF time < 50 and Concave point worst >= 0.12 and Compactness se >= 0.02 and Compactness mean >= 0.12 and Fractal dimension mean >= 0.06 and Concavity mean < 0.21 and Compactness se < 0.06 and Smoothness mean >= 0.11 and Tumor size >= 2.75 Then Recur |
| 21 | IF time < 50 and Concave point worst >= 0.12 and Compactness se >= 0.02 and Compactness mean >= 0.12 and Fractal dimension mean >= 0.06 and Concavity mean < 0.21 and Compactness se >= 0.06 Then Recur |
| 22 | IF time < 50 and Concave point worst >= 0.12 and Compactness se >= 0.02 and Compactness mean >= 0.12 and Fractal dimension mean >= 0.06 and Concavity mean >= 0.21 Then No Recur |
| 23 | IF time >= 50 and Fractal dimension mean < 0.05 Then Recur |
| 24 | IF time >= 50 and Fractal dimension mean >= 0.05 and Fractal dimension mean < 0.06 Then No Recur |
| 25 | IF time >= 50 and Fractal dimension mean >= 0.05 and Fractal dimension mean >= 0.06 and Compactness mean < 0.17 and Compactness mean < 0.16 and Compactness se < 0.04 and Concavity mean < 0.14 Then No Recur |
| 26 | IF time >= 50 and Fractal dimension mean >= 0.05 and Fractal dimension mean >= 0.06 and Compactness mean < 0.17 and Compactness mean < 0.16 and Compactness se < 0.04 and Concavity mean >= 0.14 and Fractal dimension mean < 0.06 Then Recur |
| 27 | IF time >= 50 and Fractal dimension mean >= 0.05 and Fractal dimension mean >= 0.06 and Compactness mean < 0.17 and Compactness mean < 0.16 and Compactness se < 0.04 and Concavity mean >= 0.14 and Fractal dimension mean >= 0.06 and Area worst < 879.1 Then Recur |
| 28 | IF time >= 50 and Fractal dimension mean >= 0.05 and Fractal dimension mean >= 0.06 and Compactness mean < 0.17 and Compactness mean < 0.16 and Compactness se < 0.04 and Concavity mean >= 0.14 and Fractal dimension mean >= 0.06 and Area worst >= 879.1 Then No Recur |
| 29 | IF time >= 50 and Fractal dimension mean >= 0.05 and Fractal dimension mean >= 0.06 and Compactness mean < 0.17 and Compactness mean < 0.16 and Compactness se >= 0.04 Then Recur |
| 30 | IF time >= 50 and Fractal dimension mean >= 0.05 and Fractal dimension mean >= 0.06 and Compactness mean < 0.17 and Compactness mean >= 0.16 Then Recur |
| 31 | IF time >= 50 and Fractal dimension mean >= 0.05 and Fractal dimension mean >= 0.06 and Compactness mean >= 0.17 Then No Recur |

**Figure C.13:** The rules of Random tree without PCA in WPBC data set

| No. | Classification rules of J48 Decision Tree |
|---|---|
| 1 | IF PC9 <= -0.776767 Then No Recur |
| 2 | IF PC9 > -0.776767 and PC2 <= -4.136541 Then No Recur |
| 3 | IF PC9 > -0.776767 and PC2 > -4.136541 and PC3 <= 0.084338 and PC5 <= -0.5713 Then No Recur |
| 4 | IF PC9 > -0.776767 and PC2 > -4.136541 and PC3 <= 0.084338 and PC5 > -0.5713 and PC4 <= 2.134624 and PC6 <= 1.734945 and PC1 <= 1.631704 Then No Recur |
| 5 | IF PC9 > -0.776767 and PC2 > -4.136541 and PC3 <= 0.084338 and PC5 > -0.5713 and PC4 <= 2.134624 and PC6 <= 1.734945 and PC1 > 1.631704 and PC2 <= 1.112113 Then No Recur |
| 6 | IF PC9 > -0.776767 and PC2 > -4.136541 and PC3 <= 0.084338 and PC5 > -0.5713 and PC4 <= 2.134624 and PC6 <= 1.734945 and PC1 > 1.631704 and PC2 > 1.112113 Then Recur |
| 7 | IF PC9 > -0.776767 and PC2 > -4.136541 and PC3 <= 0.084338 and PC5 > -0.5713 and PC4 <= 2.134624 and PC6 > 1.734945 Then Recur |
| 8 | IF PC9 > -0.776767 and PC2 > -4.136541 and PC3 <= 0.084338 and PC5 > -0.5713 and PC4 > 2.134624 Then Recur |
| 9 | IF PC9 > -0.776767 and PC2 > -4.136541 and PC3 > 0.084338 and PC9 <= 1.198678 and PC1 <= -3.01202 Then No Recur |
| 10 | IF PC9 > -0.776767 and PC2 > -4.136541 and PC3 > 0.084338 and PC9 <= 1.198678 and PC1 > -3.01202 and PC1 <= -2.242518 Then Recur |
| 11 | IF PC9 > -0.776767 and PC2 > -4.136541 and PC3 > 0.084338 and PC9 <= 1.198678 and PC1 > -3.01202 and PC1 > -2.242518 and PC2 <= 4.017542 and PC6 <= 1.583503 and PC8 <= -0.396361 and PC1 <= -0.761937 Then No Recur |
| 12 | IF PC9 > -0.776767 and PC2 > -4.136541 and PC3 > 0.084338 and PC9 <= 1.198678 and PC1 > -3.01202 and PC1 > -2.242518 and PC2 <= 4.017542 and PC6 <= 1.583503 and PC8 <= -0.396361 and PC1 > -0.761937 and PC6 <= 0.418143 Then Recur |
| 13 | IF PC9 > -0.776767 and PC2 > -4.136541 and PC3 > 0.084338 and PC9 <= 1.198678 and PC1 > -3.01202 and PC1 > -2.242518 and PC2 <= 4.017542 and PC6 <= 1.583503 and PC8 <= -0.396361 and PC1 > -0.761937 and PC6 > 0.418143 and PC2 <= 1.157287 Then No Recur |
| 14 | IF PC9 > -0.776767 and PC2 > -4.136541 and PC3 > 0.084338 and PC9 <= 1.198678 and PC1 > -3.01202 and PC1 > -2.242518 and PC2 <= 4.017542 and PC6 <= 1.583503 and PC8 <= -0.396361 and PC1 > -0.761937 and PC6 > 0.418143 and PC2 > 1.157287 Then Recur |
| 15 | IF PC9 > -0.776767 and PC2 > -4.136541 and PC3 > 0.084338 and PC9 <= 1.198678 and PC1 > -3.01202 and PC1 > -2.242518 and PC2 <= 4.017542 and PC6 <= 1.583503 and PC8 > -0.396361 and PC2 <= 0.724495 and PC5 <= -0.559847 Then No Recur |
| 16 | IF PC9 > -0.776767 and PC2 > -4.136541 and PC3 > 0.084338 and PC9 <= 1.198678 and PC1 > -3.01202 and PC1 > -2.242518 and PC2 <= 4.017542 and PC6 <= 1.583503 and PC8 > -0.396361 and PC2 <= 0.724495 and PC5 > -0.559847 Then Recur |
| 17 | IF PC9 > -0.776767 and PC2 > -4.136541 and PC3 > 0.084338 and PC9 <= 1.198678 and PC1 > -3.01202 and PC1 > -2.242518 and PC2 <= 4.017542 and PC6 <= 1.583503 and PC8 > -0.396361 and PC2 > 0.724495 Then No Recur |
| 18 | IF PC9 > -0.776767 and PC2 > -4.136541 and PC3 > 0.084338 and PC9 <= 1.198678 and PC1 > -3.01202 and PC1 > -2.242518 and PC2 <= 4.017542 and PC6 > 1.583503 Then Recur |
| 19 | IF PC9 > -0.776767 and PC2 > -4.136541 and PC3 > 0.084338 anPC9 <= 1.198678 and PC1 > -3.01202 and PC1 > -2.242518 and PC2 > 4.017542 Then Recur |
| 20 | IF PC9 > -0.776767 and PC2 > -4.136541 and PC3 > 0.084338 and PC9 > 1.198678 Then Recur |

**Figure C.14:** The rules of J48 decision tree with PCA in WPBC data set

| No. | Classification rules of REP Tree |
|---|---|
| 1 | IF PC5 < -0.34 Then No Recur |
| 2 | IF PC5 >= -0.34 and PC1 < -0.05 Then No Recur |
| 3 | IF PC5 >= -0.34 and PC1 >= -0.05 and PC9 < 0.12 and PC2 < 2.86 and PC3 < 0.46 Then No Recur |
| 4 | IF PC5 >= -0.34 and PC1 >= -0.05 and PC9 < 0.12 and PC2 < 2.86 and PC3 >= 0.46 Then Recur |
| 5 | IF PC5 >= -0.34 and PC1 >= -0.05 and PC9 < 0.12 and PC2 >= 2.86 Then Recur |
| 6 | IF PC5 >= -0.34 and PC1 >= -0.05 and PC9 >= 0.12 and PC6 < 0.17 and PC4 < -0.14 Then Recur |
| 7 | IF PC5 >= -0.34 and PC1 >= -0.05 and PC9 >= 0.12 and PC6 < 0.17 and PC4 >= -0.14 Then No Recur |
| 8 | IF PC5 >= -0.34 and PC1 >= -0.05 and PC9 >= 0.12 and PC6 >= 0.17 Then Recur |

**Figure C.15:** The rules of REP tree with PCA in WPBC data set

| No. | Classification rules of Random Tree |
|---|---|
| 1 | IF PC2 < -4.12 Then No Recur |
| 2 | IF PC2 >= -4.12 and PC2 < -3.98 Then Recur |
| 3 | IF PC2 >= -4.12 and PC2 >= -3.98 and PC9 < -0.78 Then No Recur |
| 4 | IF PC2 >= -4.12 and PC2 >= -3.98 and PC9 >= -0.78 and PC3 < 2.89 and PC5 < -0.38 and PC6 < 1.8 and PC6 < -0.6 and PC3 < -2.12 Then Recur |
| 5 | IF PC2 >= -4.12 and PC2 >= -3.98 and PC9 >= -0.78 and PC3 < 2.89 and PC5 < -0.38 and PC6 < 1.8 and PC6 < -0.6 and PC3 >= -2.12 and PC8 < 1.92 and PC3 < 1.23 and PC1 < -1.82 and PC2 < -3.64 Then No Recur |
| 6 | IF PC2 >= -4.12 and PC2 >= -3.98 and PC9 >= -0.78 and PC3 < 2.89 and PC5 < -0.38 and PC6 < 1.8 and PC6 < -0.6 and PC3 >= -2.12 and PC8 < 1.92 and PC3 < 1.23 and PC1 < -1.82 and PC2 >= -3.64 Then Recur |
| 7 | IF PC2 >= -4.12 and PC2 >= -3.98 and PC9 >= -0.78 and PC3 < 2.89 and PC5 < -0.38 and PC6 < 1.8 and PC6 < -0.6 and PC3 >= -2.12 and PC8 < 1.92 and PC3 < 1.23 and PC1 >= -1.82 Then No Recur |
| 8 | IF PC2 >= -4.12 and PC2 >= -3.98 and PC9 >= -0.78 and PC3 < 2.89 and PC5 < -0.38 and PC6 < 1.8 and PC6 < -0.6 and PC3 >= -2.12 and PC8 < 1.92 and PC3 >= 1.23 Then No Recur |
| 9 | IF PC2 >= -4.12 and PC2 >= -3.98 and PC9 >= -0.78 and PC3 < 2.89 and PC5 < -0.38 and PC6 < 1.8 and PC6 < -0.6 and PC3 >= -2.12 and PC8 >= 1.92 Then Recur |
| 10 | IF PC2 >= -4.12 and PC2 >= -3.98 and PC9 >= -0.78 and PC3 < 2.89 and PC5 < -0.38 and PC6 < 1.8 and PC6 >= -0.6 and PC9 < 0.37 Then No Recur |
| 11 | IF PC2 >= -4.12 and PC2 >= -3.98 and PC9 >= -0.78 and PC3 < 2.89 and PC5 < -0.38 and PC6 < 1.8 and PC6 >= -0.6 and PC9 >= 0.37 and PC2 < 2.22 Then No Recur |
| 12 | IF PC2 >= -4.12 and PC2 >= -3.98 and PC9 >= -0.78 and PC3 < 2.89 and PC5 < -0.38 and PC6 < 1.8 and PC6 >= -0.6 and PC9 >= 0.37 and PC2 >= 2.22 Then Recur |
| 13 | IF PC2 >= -4.12 and PC2 >= -3.98 and PC9 >= -0.78 and PC3 < 2.89 and PC5 < -0.38 and PC6 >= 1.8 and PC2 < 1.42 Then Recur |
| 14 | IF PC2 >= -4.12 and PC2 >= -3.98 and PC9 >= -0.78 and PC3 < 2.89 and PC5 < -0.38 and PC6 >= 1.8 and PC2 >= 1.42 Then Recur |
| 15 | IF PC2 >= -4.12 and PC2 >= -3.98 and PC9 >= -0.78 and PC3 < 2.89 and PC5 >= -0.38 and PC4 < -2.67 Then No Recur |
| 16 | IF PC2 >= -4.12 and PC2 >= -3.98 and PC9 >= -0.78 and PC3 < 2.89 and PC5 >= -0.38 and PC4 >= -2.67 and PC3 < 2.57 and PC7 < 0.76 and PC1 < 0.41 and PC1 < -0.06 and PC9 < 1.24 and PC1 < -4.61 Then Recur |
| 17 | IF PC2 >= -4.12 and PC2 >= -3.98 and PC9 >= -0.78 and PC3 < 2.89 and PC5 >= -0.38 and PC4 >= -2.67 and PC3 < 2.57 and PC7 < 0.76 and PC1 < 0.41 and PC1 < -0.06 and PC9 < 1.24 and PC1 >= -4.61 and PC7 < -0.38 and PC9 < -0.07 Then Recur |
| 18 | IF PC2 >= -4.12 and PC2 >= -3.98 and PC9 >= -0.78 and PC3 < 2.89 and PC5 >= -0.38 and PC4 >= -2.67 and PC3 < 2.57 and PC7 < 0.76 and PC1 < 0.41 and PC1 < -0.06 and PC9 < 1.24 and PC1 >= -4.61 and PC7 < -0.38 and PC9 >= -0.07 Then No Recur |
| 19 | IF PC2 >= -4.12 and PC2 >= -3.98 and PC9 >= -0.78 and PC3 < 2.89 and PC5 >= -0.38 and PC4 >= -2.67 and PC3 < 2.57 and PC7 < 0.76 and PC1 < 0.41 and PC1 < -0.06 and PC9 < 1.24 and PC1 >= -4.61 and PC7 >= -0.38 and PC4 < 0.95 Then No Recur |
| 20 | IF PC2 >= -4.12 and PC2 >= -3.98 and PC9 >= -0.78 and PC3 < 2.89 and PC5 >= -0.38 and PC4 >= -2.67 and PC3 < 2.57 and PC7 < 0.76 and PC1 < 0.41 and PC1 < -0.06 and PC9 < 1.24 and PC1 >= -4.61 and PC7 >= -0.38 and PC4 >= 0.95 Then Recur |
| 21 | IF PC2 >= -4.12 and PC2 >= -3.98 and PC9 >= -0.78 and PC3 < 2.89 and PC5 >= -0.38 and PC4 >= -2.67 and PC3 < 2.57 and PC7 < 0.76 and PC1 < 0.41 and PC1 < -0.06 and PC9 >= 1.24 Then Recur |
| 22 | IF PC2 >= -4.12 and PC2 >= -3.98 and PC9 >= -0.78 and PC3 < 2.89 and PC5 >= -0.38 and PC4 >= -2.67 and PC3 < 2.57 and PC7 < 0.76 and PC1 < 0.41 and PC1 >= -0.06 Then No Recur |
| 23 | IF PC2 >= -4.12 and PC2 >= -3.98 and PC9 >= -0.78 and PC3 < 2.89 and PC5 >= -0.38 and PC4 >= -2.67 and PC3 < 2.57 and PC7 < 0.76 and PC1 >= 0.41 and PC4 < 0.18 Then Recur |
| 24 | IF PC2 >= -4.12 and PC2 >= -3.98 and PC9 >= -0.78 and PC3 < 2.89 and PC5 >= -0.38 and PC4 >= -2.67 and PC3 < 2.57 and PC7 < 0.76 and PC1 >= 0.41 and PC4 >= 0.18 and PC6 < 0.36 Then Recur |
| 25 | IF PC2 >= -4.12 and PC2 >= -3.98 and PC9 >= -0.78 and PC3 < 2.89 and PC5 >= -0.38 and PC4 >= -2.67 and PC3 < 2.57 and PC7 < 0.76 and PC1 >= 0.41 and PC4 >= 0.18 and PC6 >= 0.36 and PC5 < 1.58 and PC4 < 0.26 Then No Recur |
| 26 | IF PC2 >= -4.12 and PC2 >= -3.98 and PC9 >= -0.78 and PC3 < 2.89 and PC5 >= -0.38 and PC4 >= -2.67 and PC3 < 2.57 and PC7 < 0.76 and PC1 >= 0.41 and PC4 >= 0.18 and PC6 >= 0.36 and PC5 < 1.58 and PC4 >= 0.26 Then No Recur |
| 27 | IF PC2 >= -4.12 and PC2 >= -3.98 and PC9 >= -0.78 and PC3 < 2.89 and PC5 >= -0.38 and PC4 >= -2.67 and PC3 < 2.57 and PC7 < 0.76 and PC1 >= 0.41 and PC4 >= 0.18 and PC6 >= 0.36 and PC5 >= 1.58 Then Recur |
| 28 | IF PC2 >= -4.12 and PC2 >= -3.98 and PC9 >= -0.78 and PC3 < 2.89 and PC5 >= -0.38 and PC4 >= -2.67 and PC3 < 2.57 and PC7 >= 0.76 and PC5 < 0.1 Then No Recur |
| 29 | IF PC2 >= -4.12 and PC2 >= -3.98 and PC9 >= -0.78 and PC3 < 2.89 and PC5 >= -0.38 and PC4 >= -2.67 and PC3 < 2.57 and PC7 >= 0.76 and PC5 >= 0.1 and PC1 < 1.15 and PC2 < -0.74 Then No Recur |
| 30 | IF PC2 >= -4.12 and PC2 >= -3.98 and PC9 >= -0.78 and PC3 < 2.89 and PC5 >= -0.38 and PC4 >= -2.67 and PC3 < 2.57 and PC7 >= 0.76 and PC5 >= 0.1 and PC1 < 1.15 and PC2 >= -0.74 and PC7 < 1.1 Then Recur |
| 31 | IF PC2 >= -4.12 and PC2 >= -3.98 and PC9 >= -0.78 and PC3 < 2.89 and PC5 >= -0.38 and PC4 >= -2.67 and PC3 < 2.57 and PC7 >= 0.76 and PC5 >= 0.1 and PC1 < 1.15 and PC2 >= -0.74 and PC7 >= 1.1 Then No Recur |
| 32 | IF PC2 >= -4.12 and PC2 >= -3.98 and PC9 >= -0.78 and PC3 < 2.89 and PC5 >= -0.38 and PC4 >= -2.67 and PC3 < 2.57 and PC7 >= 0.76 and PC5 >= 0.1 and PC1 >= 1.15 Then Recur |
| 33 | IF PC2 >= -4.12 and PC2 >= -3.98 and PC9 >= -0.78 and PC3 < 2.89 and PC5 >= -0.38 and PC4 >= -2.67 and PC3 >= 2.57 Then Recur |
| 34 | IF PC2 >= -4.12 and PC2 >= -3.98 and PC9 >= -0.78 and PC3 >= 2.89 Then No Recur |

**Figure C.16:** The rules of Random tree with PCA in WPBC data set

# APPENDIX D Publications

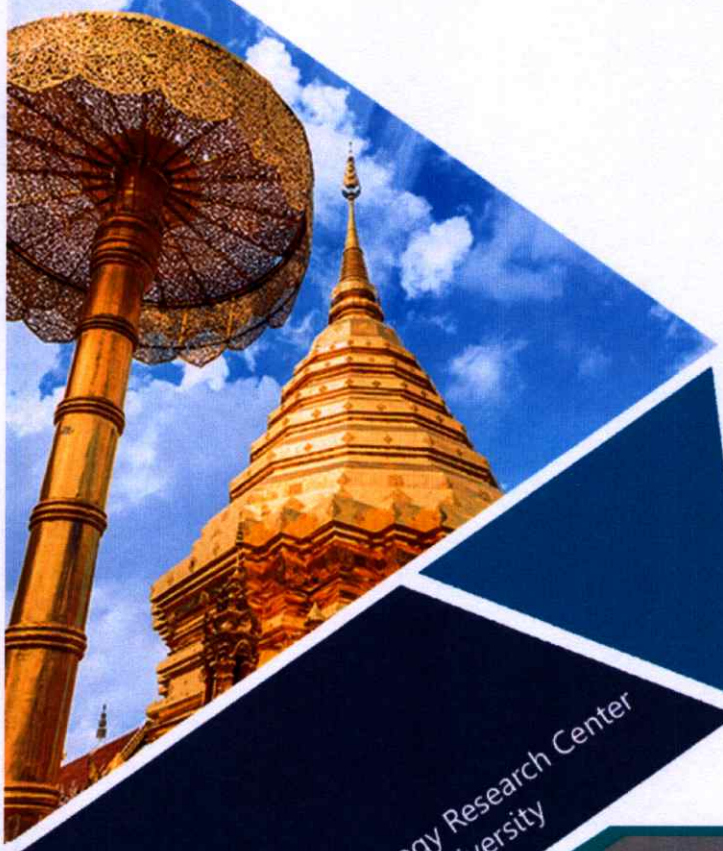
The 2018-10th International Conference on Knowledge and Smart Technology

# "Cybernetics in the Next Decades"

KST Research Center
Knowledge & Smart Technologies
since 2008

BURAPHA UNIVERSITY

KST 2018
Jan 31-Feb 3



@Kantary Hills Hotel
Chiangmai, Thailand

Organized by
Knowledge and Smart Technology Research Center
Faculty of Informatics, Burapha University

ISBN 978-1-5386-4014-2

Sponsored by IEEE IEEE Xplore Digital Library IEEE THAILAND SECTION

Patronage by ECTI depa AIAT BIA

# Building Minimal Classification Rules for Breast Cancer Diagnosis

Phonethep Douangnoulack
International College
King Mongkut's Institute of Technology Ladkrabang
Thailand
Email: phonethepdouangnoulack@hotmail.com

Veera Boonjing
International College
King Mongkut's Institute of Technology Ladkrabang
Thailand
Email: kbveera@kmitl.ac.th

***Abstract*— A rule based classifier is widely applied in breast cancer diagnosis. The classifier with a good performance of disease classification have been developed and highly required over the past decades. Since classification rules are derived from previous diagnosis with a large amount of features, it challenges to build a minimal number of rules with high performance while retaining all diagnosis information. The Principal Component Analysis (PCA) is known as a lossless data reduction technique with good classification performance. Therefore, this paper aims at finding the best performance classifier giving minimal classification rules by employing PCA. Based on experiment result on Wisconsin Breast Cancer data set, the J48 decision tree classifier is found to be the best among the three classifiers: J48 decision tree, Reduced Error Pruning Tree, and Random Tree.**

***Keywords—Rule Based Classifier; Decision Tree; PCA; Breast Cancer Diagnosis.***

## I. INTRODUCTION

A rule based classifier plays an important role in modern breast cancer diagnosis. The good classifier equips with high accurate classification rules obtaining from historical diagnosis. Since each diagnosis consists of a large amount of data features, it challenges to build minimal high accurate classification rules from such historical data. Basically, feature reduction techniques could help reduce a number of classification rules. But the trade-off is classification performance. However, if we could find a technique of feature reduction giving high classification accuracy, it would help obtain minimal high accurate classification rules. Fortunately, the Principal Component Analysis (PCA) is a data reduction technique giving new features (less than original features) that strongly differ across the classes. Hence, rules obtained from these new features always give classification performance better than rules of original features. Therefore, this research proposes to use the PCA as a data reduction technique to achieve its goal of obtaining high accurate minimal classification rules. Among decision tree classifiers, these three classifiers namely J48, REP Tree, and Random Tree are known of their ability of providing rules [8] ready to use in a rule based system. Therefore, the research aims at finding the best classifier, in terms of number of rules and classification accuracy, among them on PCA reduced data of Wisconsin Breast Cancer Data set (WBCD).

The rest of the paper is organized as follows. Section II presents related works. Section III describes the methodology. The experimental results are given in section IV. And Section V concludes the paper.

## II. RELATED WORKS

Many researches have been conducted on WBCD to obtain a high performance classifier supporting breast cancer diagnosis. Hind Elouedi et al [1] proposed a hybrid diagnosis approach of breast cancer based on decision trees and clustering. It is evaluated classification results by distinguishing different types of breast cancer. The aim is to improve the quality of classification and clustering of WBCD. The experimental results show that the splitting up of malignant instances into two clusters, and submitting them to the decision tree algorithm, they have gotten better results up to 95.14%. F.Kharbat and H.Ghalayini [2] presented a case study for building ontology from the set of rules which generated by a rule based learning system. The algorithm is used to extract and represent the rules generated from the original data based on WBCD in developing ontology elements. The results show that the rule set with only 25 rules can describe two concepts (Benign and Malignant). P. Hamsagayathri and P. Sampath [3] proposed to find the best performance of the four different decision tree algorithms for breast cancer classification such J48, REP Tree, Random Tree, Random Forest and Priority based decision tree. The experimental results indicated that the Random Forest presents the highest accuracy of 96.70 %, while Priority based decision tree, Random tree, REP Tree, and J48 classifiers gave the accuracies of 94.70 %, 94.13 %, 94.13 %, and 93.56 %, respectively. Among them, the random forest classifier could not produce classification rules. Ronak Sumbaly et al [4] built a detection model of breast cancer in its early stages based on WBCD. J48 decision tree classifier is used to model actual diagnosis. According to the results show that the performance of J48 achieved 94.56 % classification accuracy. J48 has the ability to generate the simple rules, flexible and highly efficient algorithm for breast cancer diagnosis problem and it is also maintain the accuracy in estimation. Chandra Prasetyo Utomo et al [5] applied Artificial Neural Network (ANN) with extreme learning machine technique to compare with BP ANN for diagnosing breast cancer. These techniques are to support in medical decision.

978-1-5386-4015-9/18/$31.00 @2018 IEEE

The results show that ELM ANN classifier can classify better than BP ANN. Nevertheless, these network problems failed to give the rules and take more time in computational complexity. Smita Jhajharia et al [6] implemented hybrid prediction model which combines principal component analysis (PCA) technique with ANN for feature processing and pattern recognition. This hybrid prediction model is compared with other classification algorithms (SVM, NB, DT, IBK, OneR). The experimented results show that PCA+ANN is the most effective. Kathija and S. Nisha [7] presented the performance of selecting the smallest subset of features from WBCD by using SVM and Naive Bayes to build efficiency classifiers.

## III. METHODOLOGY

### A. Principal Component Analysis(PCA)

PCA is a mathematical method used for data analysis. It is one of the most significant features extraction techniques [9]. Normally, PCA transforms a set of dependent variables into a set of independents which handles with uncorrelated variables called Principal Component (PC). Most of the largest possible variances will be retained in the first PC and then the next PCs will decrease the possible variances [10]. The objectives of PCA are to reduce the dimension of the data and select new variables that relevant to the best outcome. There are two approaches using in PCA .i.e Eigenvalues and Eigenvectors. An eigenvector represents the direction of the line (horizontal, vertical, etc) and an eigenvalue is a number of variances in the data's direction of eigenvector. The basic process of reducing data dimension by PCA which can reduce the rules of the model. It can be explained as below:

**Algorithm 1** PCA algorithm

1: Re-center the original dataset to the origin at means zero

2: Compute the sample variance-covariance

3: Compute the eigenvalues and eigenvectors.

4: Decision which principal components should be retained based on the eigenvalues in order to select highest to lowest eigenvalues. It could achieve 95% confidence interval.

5: Find the transformation matrix based on selection of PCs

### B. Decision Tree J48

J48 is an algorithm used to create a decision tree for decision making. It is an implementation of C4.5 algorithm by using Java application in the Weka Data mining tool [11]. Many problems have been solved by decision tree classification approach based on dividing and conquering strategy. It can be used to predict an unseen data set based on the various attribute value of the available data. The decision tree is represented by a rule based (if-then rules) which described by nominal and numeric properties. The construction of decision tree is built from a root node at the top of the tree to any leaf node that defined the feature. A branch feature may stop into a leaf node when searching for subset instance in the same class or may further create the leaf node when the nodes are not the same class. Every branch from the root node to leaf node is represented as a rule. It uses Gain Ratio as a splitting condition to separate the data set for normalizing the data into the form of information gain. The highest value of information gain ratio is selected as a root node and then splitting process is continued until reaching the leaf node.

$$Gain(S,A) = Entropy(S) - \sum_{j \in Values(A)} \frac{|S_j|}{|S|} Entropy(S_j) \quad (1)$$

Where, $j$ is possible values of $A$ , and $A$ is a set of all possible attribute. $S$ is a set of samples $\{X\}$, and $S_j$ is a subset where $(X_A = j)$. There are two parts in the equation (1). For the first part is entropy of original collection $S$ and the second part is the expected value entropy which calculates the sum of the entropies of each subset weighted by the fraction of examples

$$SplitEntropy(S,A) = -\sum_{i=1}^{n} \frac{|S_j|}{|S|} \times \log_2 \frac{|S_j|}{|S|} \quad (2)$$

Where, $SplitEntropy(S,A)$ is separated information of $A$ on the value of the categorical attribute $S$ .

$$GainRatio(S,A) = \frac{Gain(S,A)}{SplitEntropy(S,A)} \quad (3)$$

### C. Reduced Error Pruning(REP) Tree

REP Tree is constructed by decision tree that is used the information gain as the splitting condition. REP is used for pruning. It is the simplest ideas by using a pruning set to evaluate the performance accuracy of node and leaf depend on the decision tree process. REP Tree could reduce the error rate when applying with unseen data. Missing values are solved by applying C4.5's method with fraction instance. Moreover, it used less time for learning a model [12]. REP Tree is explored by beginning from bottom-up strategy.

### D. Random Tree

Random tree classifier is generated by randomly select features from a set of the tree that is possible with a different instance of the training data [12]. It uses a Bagging idea which provides random data set for creating a decision tree. It is a powerful technique to make a classification which is challenging to over-fit. The combination of a large set of Random Tree generally produces a correct model. Random Tree is selected a test set depends on a given number of the random features of each node. The decision on each node is random the procedure without pruning.

### E. Research Method

The research is an experiment on WBCD. It builds three classifiers (J48, REP Tree, and Random Tree) on the original

dataset and on the PCA-based reduced data set. Evaluation of each classifier is in terms of number of rules obtained and classification accuracy with 10-fold cross-validation. WEKA tool version 3.8.1 framework [13] is used as a tool in this study.

### F. Data Description

The data set of Breast Cancer reported by Dr. Wllliam H. Wolberg was collected as samples of clinical provided by the University of Wisconsin Hospitals. The data set has been stored in the UCI Machine Learning repository [14]. The total number of instances consists of 699 samples with 11 attributes, but some data have missed about 16 samples. There are two classes such as Benign (444 instances) and Malignant (239 instances). The numbers between 1 to 10 are used to record in the domain of each attribute. Sample code number is only an id number of the instance that does not affect the model. In the training and testing phases of the classification do not include the sample code number as in Table I.

### G. Data Preprocessing

Data preprocessing is a preliminary phase to perform on raw data which applies data normalization and separates incomplete data, outliers data and inconsistent data before the data is used to other procedures. PCA is a procedure to transform the data dimension and find a new set of variables by selecting the subset of principal component without losing the important feature. The best significant subset evaluation has been collected and used in the next step for the experiment.

TABLE I. ATTRIBUTES DESCRIPTION

| **Breast Cancer Dataset** | | |
|---|---|---|
| ***Attributes*** | ***Range*** | ***Data Type*** |
| Sample Code Number | No | Numeric |
| Clump thickness | 1-10 | Numeric |
| Uniformity of cell size | 1-10 | Numeric |
| Uniformity of cell shape | 1-10 | Numeric |
| Marginal adhesion | 1-10 | Numeric |
| Sigle epithelial cell size | 1-10 | Numeric |
| Bare nuclei | 1-10 | Numeric |
| Bland chromatin | 1-10 | Numeric |
| Normal nucleoli | 1-10 | Numeric |
| Mitosis | 1-10 | Numeric |
| Class | 2: benign<br>4: malignant | Nominal |

### H. Classification Algorithms

The classification phase learns the data set from previous step by using three different decision tree classification algorithms. There are Decision Tree (J48), REP Tree, and Random Tree.

### I. Performance Evaluation

In this section, three classifiers have been compared the accuracy results between data preprocessing with and without PCA. Among these two techniques, there will be a technique provide the minimal number of rules for breast cancer diagnosis.

The most effective approach to evaluate the performance of the model is based on the confusion matrix as shown in Table II. The confusion matrix is a specific table layout that shows the information about actual and predicted value by a classification model. There are four possible classification methods for each instance: a true positive (TP), a true negative (TN), a false positive (FP) and false negative (FN). Accuracy is computed by the equation (4)

TABLE II. CONFUSION MATRIX [15]

| | Predicted Positive | Predicted Negative |
|---|---|---|
| Actual Posituve | TP | FP |
| Actual Negative | FN | TN |

$$Accuarcy = \frac{TP + TN}{TP + FN + FP + TN} \quad (4)$$

## IV. Experimental Results

The Table III shows the classification's accuracy of rule based system in breast cancer data set without applying PCA between three classification algorithms. For this implementation 9 features from the data set are used. The results show that J48 classifier can classify 96.04 % correctly while REP Tree classifier performed 95.31% and Random Tree classifier computed 94.14%.

TABLE III. THE PERFORMANCE CLASSIFICATION WITHOUT PCA

| ***Classifiers*** | ***J48*** | ***REPTree*** | ***Random Tree*** |
|---|---|---|---|
| Correctly classified | 656 | 651 | 643 |
| Incorrectly classified | 27 | 32 | 40 |
| Accuracy (%) | 96.04 | 95.31 | 94.14 |

From Table IV indicates the results of three classifiers with PCA. From the data set, there are only 7 features used for classification. It shows that J48 classifier proved to be the most accurate classifier for WBCD with the accuracy of 97.36% by 1.32 % increased. In addition, REP tree and Random tree classifiers got the improvement by 1.46 % and 0.58 %, respectively.

TABLE IV. THE PERFORMANCE CLASSIFICATION WITH PCA

| ***Classifiers*** | ***J48*** | ***REPTree*** | ***Random Tree*** |
|---|---|---|---|
| Correctly classified | 665 | 661 | 647 |
| Incorrectly classified | 18 | 22 | 36 |
| Accuracy (%) | 97.36 | 96.77 | 94.72 |

The bar chart gives the information about the comparison of rule based system of three proposed classifiers with and without PCA approaches as shown in Figure 2. According to the chart, the number of rules are decreased in case of the classifiers with PCA. Especially, J48 and REP Tree classifier performed significantly in producing the rules better than Random Tree classifiers. They provide only 2 rules while Random Tree classifier needs 30 rules. However, the number of rules still high when PCA is not used i.e. REP Tree, J48, and Random tree classifiers need 7, 11, 34 rules, respectively.

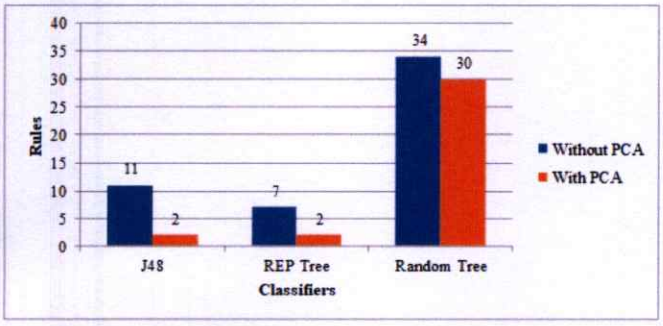



Fig. 2. Comparison performance of three classifiers

## V. CONCLUSION

A rule based system for breast cancer diagnosis has been a powerful tool supporting doctor diagnosis. Such a system requires classification rules derived from historical diagnosis. The desirable rules should be minimal in their number and give a good performance. This paper is to obtain such rules from the Wisconsin Breast Cancer data set. It performed experiments on the data set with PCA reduction to determine the best classifier among J48 decision tree, REP Tree, and Random Tree. It found that J48 classifier giving the best accuracy and smallest number of rules which are 97.36% and 2, respectively.

## ACKNOWLEDGMENT

The author would like to be grateful to International College, King Mongkut's Institute of Technology Lardkrabang for supporting and motivating during my research and thankful to Dr. William H. Wolberg at the University of Wisconsin for providing breast cancer data set which used in this paper.

## REFERENCES

[1] H. Elouedi, W. Meliani, and Z. Elouedi, "A hybrid approach based o decision trees and clustering for breast cancer classification," in 6t International Conference of Soft Computing and Pattern Recognitio (SoCPaR), 2014.

[2] F.Kharbat, H.Ghalayini, "New algorithm for Building Ontology fron Existing Rules: A Case Study," in International Conference o Information Management and Engineering, 2009, pp. 12-16.

[3] P. Hamsagayathri and P. Sampath, "Performance analysis of breas cancer classification using decision tree classifier," International Journa Of Current Pharmaceutical Research (IRCPR), vol. 9, 2017.

[4] R. Sumbaly, N. Vishnusri, and S. Jeyalatha, "Diagnosis of breast cance using decision tree data mining technique," International Journal o Computer Applications, vol. 98, no. 10, p. 0975 8887, 2014.

[5] C. P. Utomo, A. Kardina, and R. Yuliwulandari, "Breast cance diagnosis using artificial neural networks with extreme learnin techniques," International Journal of Advanced Research in Artificia Intelligence (IJARAI), vol. 3, no. 7, pp. 10–14, 2014.

[6] S.Jhajharia, H.K.Varshney, S.Verma, and R.Kumar, "A neural networ based breast cancer prognosis model with PCA processed features," i International Conference on Advances in Computing, Communication and Informatics (ICACCI), 2016.

[7] Kathija and S.Nisha,"Breast cancer Data Classification Using SVN an Naive Bayes Techniques," International Journal of Innovative Researc in Computer and Communication Engineering(IJIRCCE), vol. 4, pp. 2 167–21 175, 2016.

[8] D.Lavanya and D. Rani, "Evaluation of Decision Tree Classifiers o Tumor Datasets," International Journal of Emerging Trends Technolog in Computer Science (IJETTCS), vol. 2, pp. 418–423, 2013.

[9] T. M. Mohamed, "Efficient breast cancer detection using sequential fea ture selection technique," in 7th International Conference on Intelliger Computing and Information Systems (ICICIS), 2015.

[10] J.-W.Liu, Y.H.Chen, and C.H.Cheng, "Owa based information fusio method with PCA preprocessing for data classification," in Internationa Conference on Machine Learning and Cybernetics, 2012, pp. 3322 3327.

[11] T. R. Patil and S. S. Sherekar, "Performance analysis of naive bayes an j48 classification algorithm for data classification," International Journa Of Computer Science And Applications, vol. 6, no. 2, 2013.

[12] I. H. Witten and E. Frank, Data Mining: Practical machine learning tool and techniques. Morgan Kaufmann Publishers Inc., San Francisco,CA USA, 2nd edition, 2005.

[13] A. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H Witten. The weka data mining software:. [Online]. Available http://www.cs.waikato.ac.nz/ml/weka/downloading.html

[14] W.H.Wolberg, O.Mangasarian, and D.W.Aha.UCI Machine Learnin Repository. [Online]. Available: https://archive.ics.uci.edu/ml/index.php

[15] K. M. Ting, Confusion Matrix. Springer US, 2010.

# AUTHOR BIOGRAPHY

| | |
|---|---|
| Name | Phonethep Douangnoulack |
| Degree | Master of Engineering |
| Date of graduation | 18th July 2018 |
| Date of Birth | 01st December 1994 |
| Place of Birth | Thoulakhome District, Vientiane Province, Laos |

**Undergraduate and Graduate Education:**

Master of Engineering in Computing in Engineering Systems,

King Mongkuts Institute of Technology Ladkrabang, Bangkok 10520, Thailand, 2018

Bachelor degree in Computer Science,

National University of Laos, Vientiane Capital, Laos, 2015

**Major:** Computing in Engineering Systems

**Presentations and Publications:**

1.) P. Douangnoulack and V. Boonjing, Building Minimal Classification Rules for Breast Cancer Diagnosis,

10th International Conference on Knowledge and Smart technology, pp. 278-281, March 2018