PERFORMANCE COMPARISONS OF MACHINE LEARNING APPROACHES FOR NUMERICALLY STRUCTURED BREAST CANCER DATA CLASSIFICATIONS

Nattaphon Talmongkol

A Thesis Submitted in Partial Fulfillment of the Requirements
For the Degree of Master of Engineering Program in Engineering Technology
Graduate School
Thai-Nichi Institute of Technology
Academic Year 2018

Thesis Topic	Performance Comparisons of Machine Learning Approaches for Numerically Structured Breast Cancer Data Classifications.
By	Nattaphon Talmongkol
Field of Study	Engineering Technology
Thesis Advisor	Asst. Prof. Dr. Wimol San-Um

The Graduate School of Thai-Nichi Institute of Technology has been approved and accepted as partial fulfillment of the requirement for the Master's Degree

…….. Dean of the Graduate School

(Assoc. Prof. Dr. Pichit Sukcharoenpong)

Month.....…….. Date…....Year…….

Thesis Committees

…….. Chairperson

(Dr. Thepchai Supnithi)

…….. Committee

(Assoc. Prof. Dr. Warakorn Srichavengsup)

…….. Committee

(Dr. Pramuk Boonsieng)

…….. Advisor

(Asst. Prof. Dr. Wimol San-Um)

NATTAPHON TALMONGKOL: PERFORMANCE COMPARISONS OF LEARNING APPROACHES FOR NUMERICALLY STRUCTURED BREAST CANCER DATA CLASSIFICATIONS. ADVISOR: ASST. PROF. DR. WIMOL SAN-UM, 107 PP.

This thesis has compared the performance of a machine learning approaches for numerically structured breast cancer data classifications. The data classification is one of data analytics methods which common use to encourage the decision-making aim of many businesses, medical, healthcare or any requirement that impact to the human life. Furthermore, this thesis has been perceived the motivation from the big data analytics technologies which can drive to improve Thailand technologies in the present trend. Subsequently, the breast cancer dataset was selected from the UCI data repository in the large of a dataset stock to support many researchers to perform the data analytics purposes. Moreover, this dataset contains 30 features of three breast cell nucleus which shown the 569 instances of historical diagnostic. Nonetheless, the machine learning techniques which selected to perform in this thesis can be divided into four algorithms of the Decision Tree (DT), Naive Bayes (NB), Artificial Neural Network (ANN), and Support Vector Machine (SVM). In addition, the cross-validation method is conducted to evaluate the performance of each technique by random the parameter customization to find the highest performance which separated into three categories of accuracy rate, error rate, and classification lead time. The research objective is to find the appropriate of a classifier that has provided the highest performance using the RapidMiner studio 7.4 program. The results have shown that the highest performance of machine learning approach for breast cancer data classification is SVM technique which shown the accuracy percentage of 96.84%, the F-measure (M) percentage of 95.70%, the F-measure (B) percentage of 97.50%, the RMSE of 0.194, and classification lead time of 0.52 second followed by the Decision Tree, Artificial Neural Network, and Naive Bayes respectively.

Graduate School Student's Signature……

Field of Study Engineering Technology Advisor's Signature……

Academic Year 2018

Acknowledgement

I would like to express my sincere thanks to my thesis advisor, Asst Prof. Dr. Wimol San-um for his invaluable help and constant encouragement throughout the course of this research. I am most grateful for his teaching and advice, not only the research methodologies but also many other opportunities in life. I would not have achieved this far and this thesis would not have been completed without all the support that I have always received from him.

In addition, I am grateful for the very good advice from my chairman, Dr. Thepchai Supnithi and teachers of engineering technology program in Thai Nichi Institute of Technology for suggestions and all their help.

Finally, I most gratefully acknowledge my parents and my friends for all their support throughout the period of this research.

Nattaphon Talmongkol

Table of Contents

Table of Contents (Continued)

List of Table

List of Figures

List of Figures (Continued)

List of Figures (Continued)

List of Figures (Continued)

Chapter 1
Introduction

1.1 Introduction

This chapter provides the background of a research which is the performance comparison of machine learning approach for numerically structured breast cancer data classification. The dataset of breast cancer is provided in order to analyze and use for conduct the data classification purpose. Subsequently, four machine learning techniques are selected to predict the classification result which is processed via the RapidMiner studio 7.4 program.

1.2 Background

The data classification is one of the categories of data analytics technique which is popular for an approach to predicting any requirement result of big data. In the present, the big data analytics is of most interest to support the business growth of Thailand 4.0 model. Furthermore, the big data analytics technology is used to apply for medical and healthcare because these are useful for human life. Thus, various diagnostics of disease is collected through analysis to predict the result and assist the doctor in decision making which highly advantages for new medical technology.

This research has selected the breast cancer disease dataset to study the performance comparison of machine learning approach for data classification. This is common among Thai women and over the world which is about 16% of cancer that occurred in the women. Usually, the breast cancer has occurred in the adult which can be divided into 4 periods [1]. However, there can be treated if early detecting. Currently, there is an effective way of screening the breast cancer in the women of self-breast examination or breast exams with mammograms. Furthermore, the breast cancer can be screening and diagnostic via fine needle aspirate (FNA) of a breast mass [2]. The several data of fine needle aspirate (FNA) of the breast mass have been recorded from the patients who receive diagnose. Which is the better way if can classify these data through machine learning

techniques and conduct for automatically decision making since appearing the FNA of the breast mass result. Moreover, this method is one of choice that can encourage the doctor to the decision of diagnosing of breast cancer disease. On the other hand, machine learning techniques that conduct to compute the data classification have consisted the several techniques which different way to computation. Thus, this research is focusing to compare the performance of each algorithm what highly efficiency. The machine learning techniques that are selected for this study can be divided into four techniques of a Decision tree (DT), Naive Bayes (NB), Artificial Neural Network (ANN), and Support Vector Machine (SVM).

The decision tree (DT) technique is an agent of computational modeling which decide by statistics of the dataset with an if-then rule base. The Naive Bayes (NB) technique is the algorithm which computes by using the probabilistic of the result on the historical dataset which based on the Bayes' theorem. The Artificial Neural Network (ANN) is the computation model which simulate the neuron of the human. This algorithm is popular using in the artificial intelligence technologies. Support Vector Machine (SVM) is the algorithm which decision-based on suitable distance in feature space via support vector. The original support vector machine was used for linear data. In fact, most of the information that is used in the real world was usually nonlinear which can be solved by using the Kernel function for computation. The four techniques of machine learning that selected in this research are supervised learning. Because all are computed based on the historical dataset which learning before considering the classification result. Nonetheless, this research is using the RapidMiner studio 7.4 program to conduct analyze and compare the classification performance of these four techniques. The classification performance results that focusing can be separated into three type of the accuracy percentage, Error rate, and computation time of each machine learning. The accuracy percentage is considered by accuracy value and F-measure score value. The error rate of prediction is considered by a root mean square error (RMSE).

1.3 Motivation

Currently, several machines learning method are applied to support the artificial intelligence technologies. Thus, this research would like to know about which machine learning is the highest performance in term of data classification. Nonetheless, the classification performance that analyzes in this research is decided on a numerical dataset of breast cancer from the UCI dataset repository [3].

1.4 Statement of Problems

1.4.1 The highest performance of machine learning technique for numerical structured breast cancer data classification.

1.4.2 The appropriate algorithm for the numerical breast cancer dataset classification on a binary problem.

1.5 Objective

To evaluate the performance of machine learning technique approaches for numerically structured breast cancer data classification with the binary problem.

1.6 Scope of Research

The scope of this research can be seen in the research framework as Fig. 1.1.

1.7 Expected Outcome

1.7.1 Understanding the computation method of four machine learning algorithms of Decision Tree, Naive Bayes, Artificial Neural Network, and Support Vector Machine.

1.7.2 Can be applied the machine learning technique for big data analytics.

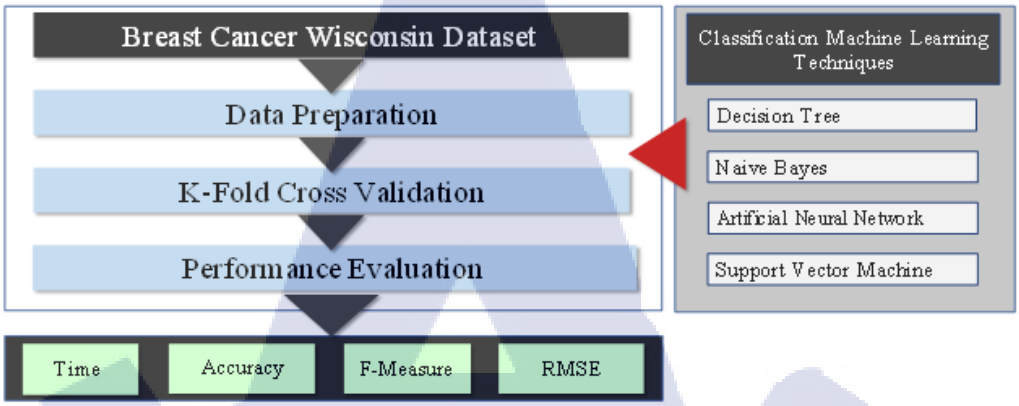


Figure 1.1 Research Framework

1.8 Conclusion

This chapter has introduced the background of a research problem and provided the framework that would evaluate and compare the performance of each machine learning approach for the breast cancer data classification. Subsequently, summarized the objective as well as the expected outcomes of the research.

1.9 Research Plan

Table 1.1 The research schedules in October 2017 until August 2018.

Research Methodology	**Year**										
	2017			**2018**							
	10	**11**	**12**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1. Dataset Collection: Breast Cancer Wisconsin Dataset from UCI Machine learning.											
2. Data Preparation: Dataset analysis and Normalization.											

Table 1.1 The research schedules in October 2017 until August 2018. (Cont.)

Research Methodology	Year										
	2017			2018							
	10	11	12	1	2	3	4	5	6	7	8
3. Design the performance evaluation method and machine learning technique selection.											
4. Conduct the performance evaluation into RapidMiner Studio 7.4 program.											
5. Summarized the performance of each machine learning technique											
6. Performance comparison and conclusion.											
7. Thesis summary											

1.10 Keyword Descriptions

1.10.1 Data Classification

Data Classification is the data modeling process that manages the data in the assigned group which shows the difference between a classes or a group of data and to predict what information should be included in any class.

1.10.2 Machine Learning Techniques

Machine learning technique is an artificial intelligence discipline which occurred from the technological development of human knowledge. Machine learning allows computers to handle new situations or decision making through the analytics process which can be divided into two types of supervised learning and unsupervised learning.

1.10.3 Performance Evaluation

Performance evaluation is the assessment process of classified results in each machine learning technique which can be separated into four methods of accuracy, F-measure, root mean square error, and time.

Chapter 2
Related Theories and Literature Reviews

This chapter has described the related theory and including of literature reviews which this research has used to meet a research target.

2.1 Related Theories

2.1.1 Data Classification

Classification is a common task in human activities that involve decision or forecast in an unknown or a future situation by using currently available information. Furthermore, classification is the process of constructing a model or function which describes and distinguishes different data classes or concepts. The propose of being able to use the model to predict the class of objects whose class label is unknown later. The derived model is based on the analysis of a set of training data. Moreover, classification is referred as a pattern recognition, discrimination, or supervised learning which contrast with unsupervised learning or clustering where no classes are predefined but they are inferred from the data. There have been several applications of classification to solve scientifically, industrial, medical, and commercial problems. However, some typical classification purposes are the detection of letter from a character image, the credit status assignment for a customer on the basis of financial and other personal information, and the preliminary diagnosis of a patient's disease during a waiting for definitive test result [4].

In learning a classification model, there exist various forms in expressing the model derived. Some common forms are IF-THEN rules, decision trees, mathematical formula, or neural network. As Figure 2.1 shows the sample dataset for data classification, Figure 2.2 shows the decision tree model which is a flow chart like tree structure where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions. Decision tree can easily convert to classification (IF-THEN) rules which can be seen in Figure 2.3. Figure 2.4 shows linear regression equations. A neural network is typically a collection of neuron-like processing

units with weighted connections between the units as can be seen in Figure 2.5. There are several other methods for constructing classification models such as Naive Bayes classification, Support Vector Machine, and K-nearest neighbor classification.

Outlook	Temp.	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Figure 2.1 A Sample Dataset (The Play-Tennis Dataset)

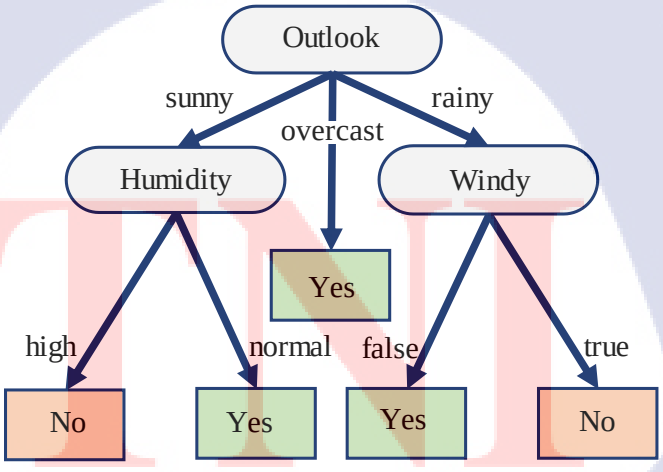


Figure 2.2 Decision Tree

1. **If** (Outlook = "Overcast") **then** Play = "Yes")
2. **If** (Humidity = "Normal" and Windy = "False") **then** Play = "Yes")
3. **If** (Temp = "Mild" and Humidity = "Normal") **then** Play = "Yes")
4. **If** (Outlook = "Rainy" and Windy = "False") **then** Play = "Yes")

Figure 2.3 Classification Rules (IF-THEN)

Play(yes) =	0.6 * outlook(sunny) + 1.0 * outlook(overcast) + 0.2 * outlook(rainy) + 0.1 * temp(hot) + 0.2 * temp(mild)+ 0.2 * temp(cool) + 0.5 * humidity(high) + 0.8 * humidity(normal) + 0.6 * windy(false) + 0.3 * windy(true)
Play(no) =	0.3 * outlook(sunny) + 0.1 * outlook(overcast) + 0.7 * outlook(rainy) + 0.2 * temp(hot) + 0.1 * temp(mild)+ 0.3 * temp(cool) + 0.7 * humidity(high) + 0.1 * humidity(normal) + 0.3 * windy(false) + 0.8 * windy(true)

Figure 2.4 Linear Regression Equations

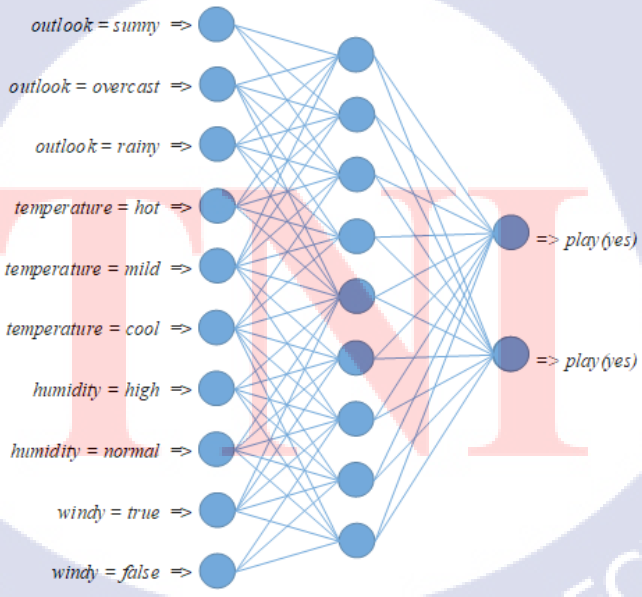


Figure 2.5 An Example of an Artificial Neural Network

2.1.2 Normalization

Normalization of the data is the data processing that transforms each attribute value become to the same range or same standard. This step is very important when dealing with parameters of different units and scales. For example, some data mining techniques use the Euclidean distance. Therefore, all parameters should have the same scale for a fair comparison between them. Two methods are usually well known for rescaling data. Normalization, which scales all numeric variables in the range (0, 1). One possible formula is given below.

$$X_{new} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

2.1.3 Machine Learning Techniques

This research has applied machine learning techniques to approach for the data classification and performance comparison, which can be divided into five techniques of Decision Tree (DT), Naïve Bayes (NB), Support Vector Machine (SVM), and Artificial Neural Network (ANN).

2.1.3.1 Decision Tree (DT)

A decision tree is a decision support tool, which is a non-parametric supervised learning method commonly applied to classification and regression of multiple variable analyses. The decision tree has a tree-shaped diagram for representing a possible decision and consequences, involving, for instance, chances, event outcomes, resource costs, and utilities. Typically, the decision tree can be constructed by Entropy (Ent) and Information Gain (IG). The Entropy is average number of binary questions which are in the form of infinitely trials to distinguish events, and can be calculated by [5]

$$Ent(ci) = -p(ci)\log 2\, p(ci) \quad (2)$$

where p(ci) is a probability of dataset i =1,2, 3...n. Generally, entropy is always nonnegative and is zero when one items ci has a unity probability. The IG is the change in entropy from prior states to a state, and is based on the decrease in entropy after a dataset is split on an attribute. The IG can be found as follows

$$IG = Ent(PR) - \{[p(c_1) \times Ent(c_1)] + [p(c_2) \times Ent(c_2)] + ...\} \quad (3)$$

where Ent (PR) is an information entropy of overall datasets before splitting.

2.1.3.2 Naïve Bayes (NB)

Naïve Bayes technique is a family of probabilistic classifiers based on Bayes' theorem with independence assumptions among features. The calculation of posterior probability is given by

$$p(\mathrm{b}|\mathrm{a}) = \frac{p(\mathrm{a}|\mathrm{b}) \times p(\mathrm{b})}{p(\mathrm{a})} \quad (4)$$

where p(b|a) is the probability that class b occurs before class a, p(a|b) is the probability that class an occurs before class b, p(a) is the probability of occurrence a, and p(b) is the probability of occurrence b. Such a Naïve Bayes technique provides uncomplicated computation process as each distribution can be independently estimated as a one-dimensional distribution [6].

2.1.3.3 Artificial Neural Network (ANN)

An artificial neural network is a computational system composed by highly interconnected processing elements based on the structure and functions of biological nervous systems. The ANN processes information through dynamic state response to external inputs and learning process. Figure 2.6 (1) and (2) show a single neuron model and a three-layer ANN, respectively. It is seen in Figure 2.6 (1) that the

number of input element vectors R, which is weighted by a gain W, is combined with a bias b, and the combination result n is fed to an activation function f, which is a sigmoidal function for this case, providing the result a. On the other hand, Figure 2.6 (2) illustrates a full diagram of the ANN composed by inputs, hidden, and outputs with a total of S layers. The generalized mathematical model of the ANN can be expressed as

$$a_S = f(w_{S,R} P_R + b_S) \tag{5}$$

Generally, the ANN can be configured for several specific applications, such as pattern recognition, data classification, clustering, and prediction.

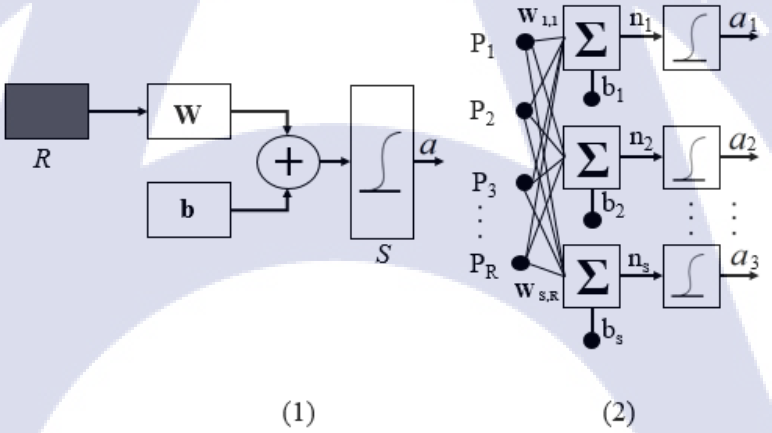


Figure 2.6 General ANN diagram; (1) a single neuron model, (2) a three-layer ANN

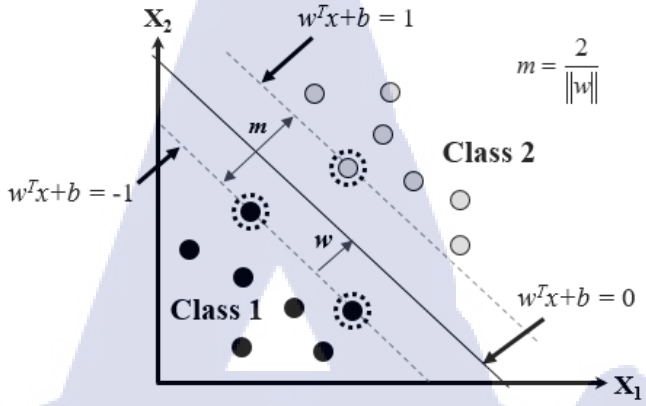


Figure 2.7 Demonstration of Linear Separating Hyper Planes for the Separable 2 Dimensional Case of SVM Technique

2.1.3.4 Support Vector Machine (SVM)

Support Vector Machine is principally a discriminative classifier that performs both regression and classification by constructing hyper planes in a multidimensional space that separates cases of different classes. Generally, SVM offers effective in high dimensional spaces and exploits less memory since a subset of training points in the decision function is realized. Moreover, SVM provides versatility in terms of Kernel functions types for any specific classification purposes. Fig. 2.7 demonstrates linear separating hyper planes for the separable 2-dimensional case. It can be seen from Fig. 2.7 that the support vectors are highlighted with large circle. Intuitively, the decision boundary should be as far away from the data of both classes as possible. This property implies the maximization of the margin (m). With reference to Fig. 2.7, given the training data {xi, yi} for i=1, 2, 3…, n, xi ϵ Rd, yi ϵ {-1, 1} where xi is datum, representing by a vector with the d dimension and y is a binary class of -1 or +1, the support vector machine finds the best hyper plane which separate the positive from the negative example, i.e. a separating hyper plane. In principle, the points x on the hyper plane satisfy the formula wTx+b=0 [7]

2.1.4 Cross Validation

Cross-validation is the evaluate method of classification model in each machine learning techniques [8]. Which separates the initial datasets into training datasets to train the classification model, and a test dataset to evaluate the classification model. The k-fold cross-validation is randomly separate the original datasets into k-set. The k-1 datasets are used as training data and the remaining single dataset is used as the validation data in-term of testing. The iteration process was performed K times and typically used in data classification. It is seen in the Figure 2.8.

2.1.5 Performance Evaluation

The performance of machine learning technique would be evaluate using multi-method, which can be separated into 4 methods of F-Measure which calculate based on the result of recall and precision, accuracy, root mean square error, and time. The first three calculation method are based on the confusion matrix as seen in Figure 2.9.

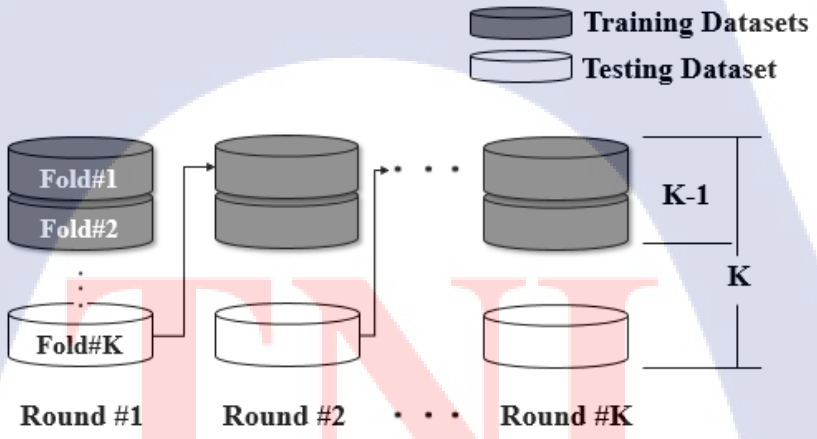


Figure 2.8 Demonstration of K-Fold Cross Validation Method

Confusion Matrix		Predicted Class	
		Yes	No
Actual Class	Yes	***TP***	*FN*
	No	*FP*	***TN***

Figure 2.9 Confusion Matrix

2.1.5.1 Recall and Precision

A recall is the ratio of a number of events that can correctly recall to number of all correct events. Precision is the ratio of a number of events that can correctly recall to a number all events recall. In other words, it is how precise of the recall. And both also can be calculated based on confusion matrix, which calculation equation as follows.

$$Recall(true) = \left| \frac{TP}{TP + FN} \right| \tag{6}$$

$$Precision(true) = \left| \frac{TP}{TP + FP} \right| \tag{7}$$

2.1.5.2 F-Measure

The F-measure is defined as the weighted harmonic mean of its precision and recall. Which is also can be calculated based on confusion matrix and the calculation equation as follows.

$$F - Measure(true) = \left| \frac{2 \times Recall(true) \times Precision(true)}{Recall(true) + Precision(true)} \right| \tag{8}$$

2.1.5.3 Accuracy

Accuracy is perhaps the most intuitive performance measure. It is simply the ratio of correctly predicted observations, which can be calculated through confusion matrix and the calculation equation as follows.

$$Accuracy = \left| \frac{TP + TN}{TP + FN + TN + FP} \right| \quad (9)$$

2.1.5.4 Root Mean Square Error (RMSE)

The RMSE is the standard deviation of the predictions from the ground-truth. Furthermore, this is one way to measure the performance of a classifier which considers the error rate or number of misclassifications. The RMSE of data classification can be calculated from confusion matrix which consists the true positive rate (TP), true negative rate (TN), false positive rate (FP), and the false negative rate (FN) in the case of a binary problem. The total size of the test dataset is Omega (TP+TN+FP+FN). Thus, the RMSE can be calculated as follows.

$$RMSE = \sqrt{\left(\frac{FP+FN}{TP+TN+FP+FN}\right)} \quad (10)$$

2.1.5.5 Classification Lead time (Time)

The lead time is one of the performance evaluation of data classification which is focusing on the processing lead time that each classifier using for computation and prediction. The processing lead time is starting since running the program until stop and shows the data classification result.

2.1.6 RapidMiner Studio7.4 program

RapidMiner Studio 7.4 is the data analytics program which can be analyzed the big data and find the useful information to support some business or research [8]. The

main functions of this program can be divided into six parts of RapidMiner Menu, Repository, Processes, Operators, Parameters, and Help. It is seen in Figure 2.10 The RapidMiner Menu is provided the tools to support general activities in the program such as save, save as, edit, and others as can be seen in Figure 2.11. The Repository of RapidMiner is provided to keep the activities of the program that user created such as Database, Process, and Model. It is seen in Figure 2.12. The Process function is provided the space which can generate the process to analyze the data as can be seen in Figure 2.13. The Operator function of RapidMiner is provided all tools that support data analytics activities. It is seen in Figure 2.14. The Parameters function is the zone which can adjust any parameters value in each operator. It is seen in Figure 2.15. The Help function is the support tools which explain in the details of each operator or parameter which make easier the understanding as can be seen in Figure 2.16.

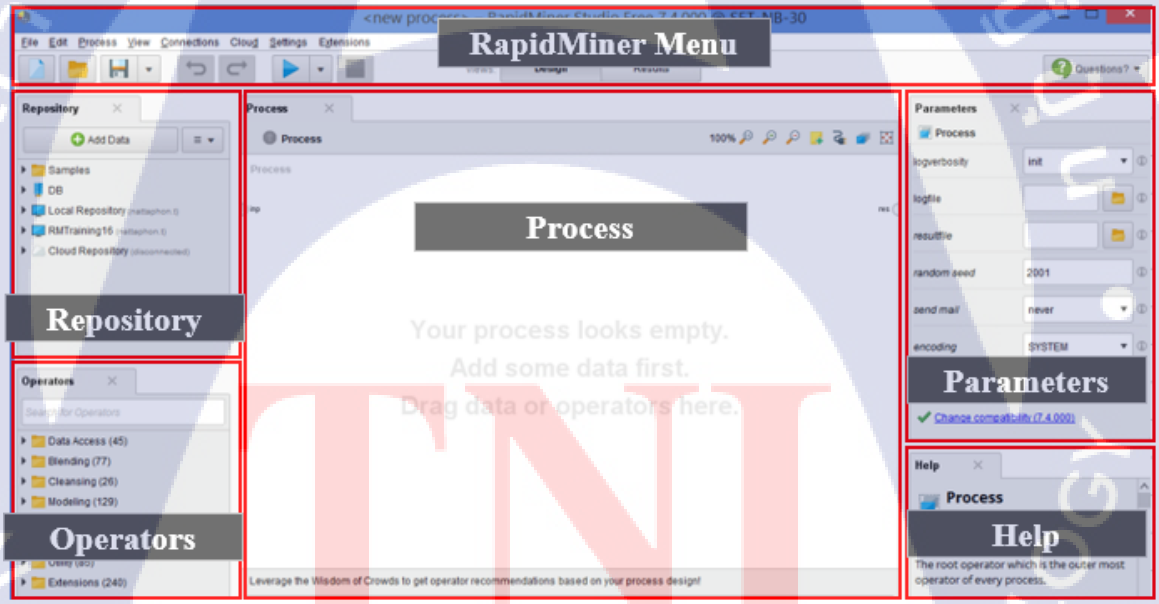


Figure 2.10 Main Functions of RapidMiner Studio7.4

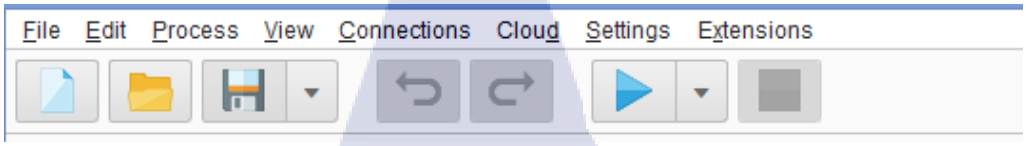


Figure 2.11 RapidMiner Studio7.4: Menu

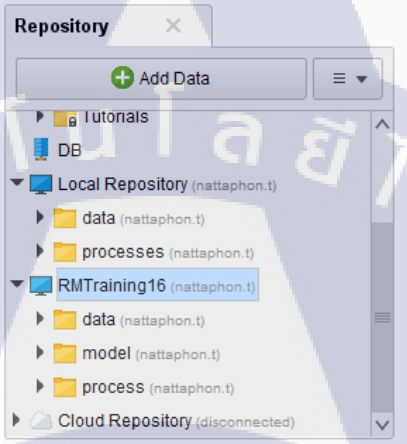


Figure 2.12 RapidMiner Studio7.4: Repository Function

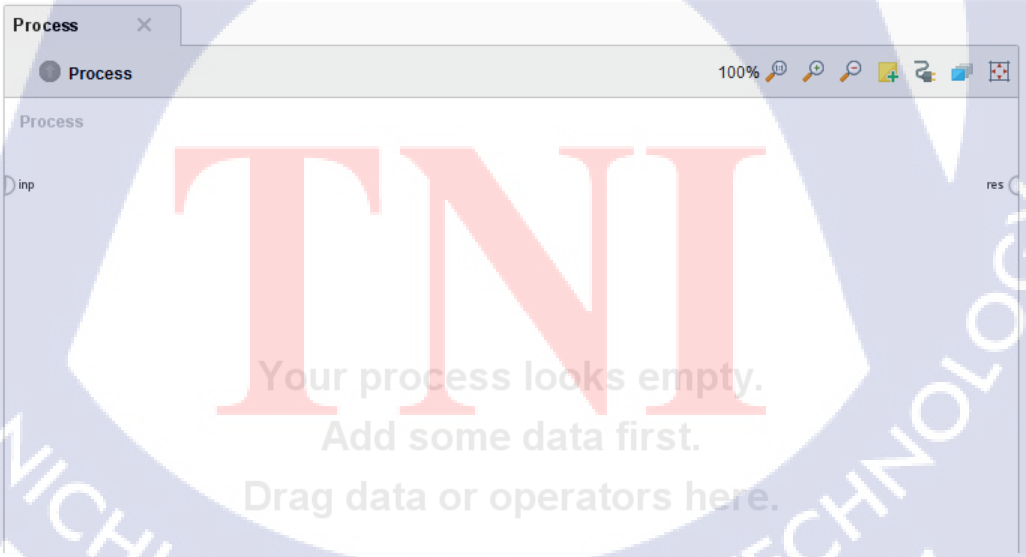


Figure 2.13 RapidMiner Studio7.4: Process Function

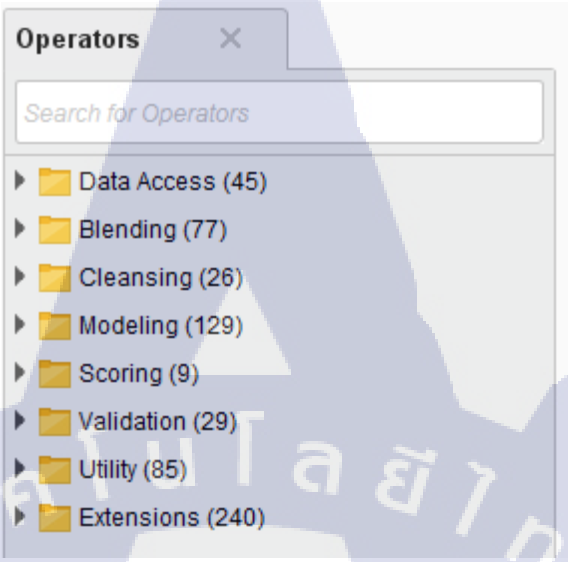


Figure 2.14 RapidMiner Studio7.4: Operators Function

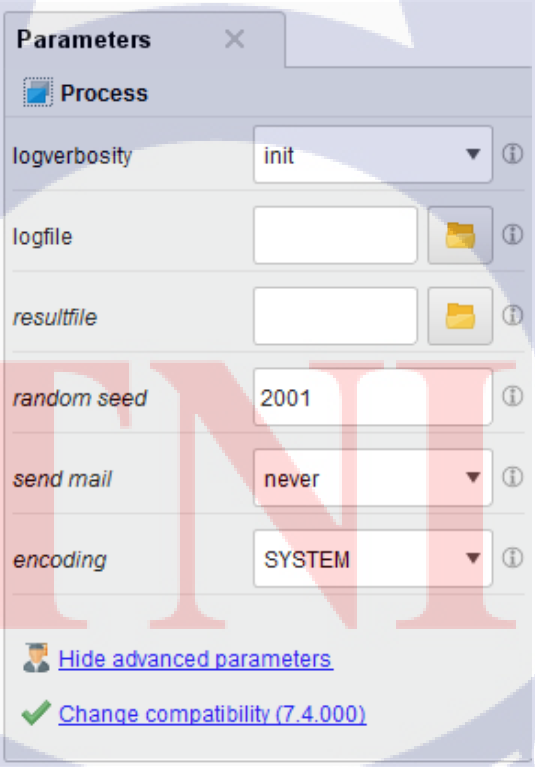


Figure 2.15 RapidMiner Studio7.4: Parameters Function

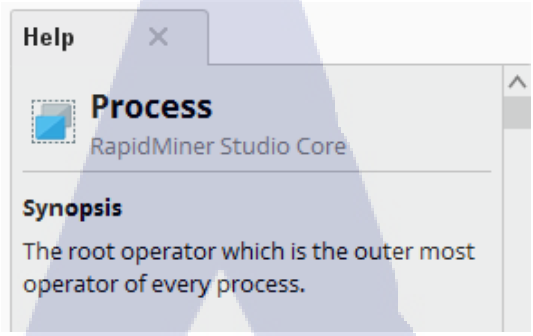


Figure 2.16 RapidMiner Studio7.4: Help

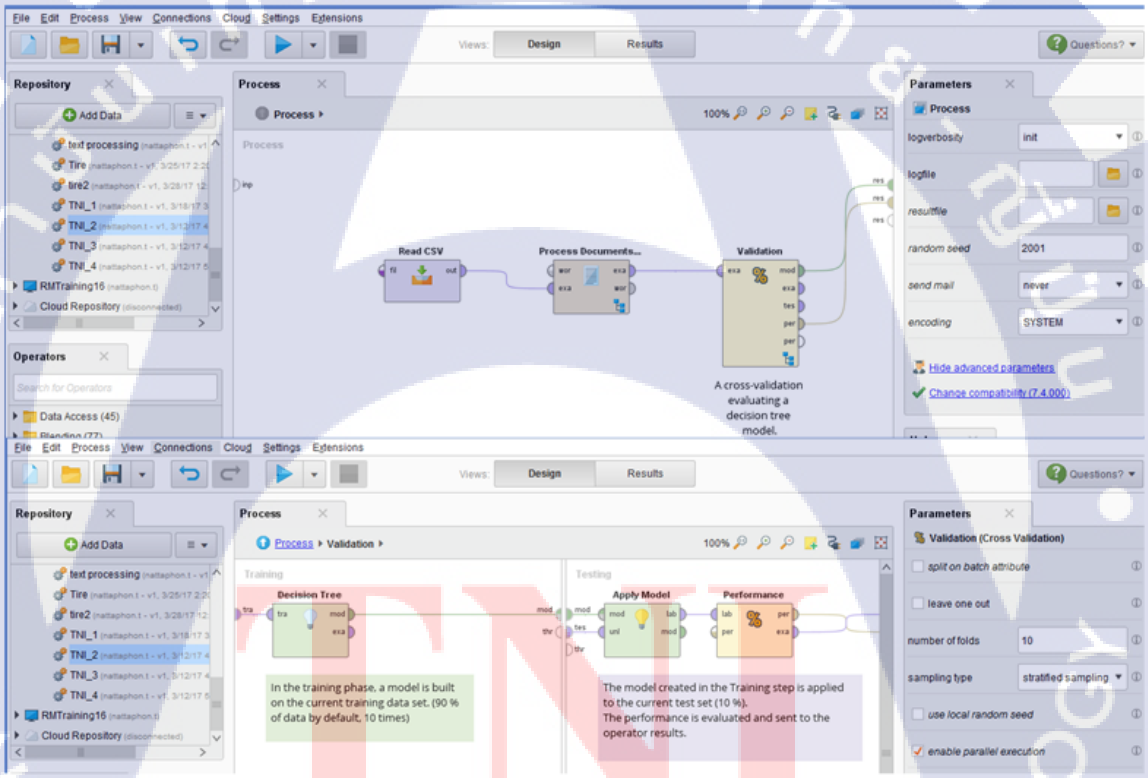


Figure 2.17 RapidMiner Studio7.4: Create the process

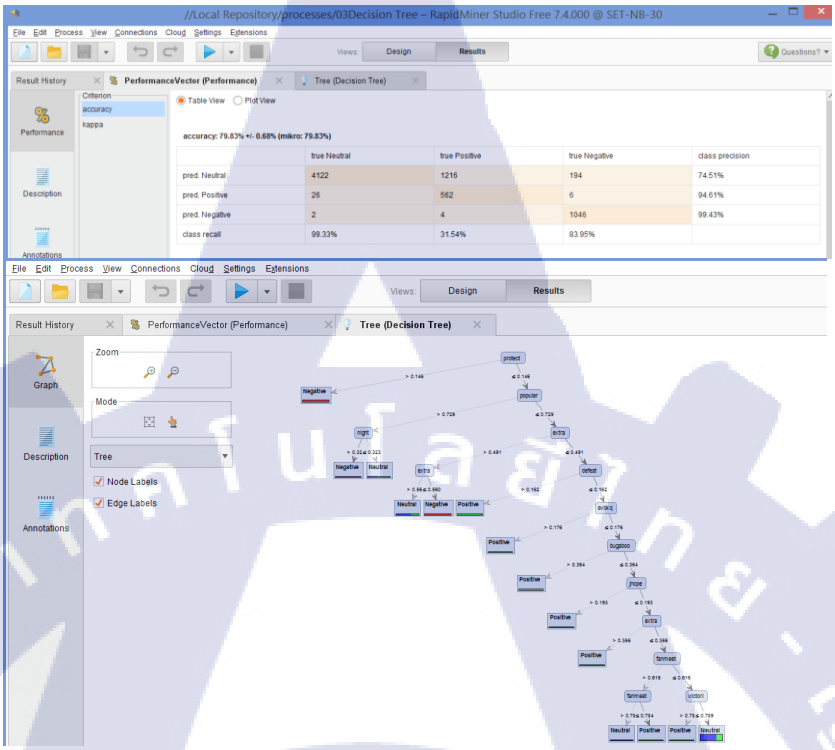


Figure 2.18 RapidMiner Studio7.4: Result

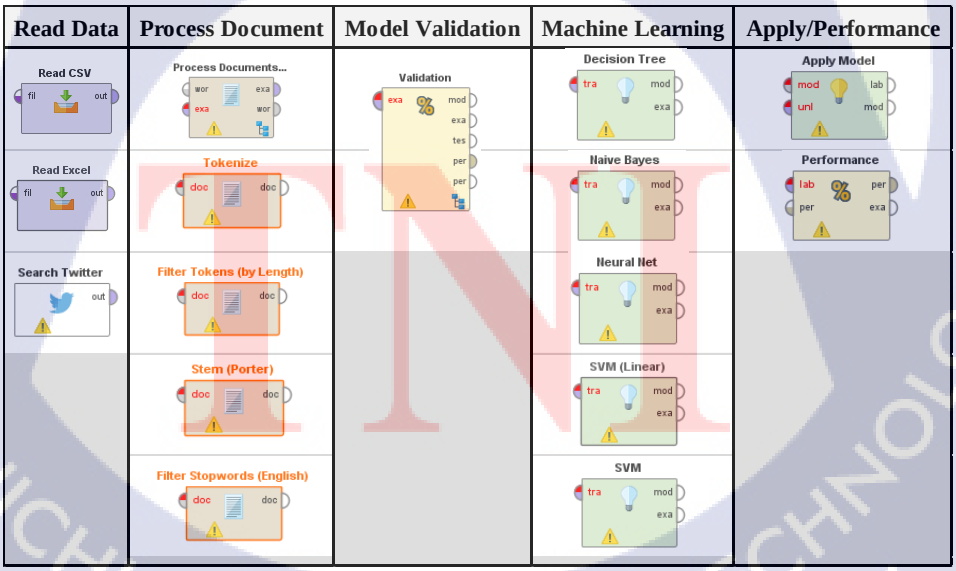


Figure 2.19 RapidMiner Studio7.4: Example of Operators

RapidMiner program can create any processes by selecting the operators and drop into the process space. Subsequently, the users must connect the operators through line connecting based on the priority of each operator as can be seen in Figure 2.17 which the data analytics result has shown in the window as can be seen in Figure 2.18. Nonetheless, the useful information or performance reports are showing in this window such as the classification result, statistical result, and confusion matrix. It can be seen in Figure 2.19.

2.2 Literature Reviews

This research has been done of the literature reviews as follows

Table 2.1. Literature Reviews

No.	Author	Year	Title
1	Tzu-Tsung Wong and Nai-Yu Yang.	2017	Dependency Analysis of Accuracy Estimates in k-fold Cross Validation
2	Arnau Mata Llenas and et al.	2017	Performance Evaluation of Machine Learning Based Signal Classification using Statistical and Multiscale Entropy Features
3	V.Shanmugarajeshwari and R. Lawrance	2016	Analysis of Students' Performance Evaluation using Classification Techniques
4	Praveen Kumar and et al.	2016	Analysis of Various Machine Learning Algorithms for Enhanced Opinion Mining using Twitter Data Streams
5	A. Swarupa Rani and S. Jyothi.	2016	Performance Analysis of Classification Algorithms Under Different Datasets

Table 2.1. Literature Reviews (Cont.)

No.	Author	Year	Title
6	Rafet Duriqi and et al.	2016	Comparative Analysis of Classification Algorithms on Three Different Datasets using WEKA
7	Govin Gaikwad and Prof. Deepali J. Joshi.	2016	Multiclass Mood Classification on Twitter Using Lexicon Dictionary and Machine Learning Algorithms
8	Sadia Zaman Mishu. and S.M. Rafi Uddin.	2016	Performance Analysis of Supervised Machine Learning Algorithms for Text Classification
9	Zehra Aysun Al tlkardes and et al.	2015	Performance Evaluation of Classification Algorithms by Excluding the Most Relevant Attributes for Dipper/Non-Dipper Pattern Estimation in Type-2 DM Patients
10	Zahra Nematzadeh and et al.	2015	Comparative Studies on Breast Cancer Classifications with k-fold Cross Validations using Machine Learning Techniques
11	Tanu Verma and et al.	2014	Tokenization and Filtering Process in RapidMiner.
12	Kavita Choudhary and et al.	2014	Glaucoma Detection using Cross Validation Algorithm: A Comparative Evaluation on RapidMiner

2.2.1 Dependency Analysis of Accuracy Estimates in k-fold Cross Validation

This research has studied the k-fold cross-validation method, which is the performance evaluation of machine learning techniques. There are twenty datasets have collected and performed. The result has shown that the many overlapping training datasets

of this method are generated the higher accuracies than is not overlapping training datasets [9].

2.2.2 Performance Evaluation of Machine Learning Based Signal Classification using Statistical and Multi-scale Entropy Features

This research has done for the performance evaluation of signal classification using Support Vector Machine technique. The researcher has used feature selection technique to increase the accuracies result and the resulting show that the feature extending is provided the high accuracies and close to the actual accuracies [10].

2.2.3 Analysis of Students' Performance Evaluation using Classification Techniques

This research has analyzed the students' performance through data classification. Subsequently, the researchers have used Decision Tree in order to meet their target. The classification results show 100% accuracy, which can be classified the performance into each student [11].

2.2.4 Analysis of Various Machine Learning Algorithms for Enhanced Opinion Mining using Twitter Data Streams

This research has conducted the sentiment analysis of people opinion from Twitter. The variety of machine learning have selected and can be divided into four techniques of Naive Bayes, Random Forest, Support Vector Machine, and Decision Tree. Two software that supports sentiment analysis of this research is MATLAB and WEKA. The Decision Tree technique is provided the highest result of 88% accuracy in WEKA and 86% MATLAB [12].

2.2.5 Performance Analysis of Classification Algorithms under Different Datasets

This research has analyzed the performance of data classification in terms of different datasets. The various classifier has selected and performed this research of eight techniques. The four different datasets can be separated of Diabetes, Nutrition, E-coli, and Mushroom. The result shows Naive Bayes is performed better than another for Diabetes classification. The MLP and IBK are provided well than another for Nutrition, E-coli and Mushroom datasets [13].

2.2.6 Comparative Analysis of Classification Algorithms on Three Different Datasets using WEKA

This research has compared the performance of classification algorithms. The three different datasets have collected and can be defined of Diabetes datasets, Spam base datasets and Credit Approval datasets. The researchers have applied two states of parameters set in WEKA tool by default and customization setting. Three machine learning algorithms have selected to conduct the research of Naive Bayes, Random Forest, and K algorithm. The result shows the custom state is provided the high performance and the highest performance is 94.89% accuracy of Random Forest technique in Spam base classification [14].

2.2.7 Multiclass Mood Classification on Twitter Using Lexicon Dictionary and Machine Learning Algorithms

This research has studied the multi-class of mood on Stanford University official site. There are selected three classifiers for conducting the research and separated into SVM, NB, and KNN techniques. The multi-mood classified based on AFFIN lexicon. The result shows SVM technique id provided the higher accuracy than NB and KNN. The highest of accuracy result is 82% [15].

2.2.8 Performance Analysis of Supervised Machine Learning Algorithms for Text Classification

This research has analyzed the performance of supervised machine learning algorithms. Seven machine learning techniques have selected for text classification on three datasets of Reuter corpus, Brown corpus, and Movie-Review corpus. Subsequently, the performance calculation through Precision, Recall, and F-Measure and used the cross-validation for validating the classification model. The highest accuracy of text classification is ANN in-term of Back propagation Network, which provided greater than 89.0% accuracy [16].

2.2.9 Performance Evaluation of Classification Algorithms by Excluding the Most Relevant Attributes for Dipper/Non-Dipper Pattern Estimation in Type-2 DM Patients

This research has performed the performance analysis of Diabetes classification through classifier in WEKA. There are proposed to skip relevant attribute for faster the diagnosis. The training set and testing set are separated into 66% and 34% respectively. The CV and split methods have used to validate the classifier model. The various evaluations have conducted to performance assessment and can be divided into four methods of Accuracy, Sensitivity, Specificity, and ROC. The finally, ANN techniques (MLP, RBF) is provided the higher performance than another technique. Which mostly 80% is of scores [17].

2.2.10 Comparative Studies on Breast Cancer Classifications with k-fold Cross Validations using Machine Learning Techniques

This research has studied the breast cancer classification with different k-fold cross-validation and different machine learning techniques. The researchers have specified three k value to validate the classification models. On the other hand, there are selected four machine learning techniques to approach for breast cancer classification of the Decision Tree, Naive Bayes, Neural Network, and Support Vector Machine with 3

kernel functions. Subsequently, performed the performance evaluation of each classifier. Finally, the highest accuracy score is neural network technique with 98% accuracy. But this research has focused on the different k-fold method, which the performance result of each k value cannot be expected to have more accurate when an increase or decrease the k. Because sometimes when increase the k value is more accurate but sometimes is not better than the previous.

2.2.11 Tokenization and Filtering Process in RapidMiner

This research has studied on the RapidMiner application for text processing. The researchers have mentioned to a filtering process, which is more suitable in case of large text datasets. The overall of this research shows how to apply the RapidMiner application and automatically of text processing, which is useful in the future [18].

2.2.12 Glaucoma Detection using Cross Validation Algorithm: A Comparative Evaluation on RapidMiner

This research has analyzed the cross-validation of Glaucoma detection through RapidMiner program and approach by Decision Tree technique. The researchers have compared the accuracy result of cross-validation and split-validation. The results show the cross-validation is provided more accuracy than split method, which is 82.83% and 46.67% respectively [19].

Nonetheless, the summarized of the literature reviews can be declared the research methods and tools and including the result which can be seen in table 2.2 – 2.3 as follows.

Table 2.2 Summarized of Data Classification Research based on Literature Reviews

Summarized of Data Classification Research based on Literature Reviews														
Research Details	Research Review No=>	1	2	3	4	5	6	7	8	9	10	11	12	**Research Proposed**
Dataset Type	Text				/			/	/			/		
	Nominal			/									/	
	Numerical		/											/
	Mixing	/				/	/			/	/			
Machine Learning	Supervised	/	/	/	/		/	/	/	/	/		/	/
	Unsupervised													
Validation	SV					/							/	
	KCV	/					/		/	/	/		/	/
Performance Evaluation	Accuracy	/		/	/	/	/	/	/	/	/		/	/
	Recall				/			/	/	/	/			/
	Precision				/			/	/	/	/			/
	F-Measure				/			/	/	/				/
	RMSE		/			/								/
	Lead Time													/
Tool	MATLAB				/						/			
	WEKA				/	/	/			/				
	RapidMiner											/	/	/
	R Language			/										
	Other	/	/					/	/					

Mostly of research reviews are related to data classification in the different objectives and all are selected the supervised learning algorithm to perform the research. Furthermore, the several studies on the performance evaluation are motivating to finding the two side compare between accuracy rate and error rate in each classifier. The various research frameworks are interesting and can refer to some methods or techniques to

encourage this research to reach the aim. However, this research is focusing to evaluate the performance of each classifier via a numerical dataset with a binary problem. In addition, this research is performed cross-validation and performance evaluation through 4 factors of accuracy, F-measure, RMSE, and time. The expected of data classification performance in this research should be greater than 90% of accuracy after reviewing the previous work results as can be seen in table 2.3.

Table 2.3 Summarized of Classification Result Based on Literature Reviews

The Performance Result of Literature Review Summary							
Literature	Study	Dataset source	Dataset type		Instance	Class	% Acc
2.3.1	k fold	UCI data repository	Numerical	Structural	Multi	Multi	NA
2.3.3	Performance	University of India	Numerical	Structural	44	2	97.7%
2.3.4	Performance	Twitter	Text	Unstructured	5,500	2	88.0%
2.3.5	Performance	UCI repository	Numerical	Structural	768	2	82.0%
2.3.6	Performance	UCI repository	Numerical	Structural	300	2	82.6%
2.3.6	Performance	UCI repository	Mixing	Structural	690	2	89.9%
2.3.6	Performance	UCI repository	Text	Unstructured	4,061	2	94.9%
2.3.7	Performance	Stanford's University	Text	Unstructured	8,000	2	82.0%
2.3.8	Performance	Reuter/Brown/Movie	Text	Unstructured	Multi	2	89.0%
Research Proposed							
New research	Performance	UCI repository	Numerical	Structural	>500	2	>90 %

Chapter 3
Methodology

3.1 Research Methodology

This chapter describes the research methodology which defines the research processing to complete the objective. The processing steps can be seen in Figure 3.1.

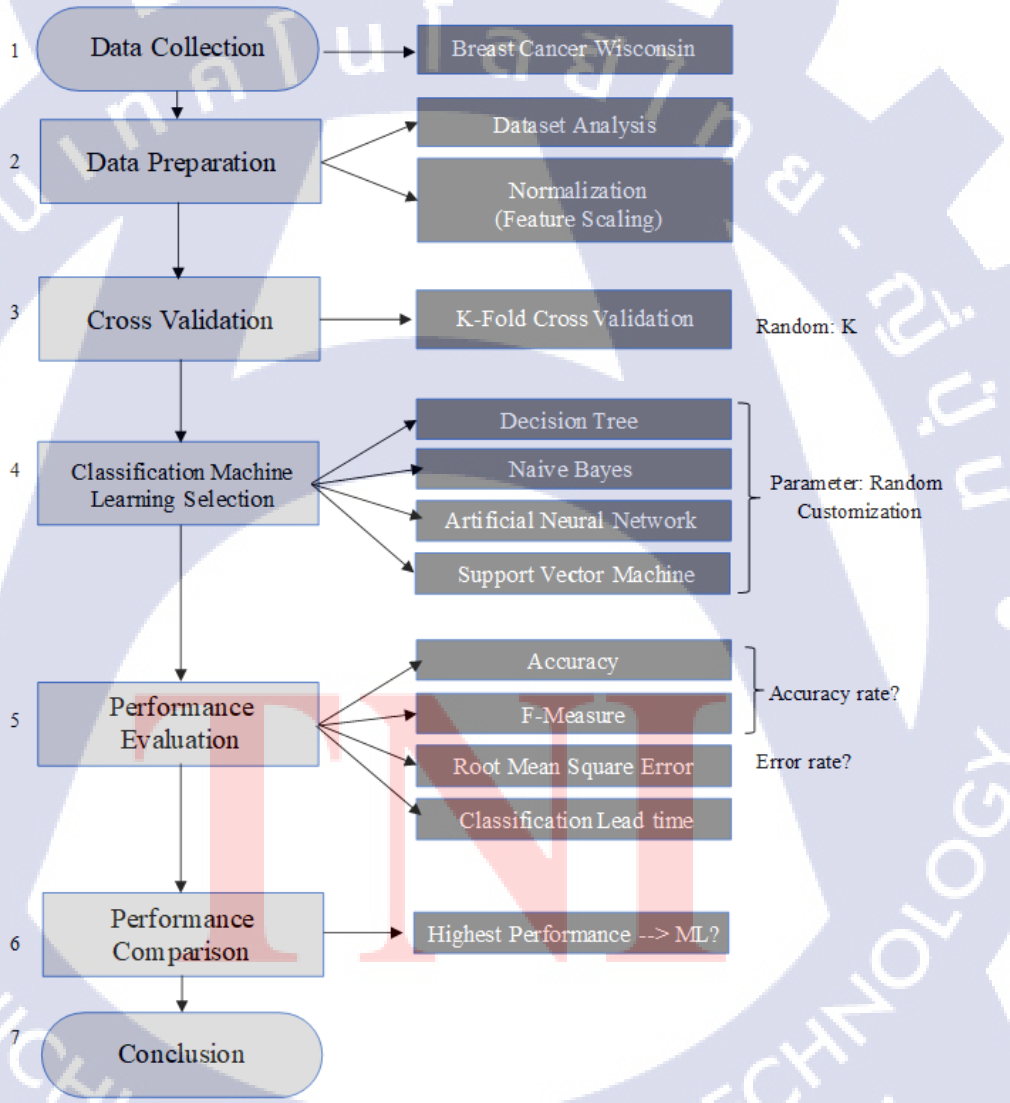


Figure 3.1 The Processing Flow of Research Methodology

3.2 Data Collection

This research has collected the dataset from UCI dataset repository which is the website that provides the sample of the various dataset for conduct the data analytics. The dataset that collected is the breast cancer Wisconsin dataset which is a real-world dataset to conduct the data classification method. The details of this dataset are existing the 30 features computed from a digitized image of a fine needle aspiration (FNA) of a breast mass which consist of FNA of three breast mass and separated into 10 features of each cell nucleus. It can be seen in Figure 3.2. The instance number of this dataset is 569 instances. Furthermore, the attribute type of this dataset is numerical of ten real-valued which decide a binary problem of Benign and Malignant as can be seen in Figure 3.3. Moreover, the ten real-valued features that are computed for each of three different cell nucleuses are the following.

a) Radius: Mean of distances from center to points on perimeter

b) Texture: The standard deviation of gray-scale values

c) Perimeter:

d) Area

e) Smoothness: local variation in radius lengths

f) Compactness: perimeter^2/area - 1.0

g) Concavity: severity of concave portions of contour

h) Concave points: number of concave portions of the contour

i) Symmetry

j) Fractal dimension: "coastline approximation" – 1

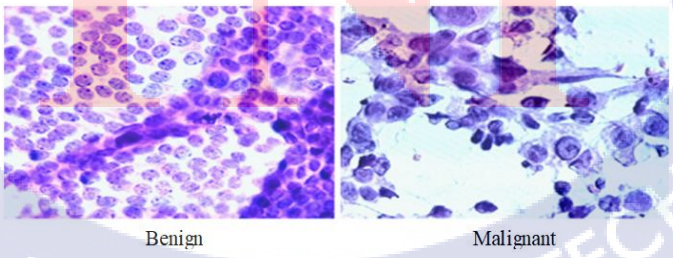


Figure 3.2 The Sample of Fine Needle Aspiration of Breast Mass

Dataset Type	Classification
Origin	Real world
Instances	569
Features/Attriutes	30
Classes	2

Attribute	Cell Nucleus 1		Cell Nucleus 2		Cell Nucleus 3	
	Domain		Domain		Domain	
radius	6.9810	28.1100	0.1140	2.8730	7.9300	36.0400
texture	9.7100	39.2800	0.3600	4.8850	12.0200	49.5400
perimeter	43.7900	188.5000	0.7710	21.9800	50.4100	251.2000
area	143.5000	2501.0000	6.8020	542.2000	185.2000	4254.0000
smoothness	0.0530	0.1630	0.0030	0.0310	0.0710	0.2230
compactness	0.0190	0.3450	0.0020	0.1350	0.0270	1.0580
concavity	0.0000	0.4270	0.0000	0.3960	0.0000	1.2520
concave points	0.0000	0.2010	0.0000	0.0530	0.0000	0.2910
symmetry	0.1060	0.3040	0.0080	0.0790	0.1560	0.6640
fractal dimension	0.0500	0.0970	0.0010	0.0300	0.0550	0.2080
Classes	**Benign**	**Malignant**				

Figure 3.3 Brest Cancer Wisconsin Dataset Descriptions

3.3 Data Preparation

The data preparation process can be divided into 2 steps of dataset analysis and normalization. This step is using the Excel and Minitab program to performing the analyzing and normalization.

3.3.1 Dataset analysis

The dataset analysis has proposed the step that would analyze the structure and cleansing of a dataset which used in this research.

3.3.2 Normalization

This step is conducted to transform the dataset which used to evaluate the performance become to the same scale.

3.4 Cross Validation

The cross-validation is the method that uses to evaluate the performance of data classification in this research. However, this method has randomly searched for the suitable k-value to perform the research.

3.5 Classification Machine Learning Selection

This method has selected the four machine learning techniques which supervised learning type to compare the performance result. The machine learning algorithm selection can be separated into 4 techniques of the Decision tree, Naïve Bayes, Artificial neural network, and Support vector machine. Nonetheless, the parameter of these techniques is randomly from the default.

3.6 Performance Evaluation

The performance evaluation of this research is conducted by an accuracy rate, error rate, and classification lead time which can be divided into four results of accuracy, F-measure, RMSE, and lead time.

3.7 Performance Comparison

The objective of this research is proposed to compare the classification results between four machine learning techniques which focus on the accuracy rate and error rate include with classification lead time.

3.8 Conclusion

The conclusion part of this research would be summarized the research results and discussion on each of some problem or a good point and weak point of each technique with including the opportunity in the future work.

Chapter 4
Simulation and Experimental Result

This chapter has revealed the research simulation and experimental result as follows.

4.1 Dataset Analysis

4.1.2 Dataset

The dataset consists of 30 attributes of three breast cell nucleus which can be divided into 10 attributes per one cell nucleus. Furthermore, this dataset contains 569 instances of the history from the real-world breast cancer diagnosed. Nonetheless, the classification problem of this dataset is the binary problem which predicts to 2 classes of malignant and benign. The ratio of classes that provides in this dataset is the imbalance type by declared of the benign class is 63% which consist of 357 labels and the ratio of the malignant class is 37% which consist of 212 labels of the total class. It can be seen in Figure 4.1.

4.1.3 Attributes

This dataset has provided 30 attributes which is a relational attribute type. Furthermore, the attribute declares the real number value of each feature of the breast cell nucleus. Subsequently, the analysis result of the value scale of each attribute found that the different scale of each attribute. It can be seen in Figure 4.2 - 4.4. Nonetheless, the attributes are cleansing by no missing data.

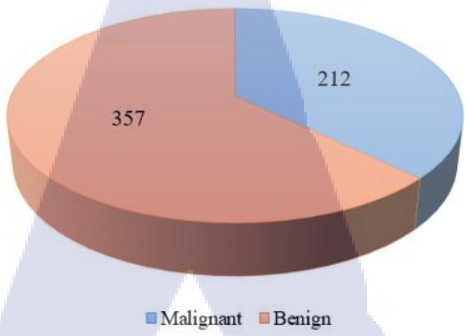


Figure 4.1 The Proportion of Class in Breast Cancer Dataset

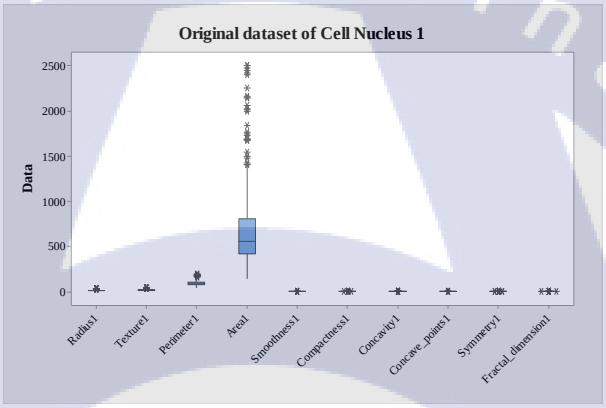


Figure 4.2 The Box Plot of Original Attribute of Cell Nucleus 1

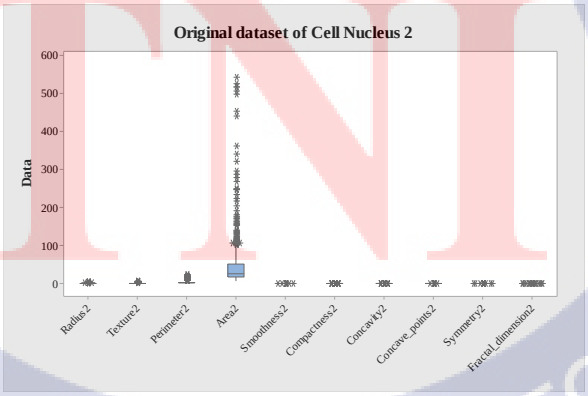


Figure 4.3 The Box Plot of Original Attribute of Cell Nucleus 2

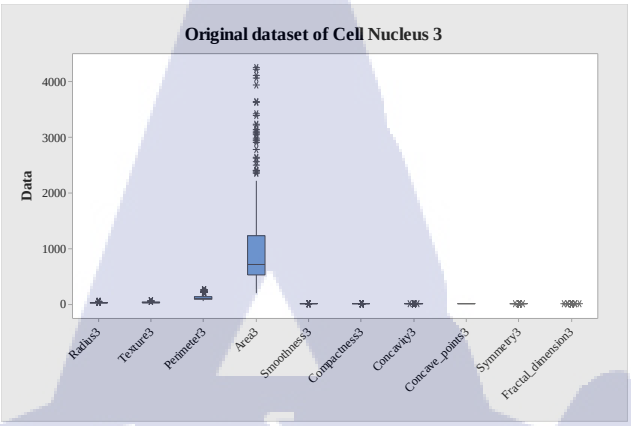


Figure 4.4 The Box Plot of Original Attribute of Cell Nucleus

Attribute	Cell Nucleus 1		Cell Nucleus 2		Cell Nucleus 3		Total	
	Domain		Domain		Domain		Min	Max
radius	6.9810	28.1100	0.1140	2.8730	7.9300	36.0400	0.1140	36.0400
texture	9.7100	39.2800	0.3600	4.8850	12.0200	49.5400	0.3600	49.5400
perimeter	43.7900	188.5000	0.7710	21.9800	50.4100	251.2000	0.7710	251.2000
area	143.5000	2501.0000	6.8020	542.2000	185.2000	4254.0000	6.8020	4254.0000
smoothness	0.0530	0.1630	0.0030	0.0310	0.0710	0.2230	0.0030	0.2230
compactness	0.0190	0.3450	0.0020	0.1350	0.0270	1.0580	0.0020	1.0580
concavity	0.0000	0.4270	0.0000	0.3960	0.0000	1.2520	0.0000	1.2520
concave points	0.0000	0.2010	0.0000	0.0530	0.0000	0.2910	0.0000	0.2910
symmetry	0.1060	0.3040	0.0080	0.0790	0.1560	0.6640	0.0080	0.6640
fractal dimension	0.0500	0.0970	0.0010	0.0300	0.0550	0.2080	0.0010	0.2080

Figure 4.5 The Summary of Min-Max Scale of Attributes.

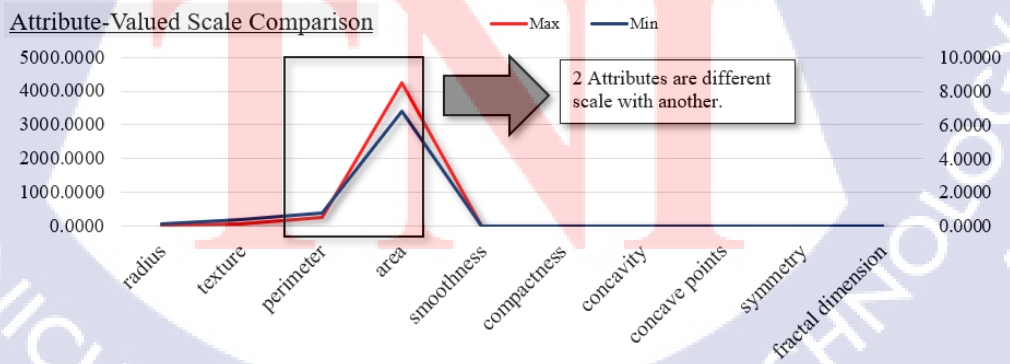


Figure 4.6 Attributes-Value Scale Comparison graph

The summary of attribute value about the min-max length of three cell nucleus has shown 2 features that are different scale with another as can be seen in Figure 4.5-4.6.

Based on the dataset analysis result, the attribute value of each cell nucleus is mostly different scale. Thus, this research has conducted the normalization of the dataset before training to each machine learning technique.

4.2 Normalization

This research has conducted the normalization of the dataset by using the min-max method which transforms each attribute-valued become to the same scale between [0-1]. The min-max normalize method can calculate from $xi_{new} = (xi - x_{min})/(x_{max}-x_{min})$ which can be seen the new attribute-valued after normalization as Figure 4.7 - 4.10.

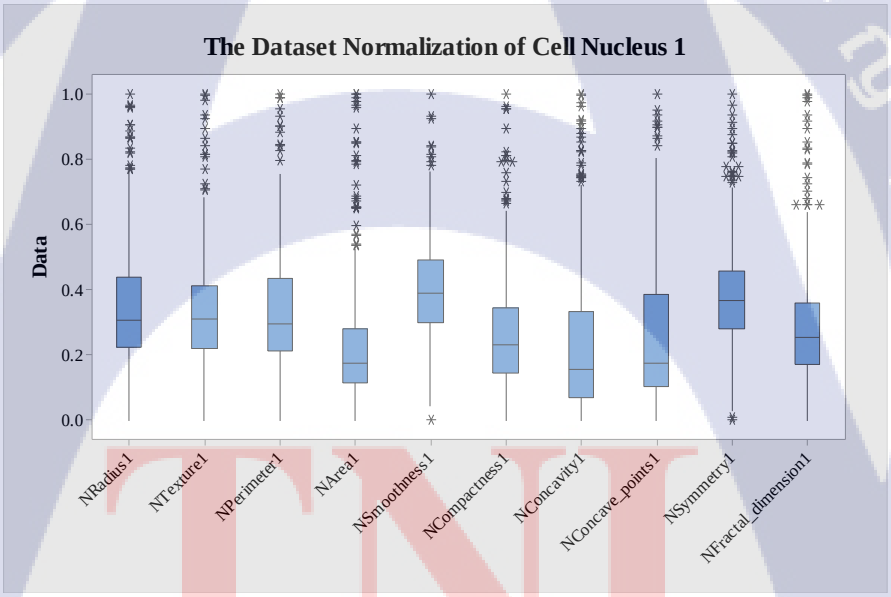


Figure 4.7 Normalization of Cell Nucleus 1

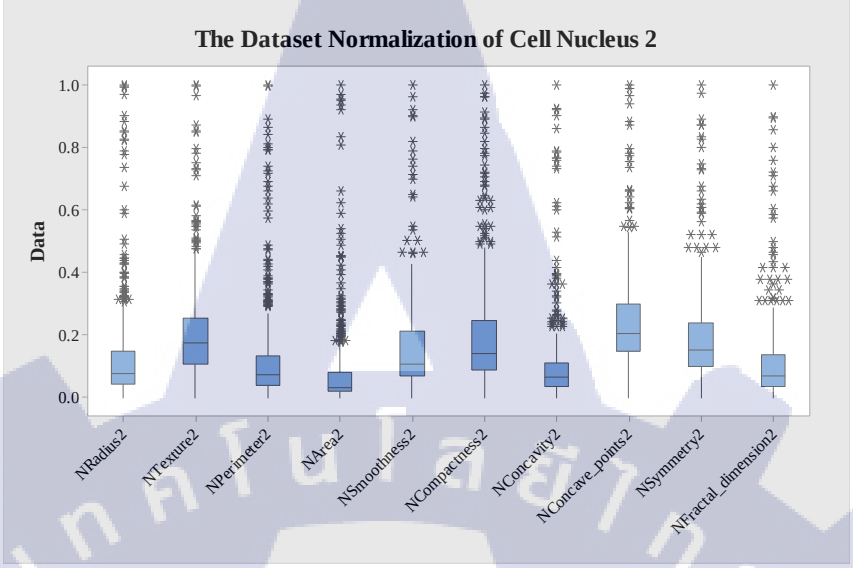


Figure 4.8 Normalization of Cell Nucleus 2

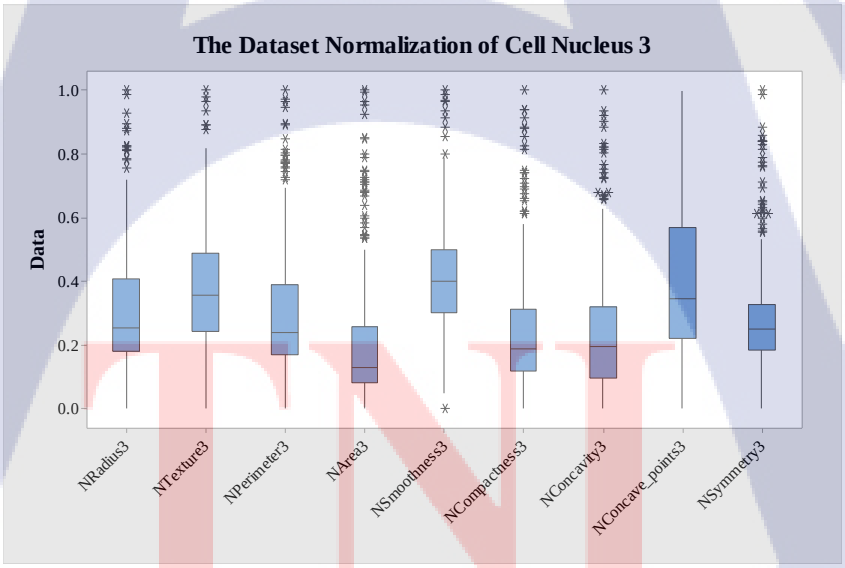


Figure 4.9 Normalization of Cell Nucleus 3

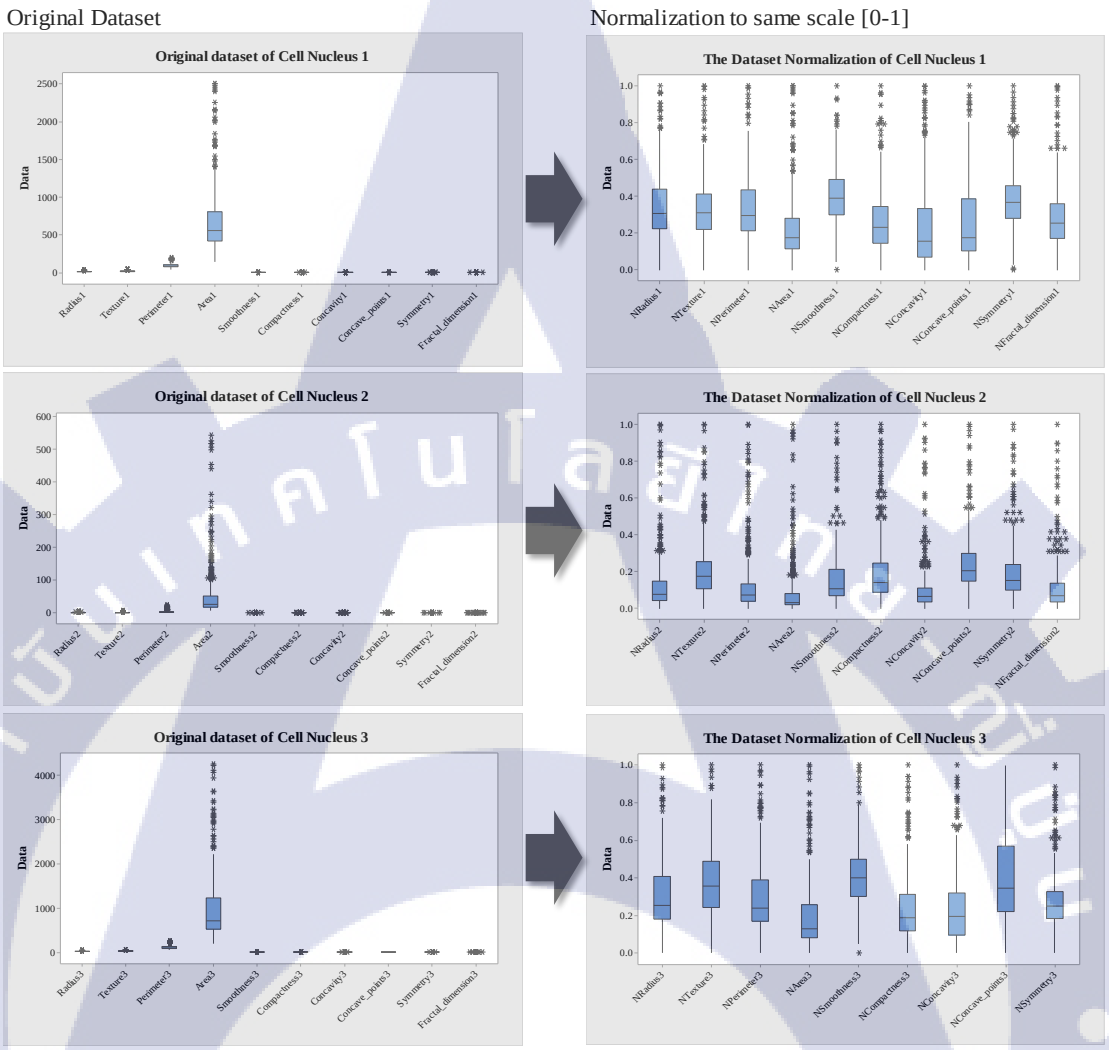


Figure 4.10 Original Dataset Value VS Normalization into (0-1) Scale

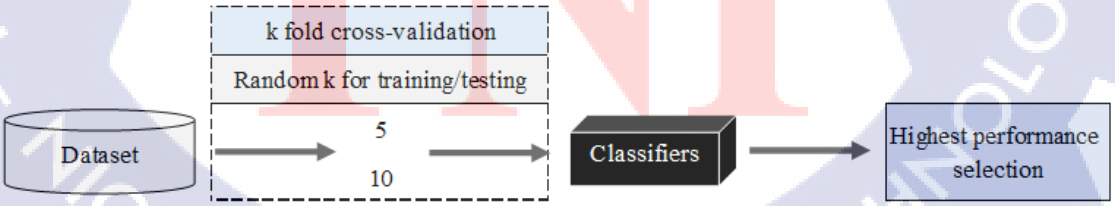


Figure 4.11 Random of The k-value of k fold Cross-Validation.

4.3 Cross Validation

This research has performed the k fold cross-validation technique to evaluate the performance of each machine learning algorithm. The k-values of this method is decided by randomly to conduct the validation process. The k-values will assign the number of the group that will perform to training and testing time. Subsequently, this method has used the stratified sampling method to consider for selecting the instance into each group with the same ratio. Nonetheless, the k-value of this method will random customizing which divided into two values of 5, and 10. It can be seen in Figure 4.11. The k fold cross-validation method has been performed in the RapidMiner Studio 7.4 program to evaluate the classification performance.

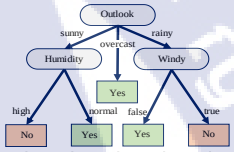
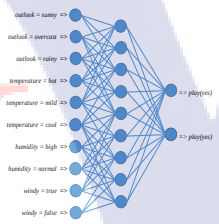
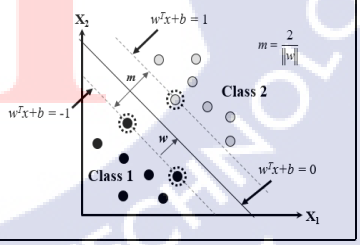
Machine Learning	Learning Type	Classification Mehtod	Model
Decision Tree (DT)	Supervised	Statistics + If-Then rule	
Naïve Bayes (NB)	Supervised	Probabilistic/Bayes theorem	$\text{Posterior} = \frac{\text{Likelihood* Prior}}{\text{Evidence}}$ $P(C_j \mid A_1, A_2, \ldots, A_n) = \frac{\left(\prod_{i=1}^{n} P(A_i \mid C_j)\right) P(C_j)}{P(A_1, A_2, \ldots, A_n)}$
Artificial Neural Network (ANN)	Supervised	Simulated the neuron system of a human	
Support Vector Machine (SVM)	Supervised	The decision of Suitable distance in feature space by vector	

Figure 4.12 Summary of the Machine Learning Technique Selection

4.4 Classification Machine Learning Selection

This research has selected the four machine learning techniques which are the supervised learning method and different how computation. A decision tree is an algorithm that represented the tree model which decision-based on historical data statistics with If-then rule. Naive Bayes is the technique that represented the decision method on the probabilistic which computed on historical data. Artificial Neural Network is the algorithm that represented the computation step which simulated the neuron of the human. Support Vector Machine is the technique that represented the classification algorithm based on the suitable distance in feature space via a vector. However, the machine learning processing has also been conducted the data classification in the RapidMiner studio 7.4 program which can be seen the summary of machine learning technique selection in Figure 4.12.

The range of parameter adjustment has been considered from the pre-test of the parameter value which impacts the data classification result change. Moreover, the function adjustment has been selected by considering the function that popularly used for data classification.

4.4.1 Decision tree (DT)

The decision tree algorithm has conducted the classification by customizing the parameter setting which focusing on 3 parameters of maximal depth, pruning, and pre-pruning.

4.4.1.1 Maximal depth

The depth of a tree varies depending upon the size and nature of the example set. This parameter is used to restrict the size of the Decision Tree. The tree generation process is not continued when the tree depth is equal to the maximal depth. If its value is set to -1, the maximal depth parameter puts no bound on the depth of the tree, a tree of maximum depth is generated. If its value is set to 1, a Tree with a single node is generated.

4.4.1.2 Pruning

Normally, the Decision Tree is generated with pruning. Setting this parameter to false disables the pruning and delivers an unpruned Tree. Thus, this parameter

Decision tree Parameter		Random	
		Default	Customization
Maximum Depth		20	±5
Pruning: Confidence Level		25%	±5%
Pre-pruning	Minimal Gain	10%	±5%
	Minimal Leaf Size	2	3,4

Figure 4.13 Decision Tree Parameter Customization

is must setting to prevent the noise in this algorithm. This parameter specifies the confidence level used for the pessimistic error calculation of pruning.

4.4.1.3 Pre-pruning

Usually, the Decision Tree is generated with pre-pruning. Setting this parameter to false disables the pre-pruning and delivers a tree without any pre-pruning. Thus, this parameter setting can be separate into 2 factors of minimal gain and minimal leaf size. The minimal gain is the gain of a node is calculated before splitting it. The node is split if its gain is greater than the minimal gain. A higher value of minimal gain results in fewer splits and thus a smaller tree. A too high value will completely prevent splitting and a tree with a single node is generated. The minimal leaf size is the size of a leaf node is the number of examples in its subset. The tree is generated in such a way that every leaf node subset has at least the minimum leaf size number of instances.

Nonetheless, all parameter of decision tree techniques that focus is random customized from the default which can be seen in Figure 4.13.

4.4.2 Naïve Bayes (NB)

A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be 'independent feature model'. In simple terms, a Naive Bayes classifier assumes that the

presence (or absence) of a particular feature of a class (i.e. attribute) is unrelated to the presence (or absence) of any other feature. The advantage of the Naive Bayes classifier is that it only requires a small amount of training data to estimate the means and variances of the variables necessary for classification. Because independent variables are assumed, only the variances of the variables for each label need to be determined and not the entire covariance matrix. Thus, the parameter setting of this method can be separated into 2 types of using Laplace correction or not using Laplace correction. The Laplace correction is an expert parameter. This parameter indicates if Laplace correction should be used to prevent the high influence of zero probabilities. There is a simple trick to avoid zero probabilities. We can assume that our training set is so large that adding one to each count that we need would only make a negligible difference in the estimated probabilities, yet would avoid the case of zero probability values. This technique is known as Laplace correction.

4.4.3 Artificial Neural Network (ANN)

An artificial neural network (ANN) is a mathematical model or computational model that is inspired by the structure and functional aspects of biological neural networks. A neural network consists of an interconnected group of artificial neurons and it processes information using a connectionist approach to computation. In most cases, an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. Modern neural networks are usually used to model complex relationships between inputs and outputs or to find patterns in data. The parameter customizing of this algorithm is can be divided into 3 categories of hidden layers, training cycles, and Learning rate. An overall parameter that using in RapidMiner Studio program can be described as follows. It can be seen in Figure 4.14.

4.4.3.1 Hidden Layers

This parameter describes the name and the size of all hidden layers which can define the structure of the neural network with this parameter. Each list entry describes a new hidden layer. Each entry requires the name and size of the hidden layer.

The layer name can be chosen arbitrarily. It is only used for displaying the model. Note that the actual number of nodes will be one more than the value specified as hidden layer size because an additional constant node will be added to each layer. This node will not be connected to the preceding layer. If the hidden layer size value is set to -1 the layer size would be calculated from the number of attributes of the input example set. In this case, the layer size will be set to (number of attributes + number of classes) / 2 + 1. If the user does not specify any hidden layers, a default hidden layer with sigmoid type and size equal to (number of attributes + number of classes) / 2 + 1 will be created and added to the net. If only a single layer without nodes is specified, the input nodes are directly connected to the output nodes and no hidden layer will be used.

4.4.3.2 Training cycles

This parameter specifies the number of training cycles used for the neural network training. In back-propagation, the output values are compared with the correct answer to compute the value of some predefined error-function. The error is then fed back through the network. Using this information, the algorithm adjusts the weights of each connection in order to reduce the value of the error function by some small amount. This process is repeated n number of times. n can be specified using this parameter.

4.4.3.3 Learning rate

This parameter determines how much we change the weights at each step. It should not be 0.

4.4.4 Support Vector Machine (SVM)

A support vector machine constructs a hyper plane or set of hyper planes in a high- or infinite- dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyper plane that has the largest distance to the nearest training data points of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier. Whereas the original problem may be stated in a finite dimensional space, it often happens that the sets to discriminate are not linearly separable in that space. For this reason, it was proposed

that the original finite-dimensional space would be mapped into a much higher-dimensional space, presumably making the separation easier in that space. To keep the computational load reasonable, the mapping used by SVM schemes are designed to ensure

ANN Parameter	Randomization	
	Default	Customization
Hidden Layers	1	2,3,4
Training Cycles	500	±100
Learning rate	30%	±5

Figure 4.14 ANN Parameter Random Customization

that dot products may be computed easily in terms of the variables in the original space, by defining them in terms of a kernel function K (x, y) selected to suit the problem. The hyper planes in the higher dimensional space are defined as the set of points whose inner product with a vector in that space is constant. However, this algorithm is randomly customization the parameter in RapidMiner studio program which can be divided into three parameters of kernel type, gamma-value, C-value.

4.4.4.1 Kernel type

The type of the kernel function is selected through this parameter which is provided into 4 functions of Linear function, Polynomial function (Poly), Radial Basis function (RBF), and Sigmoid function. The RBF kernel type is the default value. In general, the RBF kernel is a reasonable first choice. Here are a few guidelines regarding different kernel types.

4.4.4.2 Gamma-value

This parameter is only available when the kernel type parameter is set to Linear, Poly, RBF or sigmoid. This parameter specifies gamma for polynomial, RBF, and sigmoid kernel functions. The value of gamma may play an important role in the SVM

model. Changing the value of gamma may change the accuracy of the resulting SVM model. So, it is a good practice to use cross-validation to find the optimal value of gamma.

SVM Parameter	Randomization			
	Default	Customization		
Kernel type	RBF	Poly	Linear	Sigmoid
Gamma-value	0	0.1, 0.2		
C-value	0	50, 100		

Figure 4.15 SVM Parameter Random Customization

4.4.4.3 C-value

This parameter is only available when the svm type parameter is set to c-SVC, epsilon-SVR or nu-SVR. This parameter specifies the cost parameter C for c-SVC, epsilon-SVR and nu-SVR. C is the penalty parameter of the error term. Furthermore, an overall parameter customizing from the default can be seen in Figure 4.15.

4.5 Performance Evaluation

This research has proposed to evaluate the classification performance which can be divided into four categories of Accuracy, F-measure, Root mean square error, and Classification lead time.

4.5.1 Accuracy

The relative number of correctly classified examples or in other words percentage of correct predictions which can be calculated based on the confusion matrix.

4.5.2 F-measure

The F-measure is defined as the weighted harmonic mean of its precision and recall which is also can be calculated based on the confusion matrix. A recall is the

ratio of a number of events that can correctly recall to number of all correct events. Precision is the ratio of a number of events that can correctly recall to a number all events recall. In other words, it is how precise of the recall. And both also can be calculated based on confusion matrix.

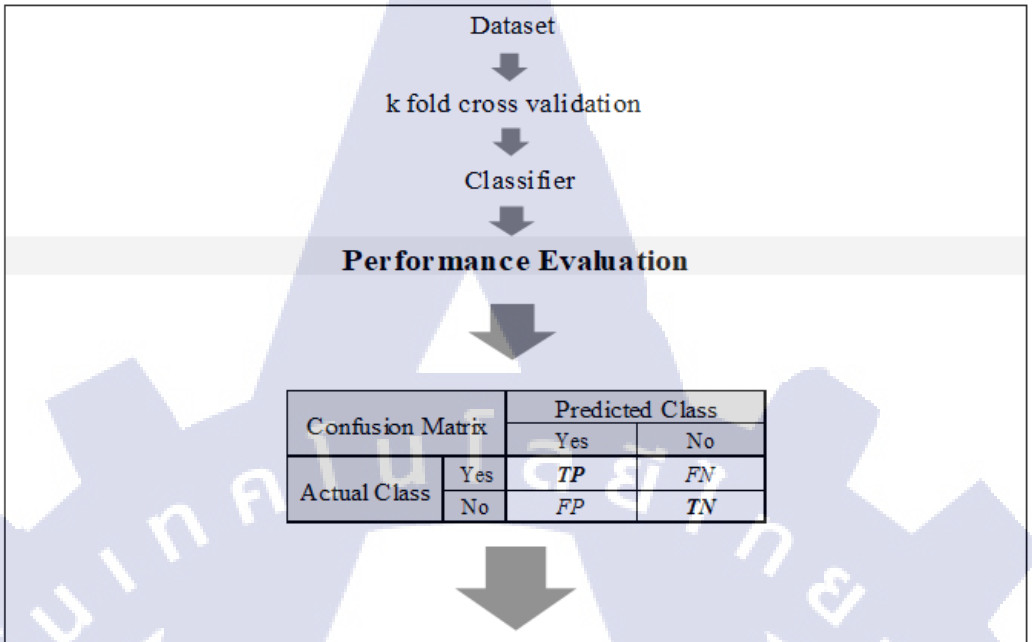
4.5.3 Root Mean Square Error (RMSE)

The RMSE is the standard deviation of the predictions from the ground-truth. Furthermore, this is one way to measure the performance of a classifier which considers the error rate or number of misclassifications.

4.5.4 Classification Lead time (Time)

The lead time is one of the performance evaluation of data classification which is focusing on the processing lead time that each classifier using for computation and prediction. The processing lead time is starting since running the program until stop and shows the data classification result.

Nonetheless, this research has proposed to compare the classification performance of each classifier by considering the highest score of accuracy rate and the lowest score of error rate which conduct the performance evaluation via the RapidMiner studio 7.4 can be It can be seen the summary of performance evaluation in Figure 4.16.



Accuracy	(TP+TN)/(TP+FN+TN+FP)
Precision (Yes)	TP/(TP+FP)
Precision (No)	TN/(FN+TN)
Recall (Yes)	TP/(TP+FN)
Recall (No)	TN/(FP+TN)
F-measure(Yes)	[2 x recall (yes) x precision (yes)] / [recall (yes) + precision (yes)]
F-measure(No)	[2 x recall (no) x precision (no)] / [recall(no) + precision (no)]
RMSE	$\sqrt{(FP+FN)/(TP+TN+FP+FN)}$
Lead Time	Δ Start - Stop

Figure 4.16 Summary of Performance Evaluation Method

4.6 Data Classification via RapidMiner Studio 7.4

The classification process of this research has been performed in RapidMiner studio 7.4 program which can be seen the programming diagram in Figure 4.17.

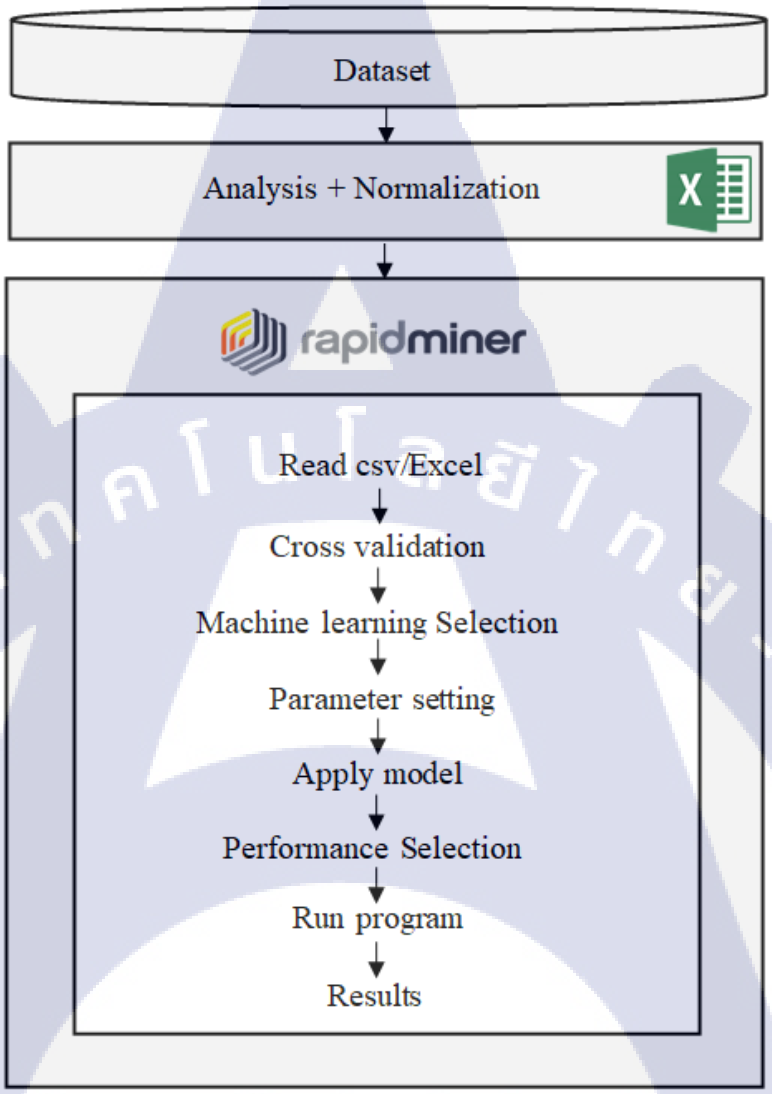


Figure 4.17 Process diagram of Data Classification in RapidMiner studio 7.4 Program

4.6.1 Read csv/excel

The starting of data classification in RapidMiner studio program retrieves the dataset which prepared in excel file or csv file type. The dataset that prepared in excel file will be passed the preparation process which arrange the data of attribute in each column of excel and starting from first attribute to the labeled or the result of classes. It can be seen in Figure 41. First, performed to select the read csv/excel in operators' module as

step number1-2. Second, click the Read csv operators to process space as step number3 which these processes can be seen in Figure 4.18 – 4.20.

NRadius1	NTexture1	NPerimeter1	NArea1	NSmoothness1	NCompactness1	NConcavity1	NConcave_points1	NSymmetry1	NFractal_dimension1	Result
0.21009	0.36084	0.72459	0.10291	0.80909	0.81288	0.5644	0.52239	0.77778	1	M
0.25884	0.20257	0.26798	0.14151	0.68182	0.46319	0.37002	0.40299	0.5202	0.55319	M
0.67107	0.45079	0.6455	0.53468	0.37273	0.2546	0.25761	0.09733	0.35859	0.06383	M
0.37905	0.52384	0.35761	0.23555	0.37273	0.09816	0.05621	0.14428	0.25253	0.10638	M
0.57215	0.51065	0.584	0.40742	0.34545	0.6135	0.49415	0.49751	0.63131	0.2766	M
0.29812	0.35746	0.30081	0.17039	0.46364	0.38344	0.22951	0.30846	0.4596	0.17343	M
0.56601	0.4092	0.28768	0.16458	0.4	0.2638	0.19438	0.26368	0.34848	0.25532	M
0.3658	0.40176	0.35768	0.21777	0.55455	0.3589	0.89513	0.40299	0.4899	0.38298	M
0.07653	0.38113	0.07512	0.03321	0.64545	0.21779	0.07026	0.04478	0.38889	0.38298	B
0.26972	0.47785	0.26881	0.1508	0.53636	0.32822	0.26464	0.34328	0.42424	0.34043	M
0.11756	0.38214	0.11278	0.0534	0.46364	0.07048	0.05152	0.07463	0.33333	0.40426	B
0.27446	0.22929	0.25969	0.15228	0.40909	0.10123	0.08665	0.14428	0.26768	0.14894	B
0.48365	0.66901	0.48656	0.33336	0.49091	0.50307	0.39578	0.39303	0.43939	0.31915	M
0.25221	0.2303	0.24463	0.13888	0.35455	0.15031	0.07963	0.11443	0.33333	0.19149	B
0.52388	0.21779	0.52802	0.36585	0.48182	0.60123	0.39344	0.53731	0.55051	0.3617	M
0.39704	0.44133	0.38981	0.24802	0.35455	0.26074	0.2623	0.37313	0.33333	0.23404	M
0.1328	0.34596	0.1266	0.06405	0.44545	0.10429	0.01639	0.0398	0.14646	0.40426	B
0.31374	0.47555	0.30585	0.1863	0.38182	0.20245	0.20141	0.22388	0.27778	0.19149	M
0.24606	0.36557	0.23101	0.1337	0.24545	0.06442	0.05621	0.08955	0.34343	0.14894	B
0.29008	0.19784	0.30005	0.16403	0.79091	0.4816	0.48478	0.47761	0.43434	0.57447	M
0.81731	0.35475	0.8459	0.68611	0.83636	0.82209	1	1	0.80808	0.40426	M
0.30238	0.2259	0.29238	0.17391	0.16364	0.16564	0.1897	0.1393	0.18182	0.17021	B

Figure 4.18 The Sample of Dataset Preparation in Excel File

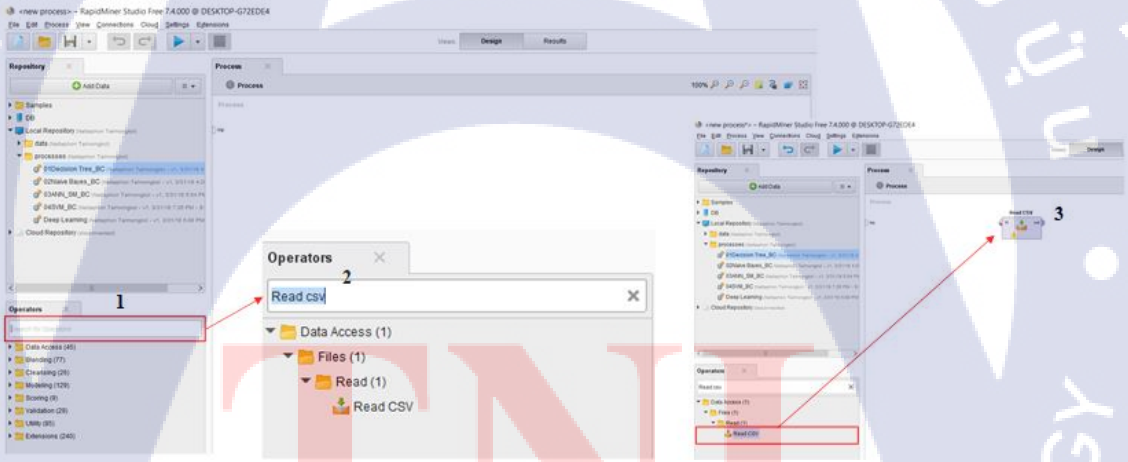
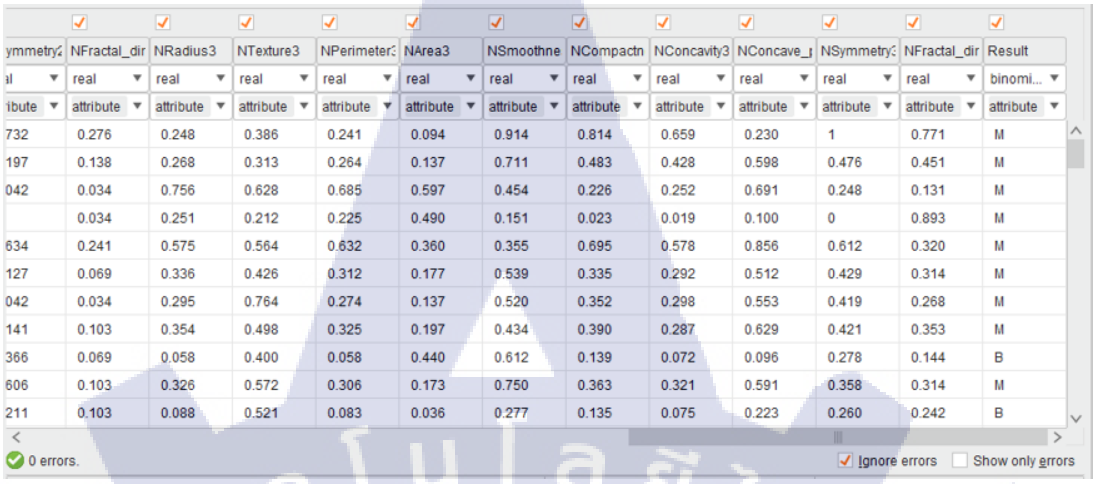


Figure 4.19 Read Excel/CSV File Process-1



ymmetry2	NFractal_dir	NRadius3	NTexture3	NPerimeter3	NArea3	NSmoothne	NCompactn	NConcavity3	NConcave_1	NSymmetry3	NFractal_dir	Result
al	real	real	real	real	real	real	real	real	real	real	real	binomi...
ibute	attribute	attribute	attribute	attribute	attribute	attribute	attribute	attribute	attribute	attribute	attribute	attribute
732	0.276	0.248	0.386	0.241	0.094	0.914	0.814	0.659	0.230	1	0.771	M
197	0.138	0.268	0.313	0.264	0.137	0.711	0.483	0.428	0.598	0.476	0.451	M
042	0.034	0.756	0.628	0.685	0.597	0.454	0.226	0.252	0.691	0.248	0.131	M
	0.034	0.251	0.212	0.225	0.490	0.151	0.023	0.019	0.100	0	0.893	M
634	0.241	0.575	0.564	0.632	0.360	0.355	0.695	0.578	0.856	0.612	0.320	M
127	0.069	0.336	0.426	0.312	0.177	0.539	0.335	0.292	0.512	0.429	0.314	M
042	0.034	0.295	0.764	0.274	0.137	0.520	0.352	0.298	0.553	0.419	0.268	M
141	0.103	0.354	0.498	0.325	0.197	0.434	0.390	0.287	0.629	0.421	0.353	M
366	0.069	0.058	0.400	0.058	0.440	0.612	0.139	0.072	0.096	0.278	0.144	B
606	0.103	0.326	0.572	0.306	0.173	0.750	0.363	0.321	0.591	0.358	0.314	M
211	0.103	0.088	0.521	0.083	0.036	0.277	0.135	0.075	0.223	0.260	0.242	B

0 errors. Ignore errors Show only errors

Figure 4.20 Read Excel/CSV File Process-2

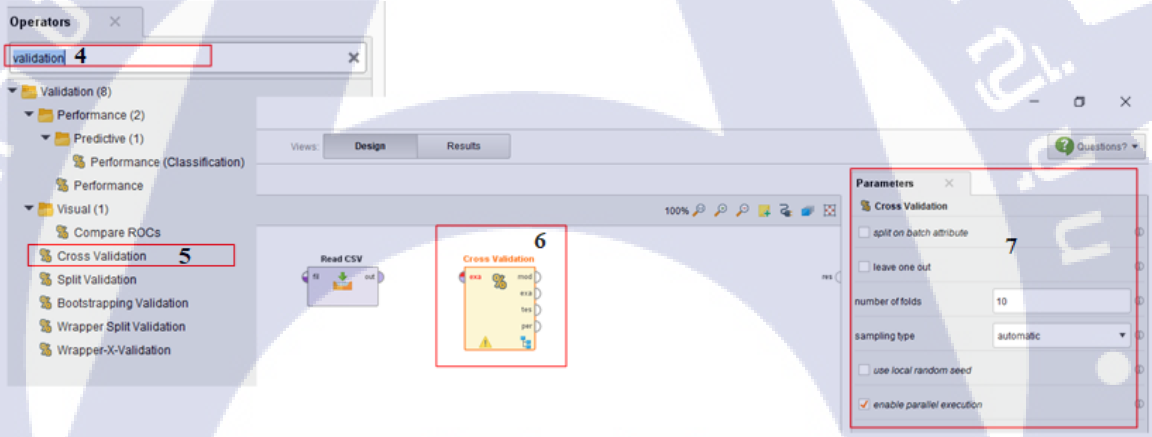


Figure 4.21 The Cross-Validation Process

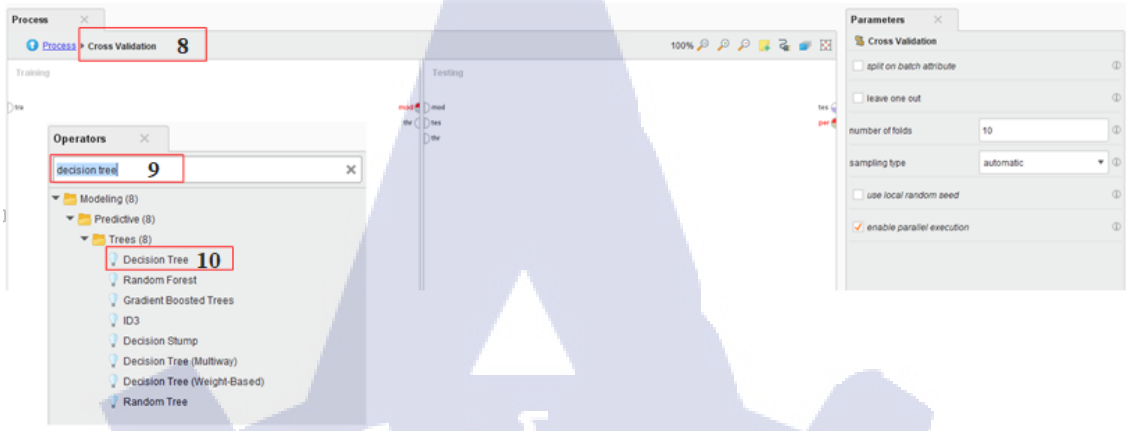


Figure 4.22 Machine Learning Selection Process

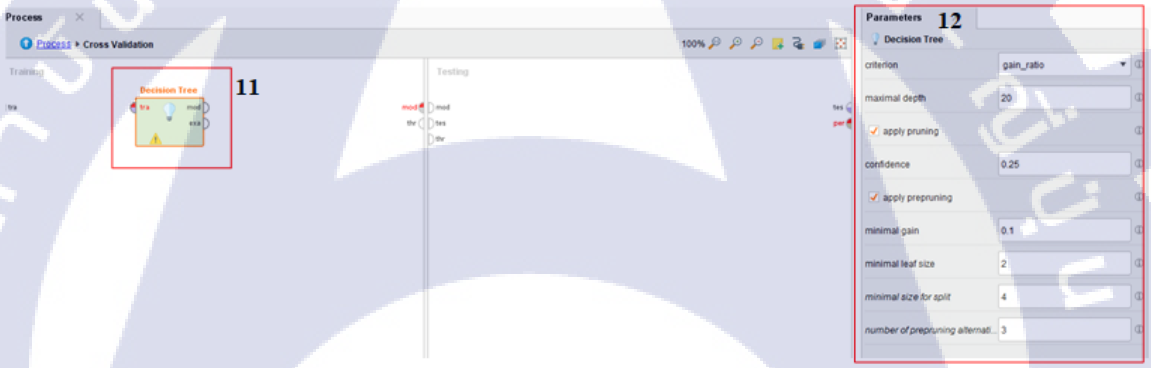


Figure 4.23 Parameter Setting Process

4.6.2 Cross validation

The cross-validation process in RapidMiner program is conducted by select the validation operator in operator's module, thus drag them to process space module. it can be seen in number 4-6 and after that, a user can adjust the cross-validation parameter which shown on a number 7. It can be seen in Figure 4.21.

4.6.3 Machine Learning Selection

Based on the cross-validation process that completed, the user can be done of the machine learning selection process by double click for cross-validation block in number 6 of Figure 4.21 and after that select the machine learning operator in operator's module and drag them to cross-validation space that shown in number 8-11. It can be seen in Figure 4.22 – 4.23.

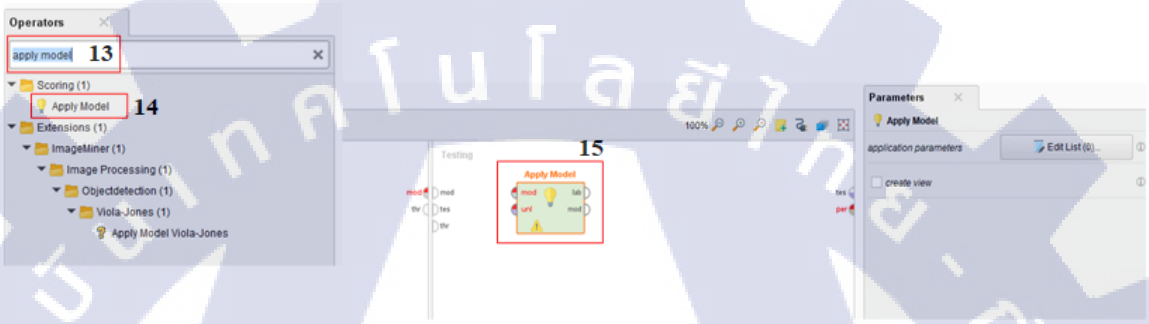


Figure 4.24 Apply Model Process

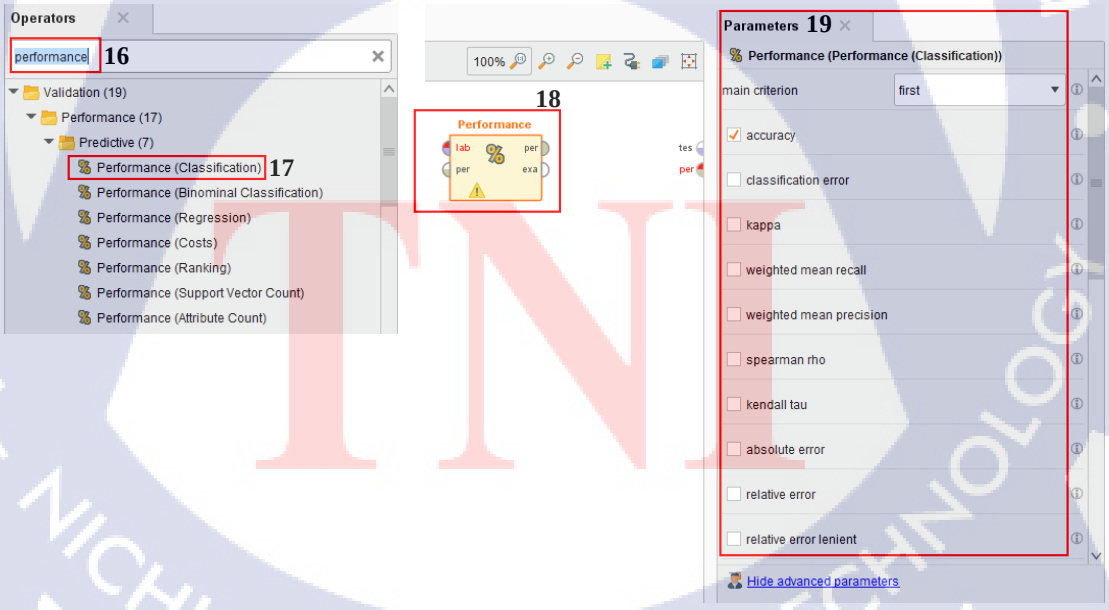


Figure 4.25 Performance Selection Process

4.6.4 Parameter setting

The parameter setting can be performed by using the parameter module as can be seen in number 12 which the parameter item will be different in each machine learning algorithm. It can be seen in Figure 4.23.

4.6.5 Apply model

After completing the machine learning selection and parameters setting, the user can apply the model by select the apply operator from the operator's module and drop on the space as can be seen in a number 13-15. It can be seen in Figure 4.24.

4.6.6 Performance Selection

The performance selection process can be done by select the performance operator and drag to space which shown on the number 16-18. After that, the user can select the performances which are the objectives of the research as can be seen in number 19. It can be seen in Figure 4.25.

4.6.7 Run program

Based on the operator put on each part of the RapidMiner program which can be seen since number 1-19. The use must be put the connection line between operators which can be seen in number 20-22. After that, the user can run the program to getting the classification result as can be seen in number 23 of Figure 4.26.

4.6.8 Results

The result of each performance proposed can be shown after the program running completed which user should be press the result's module that can be seen in a number 24. Subsequently, the performance results are contained in the performance vector of number 25. The confusion matrix is shown in number 26 and the model simulation is shown in a number 27 which can be seen in Figure 4.27.

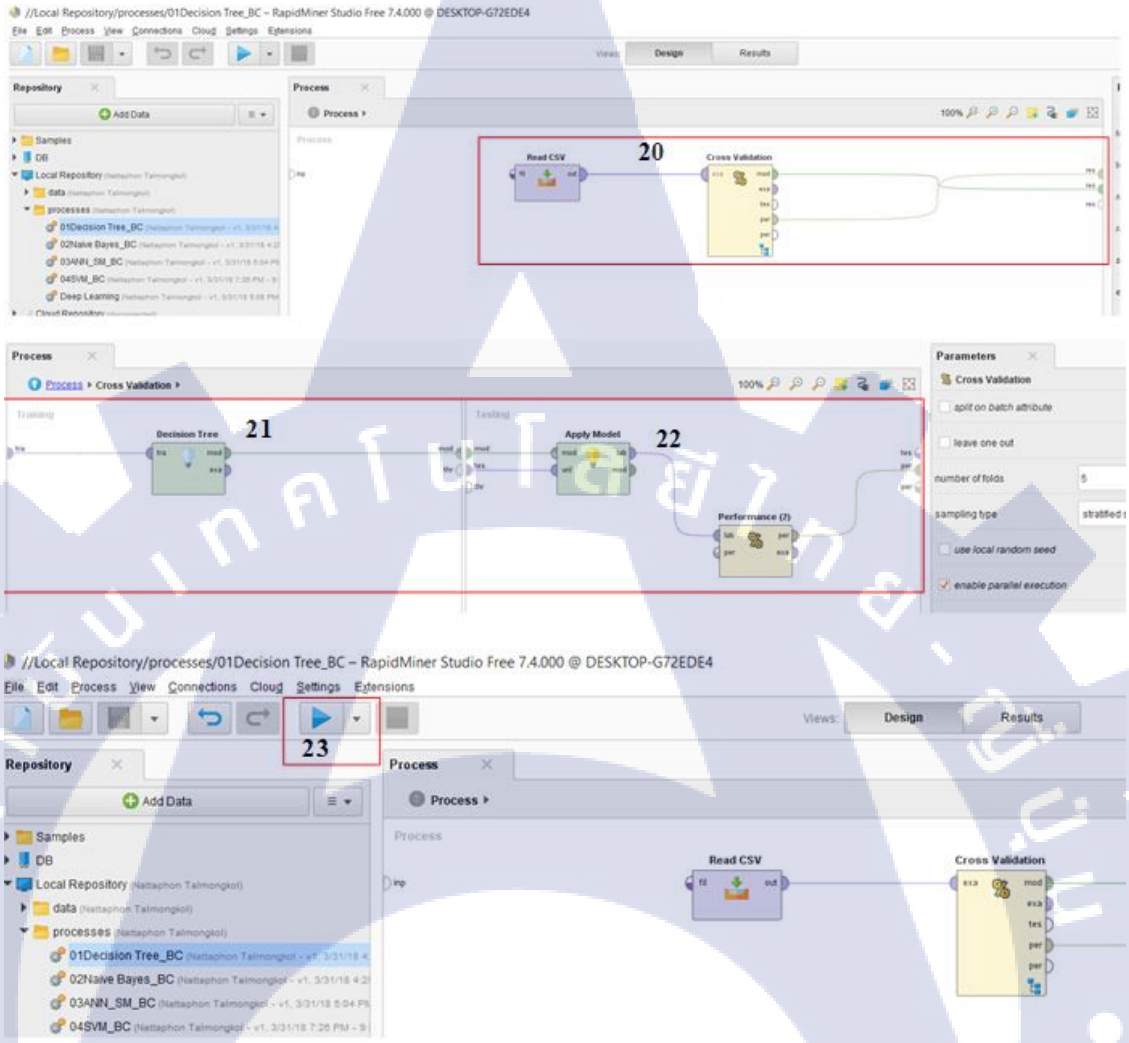


Figure 4.26 Run Program

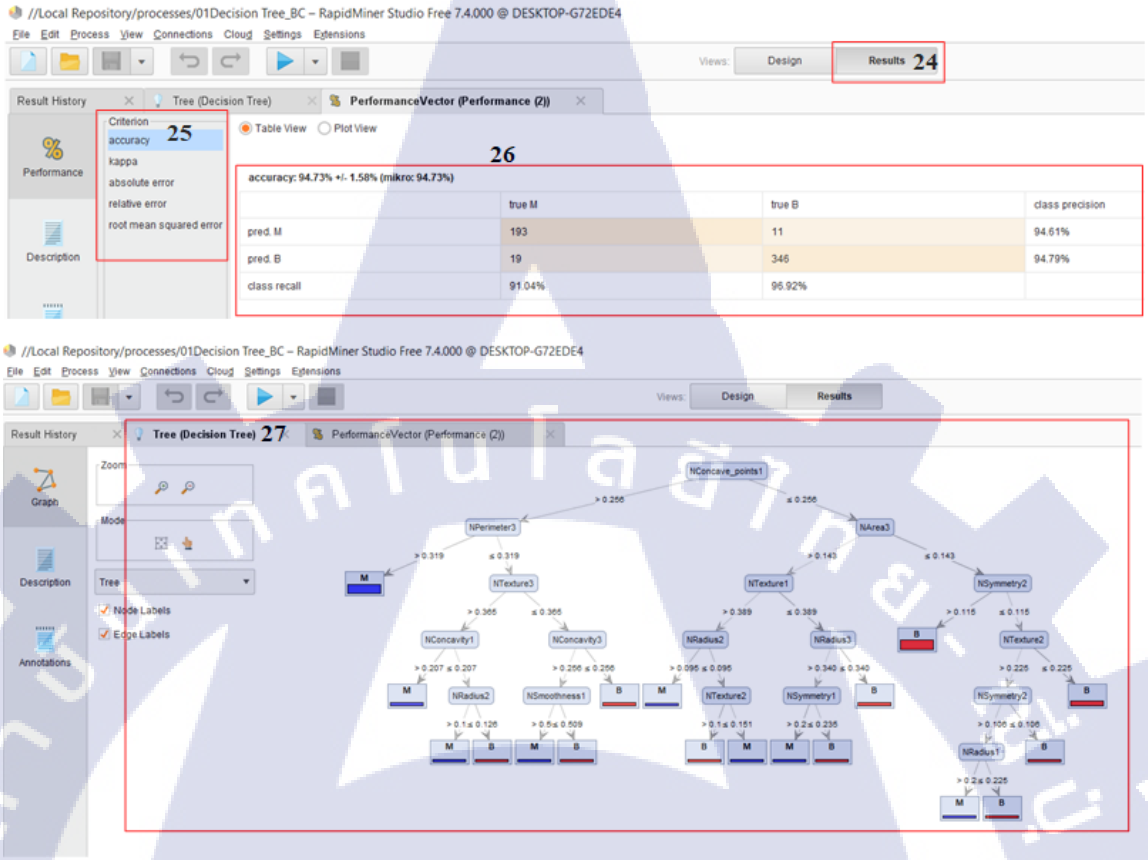


Figure 4.27 The Performance Results in RapidMiner Studio 7.4

Table 4.1 Full Combination Evaluation of Data Classification

Classifier	**Parameter**	**No. of Customize-value**	**Full Combination Evaluation**
Decision Tree	k-fold	2	162
	Max Depth	3	
	Confidence level	3	
	Min Gain	3	
	Minimal leaf size	3	

Table 4.1 Full Combination Evaluation of Data Classification (Cont.)

Classifier	**Parameter**	**No. of Customize-value**	**Full Combination Evaluation**
Naïve Bayes	k-fold Laplace correction	2 2	4
Artificial Neural Network	k-fold Hidden Layer Training Cycle Learning rate	2 4 3 3	72
Support Vector Machine	k-fold Kernel type Gamma C-Value	2 4 3 3	60 (Gamma=N/A for Linear function)

This research has conducted the performance evaluation of data classification in each classifier by random the parameter customization. However, the full combination method is performed in order to evaluate the performance which considering the number of the parameter and the number of customizing values. It can be seen in table 4.1. The highest performance can be considered from the highest percentage of accuracy, lowest the ratio of RMSE, and shortest of classification lead time.

4.7 The Performance of Decision tree (DT)

This classifier has conducted to evaluate the performance of 162 combinations which considered from five factors of k-fold, max depth, confidence level, min gain, and minimal leaf size. The classification result shown the highest percentage of accuracy with 94.90% which provided the RMSE with 0.215, and the shortest classification lead time with 1.53 sec. However, the classification result of parameter customizations can be seen

in Figure 4.28 – 4.32. Moreover, the highest performance of this technique is shown in the gray color as can be seen in Figure 4.28.

Decision tree: Evaluation performance result-1

No	k-fold	Max Depth.	Pruning	Pre-pruning		%Acc	RMSE	Lead time (sec)
			Confidence Level %	Min Gain%	Minimal Leaf size			
1	5	20	25	10%	2	92.96	0.232	1.49
2	5	20	25	10%	3	94.72	0.217	1.52
3	5	20	25	10%	4	94.38	0.225	1.52
4	5	20	25	15%	2	92.96	0.232	1.49
5	5	20	25	15%	3	94.72	0.217	1.52
6	5	20	25	15%	4	94.38	0.225	1.52
7	5	20	25	5%	2	92.96	0.232	1.49
8	5	20	25	5%	3	94.72	0.217	1.52
9	5	20	25	5%	4	94.38	0.225	1.52
10	5	20	30	10%	2	94.72	0.219	1.50
11	5	20	30	10%	3	94.90	0.215	1.75
12	5	20	30	10%	4	94.20	0.227	1.50
13	5	20	30	15%	2	94.73	0.219	1.50
14	**5**	**20**	**30**	**15%**	**3**	**94.90**	**0.215**	**1.53**
15	5	20	30	15%	4	94.20	0.227	1.58
16	5	20	30	5%	2	94.73	0.219	1.55
17	5	20	30	5%	3	94.90	0.215	1.70
18	5	20	30	5%	4	94.20	0.227	1.55
19	5	20	20	10%	2	92.96	0.232	1.59
20	5	20	20	10%	3	94.72	0.217	1.72
21	5	20	20	10%	4	94.38	0.225	1.63
22	5	20	20	15%	2	92.96	0.232	1.57
23	5	20	20	15%	3	94.72	0.217	1.53
24	5	20	20	15%	4	94.38	0.225	1.52
25	5	20	20	5%	2	92.96	0.232	1.66
26	5	20	20	5%	3	94.72	0.217	1.55
27	5	20	20	5%	4	94.38	0.225	1.62
28	5	25	25	10%	2	92.96	0.232	1.62
29	5	25	25	10%	3	94.72	0.217	1.72
30	5	25	25	10%	4	94.38	0.225	1.60
31	5	25	25	15%	2	92.96	0.232	1.96
32	5	25	25	15%	3	94.72	0.217	1.89
33	5	25	25	15%	4	94.38	0.225	1.81
34	5	25	25	5%	2	92.96	0.232	1.60
35	5	25	25	5%	3	94.72	0.217	1.63
36	5	25	25	5%	4	94.38	0.217	1.59
37	5	25	30	10%	2	94.73	0.219	1.56
38	5	25	30	10%	3	94.90	0.215	1.83
39	5	25	30	10%	4	94.20	0.227	1.58
40	5	25	30	15%	2	94.73	0.219	1.81

Decision tree: Evaluation performance result-2

No	k-fold	Max Depth.	Pruning	Pre-pruning		%Acc	RMSE	Lead time (sec)
			Confidence Level %	Min Gain%	Minimal Leaf size			
51	5	25	20	15%	4	94.38	0.225	1.62
52	5	25	20	5%	2	92.96	0.232	1.60
53	5	25	20	5%	3	94.72	0.217	1.60
54	5	25	20	5%	4	94.38	0.225	1.50
55	5	15	25	10%	2	92.96	0.232	1.66
56	5	15	25	10%	3	94.72	0.217	1.49
57	5	15	25	10%	4	94.38	0.225	1.49
58	5	15	25	15%	2	92.96	0.232	1.53
59	5	15	25	15%	3	94.72	0.217	1.56
60	5	15	25	15%	4	94.38	0.225	1.55
61	5	15	25	5%	2	92.96	0.232	1.53
62	5	15	25	5%	3	94.72	0.217	1.56
63	5	15	25	5%	4	94.38	0.225	1.58
64	5	15	30	10%	2	94.73	0.219	1.56
65	**5**	**15**	**30**	**10%**	**3**	**94.90**	**0.215**	**1.53**
66	5	15	30	10%	4	94.20	0.227	1.53
67	5	15	30	15%	2	94.73	0.219	1.76
68	**5**	**15**	**30**	**15%**	**3**	**94.90**	**0.215**	**1.53**
69	5	15	30	15%	4	94.20	0.227	1.60
70	5	15	30	5%	2	94.73	0.219	1.59
71	**5**	**15**	**30**	**5%**	**3**	**94.90**	**0.215**	**1.53**
72	5	15	30	5%	4	94.20	0.227	1.60
73	5	15	20	10%	2	92.96	0.232	1.69
74	5	15	20	10%	3	94.72	0.217	1.60
75	5	15	20	10%	4	94.38	0.225	1.58
76	5	15	20	15%	2	92.96	0.232	1.51
77	5	15	20	15%	3	94.72	0.217	1.53
78	5	15	20	15%	4	94.38	0.225	1.56
79	5	15	20	5%	2	92.96	0.232	1.62
80	5	15	20	5%	3	94.72	0.217	1.59
81	5	15	20	5%	4	94.38	0.225	1.58
82	10	20	25	10%	2	92.80	0.241	1.75
83	10	20	25	10%	3	92.80	0.238	1.62
84	10	20	25	10%	4	92.62	0.237	1.62
85	10	20	25	15%	2	92.80	0.241	1.75
86	10	20	25	15%	3	92.80	0.238	1.58
87	10	20	25	15%	4	92.62	0.237	1.63
88	10	20	25	5%	2	92.80	0.241	1.66
89	10	20	25	5%	3	92.80	0.238	1.59
90	10	20	25	5%	4	92.62	0.237	1.56

Figure 4.28 The Performance Evaluation of Decision Tree Technique-1

Decision tree: Evaluation performance result-3

No	k-fold	Max Depth.	Pruning	Pre-pruning		%Acc	RMSE	Lead time (sec)
			Confidence Level %	Min Gain%	Minimal Leaf size			
41	5	25	30	15%	3	94.90	0.215	1.65
42	5	25	30	15%	4	94.20	0.227	1.65
43	5	25	30	5%	2	94.73	0.219	1.60
44	5	25	30	5%	3	94.90	0.215	1.56
45	5	25	30	5%	4	94.20	0.227	1.53
46	5	25	20	10%	2	92.96	0.232	1.49
47	5	25	20	10%	3	94.72	0.217	1.58
48	5	25	20	10%	4	94.38	0.225	1.49
49	5	25	20	15%	2	92.96	0.232	1.66
50	5	25	20	15%	3	94.72	0.217	1.58
101	10	20	20	10%	3	92.80	0.238	1.60
102	10	20	20	10%	4	92.62	0.237	1.57
103	10	20	20	15%	2	92.80	0.241	1.55
104	10	20	20	15%	3	92.80	0.238	1.56
105	10	20	20	15%	4	92.62	0.237	1.55
106	10	20	20	5%	2	92.80	0.241	1.60
107	10	20	20	5%	3	92.80	0.238	1.59
108	10	20	20	5%	4	92.62	0.237	1.70
109	10	25	25	10%	2	92.80	0.241	1.56
110	10	25	25	10%	3	92.80	0.238	1.59
111	10	25	25	10%	4	92.62	0.237	1.56
112	10	25	25	15%	2	92.80	0.241	1.69
113	10	25	25	15%	3	92.80	0.238	1.59
114	10	25	25	15%	4	92.62	0.237	1.56
115	10	25	25	5%	2	92.80	0.241	1.68
116	10	25	25	5%	3	92.80	0.238	1.56
117	10	25	25	5%	4	92.62	0.237	1.59
118	10	25	30	10%	2	92.97	0.240	1.58
119	10	25	30	10%	3	92.80	0.239	1.56
120	10	25	30	10%	4	93.32	0.236	1.62
121	10	25	30	15%	2	92.97	0.240	1.53
122	10	25	30	15%	3	92.80	0.239	1.66
123	10	25	30	15%	4	93.32	0.236	1.62
124	10	25	30	5%	2	92.97	0.240	1.65
125	10	25	30	5%	3	92.80	0.239	1.56
126	10	25	30	5%	4	93.32	0.236	1.62
127	10	25	20	10%	2	92.80	0.241	1.63
128	10	25	20	10%	3	92.80	0.238	1.62
129	10	25	20	10%	4	92.62	0.237	1.56
130	10	25	20	15%	2	92.80	0.241	1.79
131	10	25	20	15%	3	92.80	0.238	1.56

Decision tree: Evaluation performance result-4

No	k-fold	Max Depth.	Pruning	Pre-pruning		%Acc	RMSE	Lead time (sec)
			Confidence Level %	Min Gain%	Minimal Leaf size			
91	10	20	30	10%	2	92.97	0.240	1.59
92	10	20	30	10%	3	92.80	0.239	1.63
93	10	20	30	10%	4	93.32	0.236	1.66
94	10	20	30	15%	2	92.97	0.240	1.59
95	10	20	30	15%	3	92.80	0.239	1.63
96	10	20	30	15%	4	93.32	0.236	1.76
97	10	20	30	5%	2	92.97	0.240	1.66
98	10	20	30	5%	3	92.80	0.236	1.53
99	10	20	30	5%	4	93.32	0.236	1.56
100	10	20	20	10%	2	92.80	0.241	1.66
132	10	25	20	15%	4	92.62	0.237	1.59
133	10	25	20	5%	2	92.80	0.241	1.56
134	10	25	20	5%	3	92.80	0.238	1.59
135	10	25	20	5%	4	92.62	0.237	1.56
136	10	15	25	10%	2	92.80	0.241	1.63
137	10	15	25	10%	3	92.80	0.238	1.58
138	10	15	25	10%	4	92.62	0.237	1.63
139	10	15	25	15%	2	92.80	0.241	1.59
140	10	15	25	15%	3	92.80	0.238	1.56
141	10	15	25	15%	4	92.62	0.237	1.59
142	10	15	25	5%	2	92.80	0.241	1.60
143	10	15	25	5%	3	92.80	0.238	1.59
144	10	15	25	5%	4	92.62	0.237	1.56
145	10	15	30	10%	2	92.97	0.240	1.54
146	10	15	30	10%	3	92.80	0.239	1.50
147	10	15	30	10%	4	93.32	0.236	1.62
148	10	15	30	15%	2	92.97	0.240	1.53
149	10	15	30	15%	3	92.80	0.239	1.56
150	10	15	30	15%	4	93.32	0.236	1.56
151	10	15	30	5%	2	92.97	0.240	1.66
152	10	15	30	5%	3	92.80	0.239	1.66
153	10	15	30	5%	4	93.32	0.236	1.62
154	10	15	20	10%	2	92.80	0.241	1.64
155	10	15	20	10%	3	92.80	0.238	1.59
156	10	15	20	10%	4	92.62	0.237	1.56
157	10	15	20	15%	2	92.80	0.241	1.76
158	10	15	20	15%	3	92.80	0.238	1.59
159	10	15	20	15%	4	92.62	0.237	1.56
160	10	15	20	5%	2	92.80	0.241	1.76
161	10	15	20	5%	3	92.80	0.238	1.56
162	10	15	20	5%	4	92.62	0.237	1.59

Figure 4.29 The Performance Evaluation of Decision Tree Technique-2

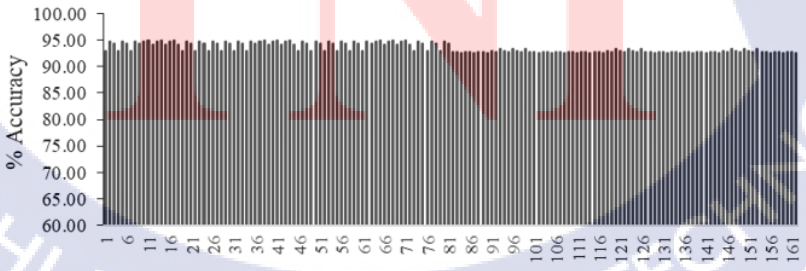


Figure 4.30 The Accuracy Result of Decision Tree

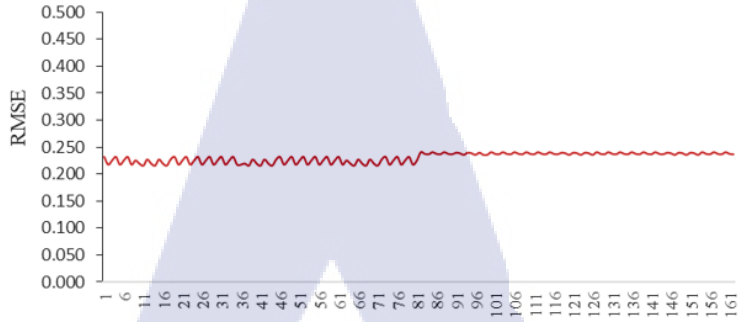


Figure 4.31 The RMSE Result of Decision Tree

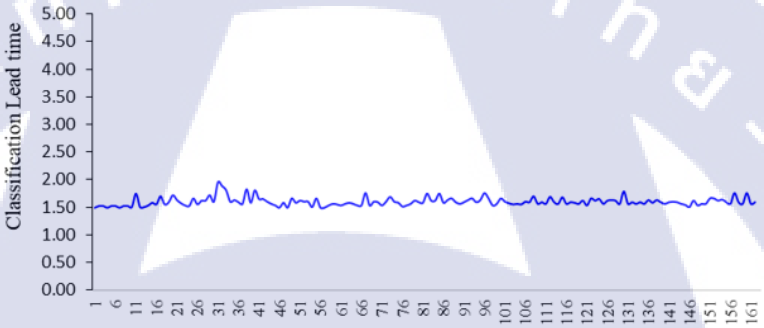


Figure 4.32 The classification lead time of Decision tree

Furthermore, the appropriate of parameter customizing can predict and show the classification result in RapidMiner studio 7.4 program which declares into two categories of confusion matrix table and decision tree model. It can be seen in Figure 4.33 - 4.34 respectively.

accuracy: 94.90% +/- 1.52% (mikro: 94.90%)

	true M	true B	class precision
pred. M	195	12	94.20%
pred. B	17	345	95.30%
class recall	91.98%	96.64%	

Figure 4.33 The Confusion Matrix of the Decision Tree

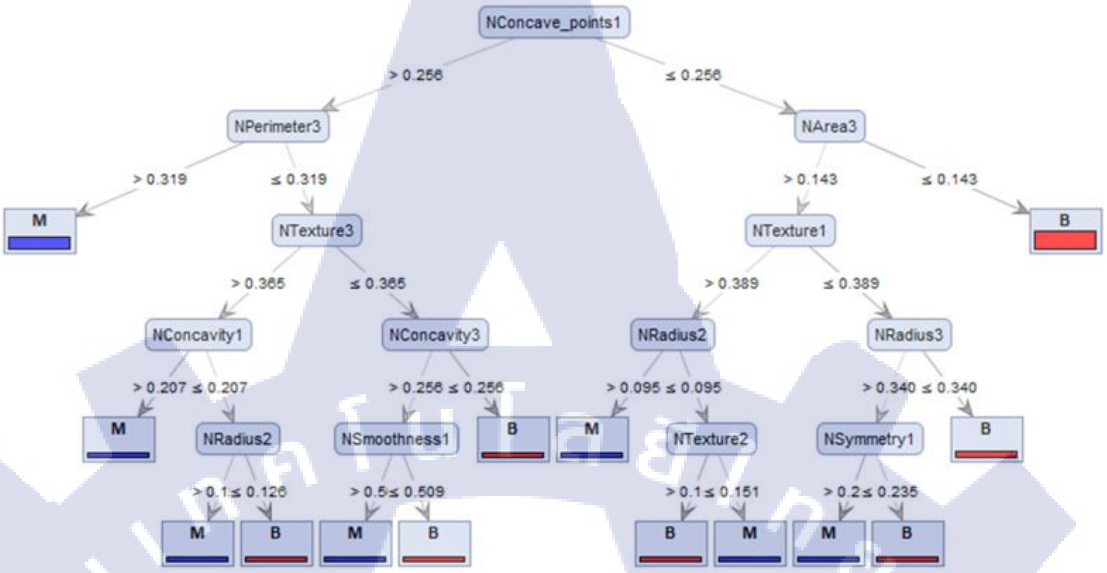


Figure 4.34 The Decision Tree Model from RapidMiner studio 7.4 Program

4.8 The Performance of Naïve Bayes (NB)

This classifier has conducted to evaluate the performance of 4 combinations which customize two factors of k-fold and Laplace correction using. The classification result shows the highest percentage of accuracy with 92.26% which provided the error ratio of RMSE with 0.258, and the classification lead time with 0.52 sec. Nonetheless, the classification result of parameter customizations can be seen in Figure 4.35-4.38 which shown the highest performance in the gray color as can be seen in Figure 4.35.

No	k-fold	Laplace Using	(%) Accuracy	RMSE	Lead time (sec)
1	**5**	**Yes**	**92.26**	**0.258**	**0.52**
2	**5**	**No**	**92.26**	**0.258**	**0.52**
3	10	Yes	92.10	0.263	0.62
4	10	No	92.10	0.263	0.62

Figure 4.35 The Performance Evaluation of Naïve Bayes Technique

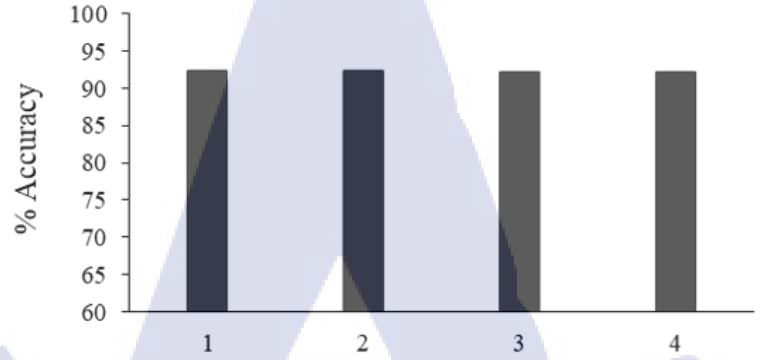


Figure 4.36 The Accuracy Result of Naïve Bayes

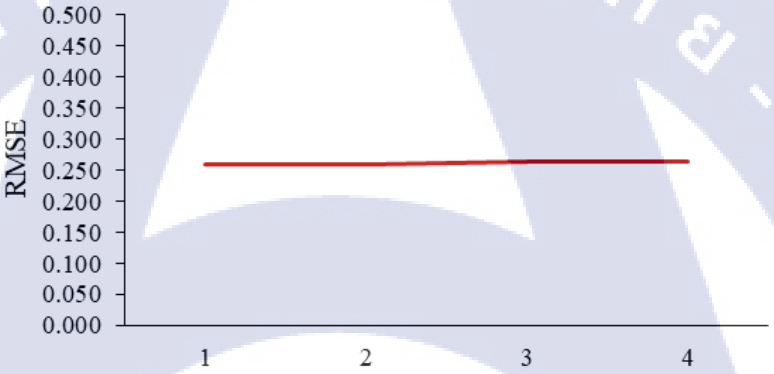


Figure 4.37 The RMSE Result of Naïve Bayes

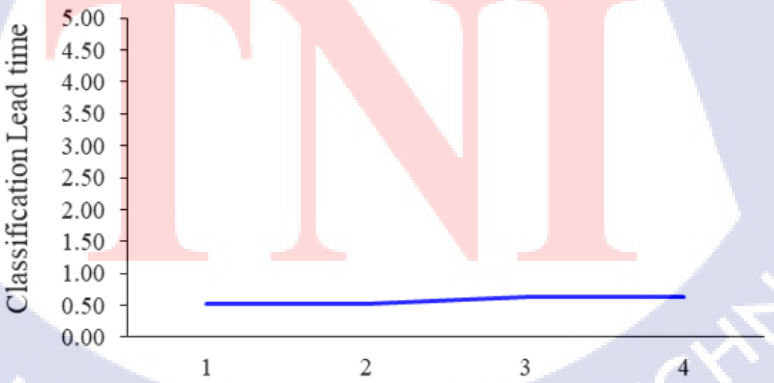


Figure 4.38 The Classification Lead Time of Naïve Bayes

Moreover, the appropriate of parameter customizing can predict and show the classification result in RapidMiner studio 7.4 program which declares into two categories of statistical and confusion matrix which can be seen in Figure 4.39-4.41 respectively.

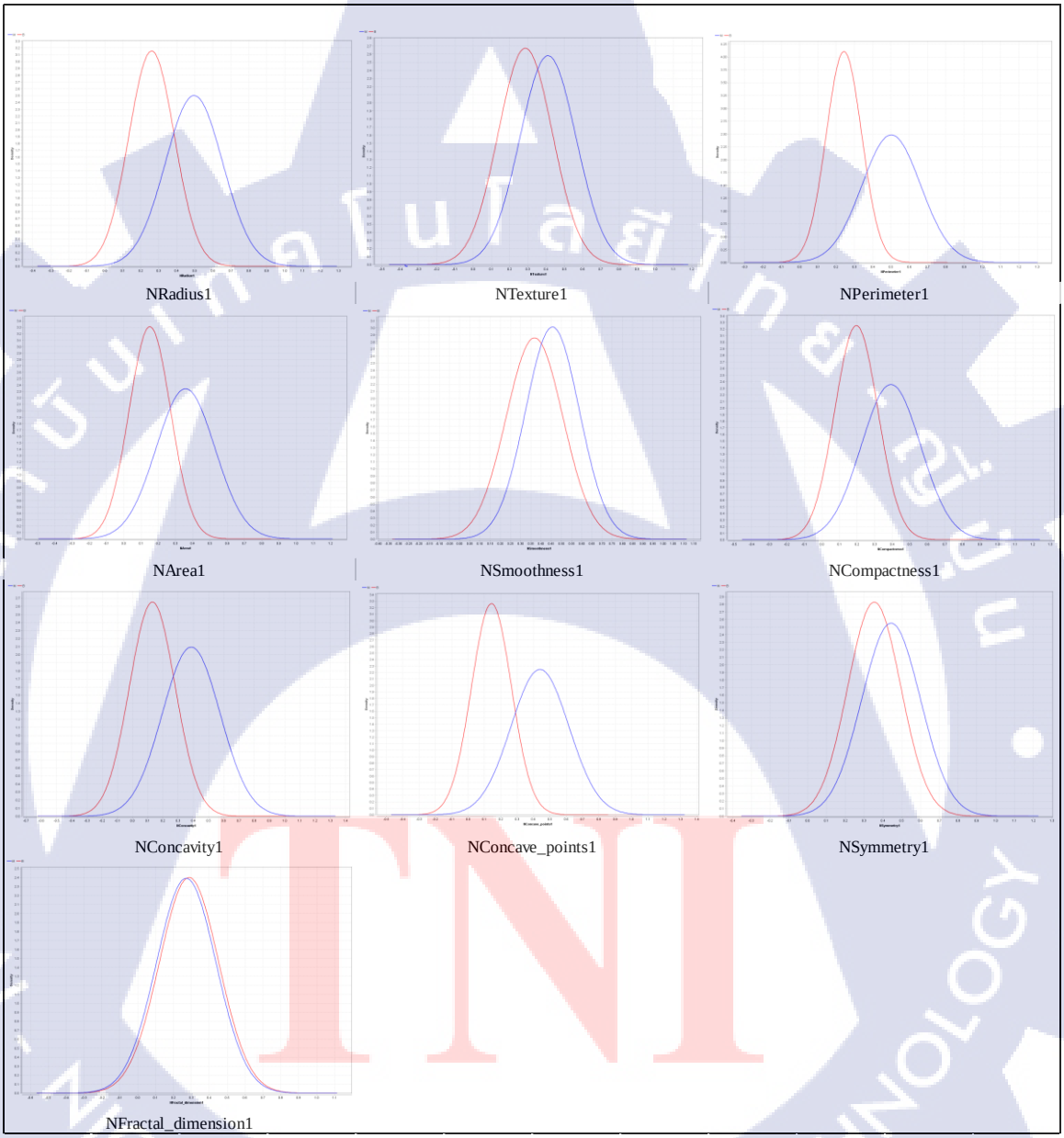


Figure 4.39 The Sample of Distribution Result of the Attribute of Naive Bayes.

Attribute	Parameter	Class: M	Class: B
NRadius1	mean	0.4958	0.2599
NRadius1	standard deviation	0.1595	0.1265
NTexture1	mean	0.4110	0.2865
NTexture1	standard deviation	0.1545	0.1492
NPerimeter1	mean	0.5014	0.2426
NPerimeter1	standard deviation	0.1607	0.0972
NArea1	mean	0.3583	0.1514
NArea1	standard deviation	0.1702	0.1204
NSmoothness1	mean	0.4561	0.3679
NSmoothness1	standard deviation	0.1324	0.1395
NCompactness1	mean	0.3929	0.1961
NCompactness1	standard deviation	0.1691	0.1227
NConcavity1	mean	0.3854	0.1299
NConcavity1	standard deviation	0.1904	0.1504
NConcave_points1	mean	0.4395	0.1432
NConcave_points1	standard deviation	0.1777	0.1223
NSymmetry1	mean	0.4435	0.3541
NSymmetry1	standard deviation	0.1563	0.1410
NFractal_dimension1	mean	0.2714	0.2873
NFractal_dimension1	standard deviation	0.1668	0.1662
NRadius2	mean	0.1952	0.0868
NRadius2	standard deviation	0.1552	0.1360
NTexture2	mean	0.2088	0.2023
NTexture2	standard deviation	0.1389	0.1542
NPerimeter2	mean	0.1919	0.0801
NPerimeter2	standard deviation	0.1636	0.1206
NArea2	mean	0.1443	0.0473
NArea2	standard deviation	0.1590	0.1087
NSmoothness2	mean	0.1508	0.1635
NSmoothness2	standard deviation	0.1433	0.1381
NCompactness2	mean	0.2387	0.1751
NCompactness2	standard deviation	0.1584	0.1742
NConcavity2	mean	0.1230	0.0894
NConcavity2	standard deviation	0.1117	0.1463
NConcave_points2	mean	0.3017	0.1990
NConcave_points2	standard deviation	0.1371	0.1389
NSymmetry2	mean	0.1855	0.1959
NSymmetry2	standard deviation	0.1582	0.1377
NFractal_dimension2	mean	0.1188	0.1105
NFractal_dimension2	standard deviation	0.1041	0.1410

Attribute	Parameter	Class: M	Class: B
NRadius3	mean	0.4650	0.2078
NRadius3	standard deviation	0.1574	0.1120
NTexture3	mean	0.4705	0.3121
NTexture3	standard deviation	0.1597	0.1552
NPerimeter3	mean	0.4522	0.1995
NPerimeter3	standard deviation	0.1580	0.1173
NArea3	mean	0.3124	0.1185
NArea3	standard deviation	0.1590	0.1326
NSmoothness3	mean	0.4899	0.3623
NSmoothness3	standard deviation	0.1514	0.1471
NCompactness3	mean	0.3439	0.1652
NCompactness3	standard deviation	0.1732	0.1264
NConcavity3	mean	0.3672	0.1525
NConcavity3	standard deviation	0.1540	0.1480
NConcave_points3	mean	0.6188	0.2659
NConcave_points3	standard deviation	0.1777	0.1426
NSymmetry3	mean	0.3345	0.2421
NSymmetry3	standard deviation	0.1575	0.1313
NFractal_dimension3	mean	0.2466	0.1735
NFractal_dimension3	standard deviation	0.1575	0.1266

Figure 4.40 The Statistical Value of Each Attribute from Naïve Bayes

accuracy: 92.26% +/- 2.65% (mikro: 92.27%)

	true M	true B	class precision
pred. M	186	18	91.18%
pred. B	26	339	92.88%
class recall	87.74%	94.96%	

Figure 4.41 The Confusion Matrix of the Naïve Bayes

No	k-fold	Hidden Layer	Training Cycle	Learning rate	%Acc	RMSE	Lead time (sec)	No	k-fold	Hidden Layer	Training Cycle	Learning rate	%Acc	RMSE	Lead time (sec)
1	5	1	500	30%	93.67	0.227	11.35	37	10	1	500	30%	94.21	0.217	20.29
2	5	1	500	35%	95.08	0.210	10.76	38	10	1	500	35%	94.56	0.211	20.33
3	5	1	500	25%	93.85	0.226	10.58	39	10	1	500	25%	94.03	0.227	20.25
4	5	1	600	30%	94.02	0.225	12.90	40	10	1	600	30%	94.21	0.216	24.33
5	5	1	600	35%	94.38	0.218	12.36	41	10	1	600	35%	94.56	0.210	24.23
6	5	1	600	25%	94.38	0.222	12.49	42	10	1	600	25%	93.86	0.227	24.55
7	5	1	400	30%	94.20	0.223	9.13	43	10	1	400	30%	94.91	0.206	16.20
8	5	1	400	35%	94.90	0.216	8.62	44	10	1	400	35%	94.73	0.210	16.17
9	5	1	400	25%	94.02	0.220	9.03	45	10	1	400	25%	93.86	0.223	16.33
10	5	2	500	30%	94.02	0.230	15.21	46	10	2	500	30%	94.03	0.229	37.93
11	5	2	500	35%	94.55	0.221	19.22	47	10	2	500	35%	93.33	0.239	38.00
12	5	2	500	25%	94.02	0.231	19.19	48	10	2	500	25%	92.98	0.245	38.37
13	5	2	600	30%	94.20	0.224	23.79	49	10	2	600	30%	94.03	0.227	46.12
14	5	2	600	35%	94.37	0.223	23.13	50	10	2	600	35%	93.50	0.234	45.92
15	5	2	600	25%	93.85	0.235	23.00	51	10	2	600	25%	92.80	0.246	45.59
16	5	2	400	30%	94.55	0.222	15.56	52	10	2	400	30%	94.03	0.227	30.68
17	5	2	400	35%	94.55	0.217	15.43	53	10	2	400	35%	94.03	0.228	30.76
18	5	2	400	25%	94.02	0.221	15.56	54	10	2	400	25%	93.86	0.230	30.80
19	5	3	500	30%	94.02	0.235	25.67	55	10	3	500	30%	94.56	0.222	48.33
20	5	3	500	35%	94.02	0.232	24.48	56	10	3	500	35%	94.21	0.219	48.50
21	5	3	500	25%	93.85	0.234	24.42	57	10	3	500	25%	94.21	0.229	49.10
22	5	3	600	30%	94.20	0.233	25.50	58	10	3	600	30%	94.21	0.226	57.89
23	5	3	600	35%	93.85	0.239	29.85	59	10	3	600	35%	94.38	0.221	58.12
24	5	3	600	25%	93.85	0.234	29.82	60	10	3	600	25%	94.21	0.233	58.17
25	5	3	400	30%	94.02	0.235	20.06	61	10	3	400	30%	94.03	0.223	38.89
26	5	3	400	35%	94.55	0.221	19.72	**62**	**10**	**3**	**400**	**35%**	**95.08**	**0.194**	**38.89**
27	5	3	400	25%	93.85	0.228	19.66	63	10	3	400	25%	94.21	0.224	38.87
28	5	4	500	30%	93.85	0.237	34.72	64	10	4	500	30%	94.38	0.221	59.69
29	5	4	500	35%	93.32	0.247	35.06	65	10	4	500	35%	94.91	0.212	59.95
30	5	4	500	25%	93.32	0.241	44.47	66	10	4	500	25%	94.56	0.21	76.99
31	5	4	600	30%	93.67	0.241	47.62	67	10	4	600	30%	94.38	0.22	58.12
32	5	4	600	35%	93.49	0.242	47.12	68	10	4	600	35%	94.03	0.225	58.72
33	5	4	600	25%	93.32	0.245	39.69	69	10	4	600	25%	94.73	0.217	59.45
34	5	4	400	30%	94.02	0.232	39.72	70	10	4	400	30%	95.06	0.201	59.01
35	5	4	400	35%	93.67	0.233	32.59	71	10	4	400	35%	94.38	0.221	59.69
36	5	4	400	25%	94.20	0.229	35.46	72	10	4	400	25%	94.38	0.209	59.98

Figure 4.42 The Performance Evaluation of Artificial Neural Network Technique

4.9 The Performance of Artificial Neural Network (ANN)

This classifier has performed to evaluate the performance of 54 combinations which considered from four factors of k-fold, hidden layer, training cycle, and learning rate. The classification result shown the highest percentage of accuracy with 95.08% which the error of RMSE with 0.194, and the classification lead time with 38.89 sec. However, the classification result of parameter customizations can be seen in Figure 4.42-4.45 Moreover, the highest performance of this technique is shown in the gray color as can be seen in Figure 4.42.

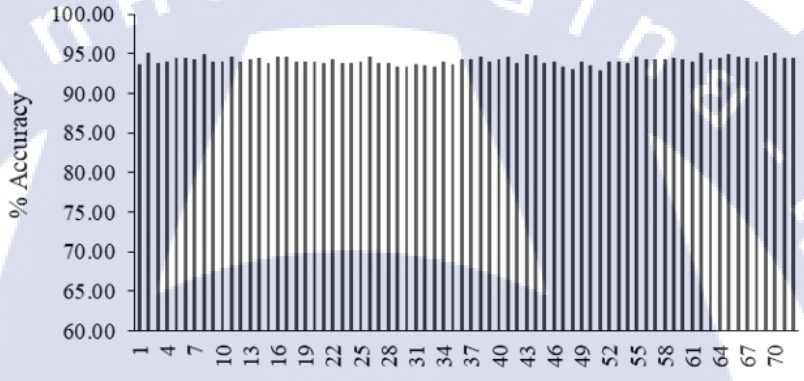


Figure 4.43 The Accuracy Result of ANN

Figure 4.44 The RMSE Result of ANN

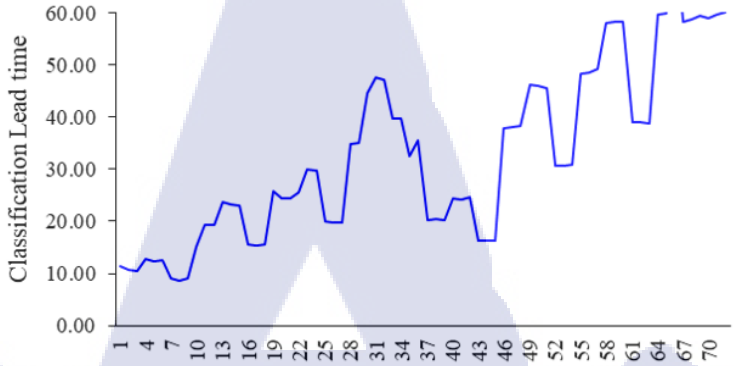


Figure 4.45 The Classification Lead Time of ANN

Furthermore, the appropriate of parameter customizing can predict and show the classification result in RapidMiner studio 7.4 program which declares into three categories of ANN model, a hidden layer and output layer, and confusion matrix which can be seen in Figure 4.46-4.52 respectively.

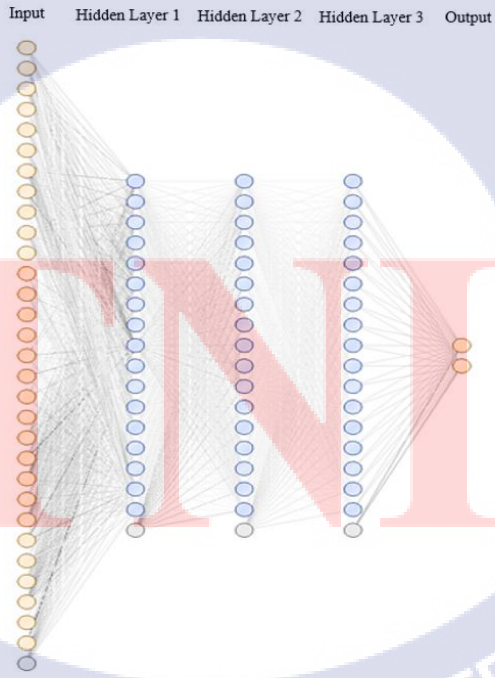


Figure 4.46 The ANN Model from RapidMiner Studio 7.4 Program

Hidden 1
========

Node 1 (Sigmoid)	Node 2 (Sigmoid)	Node 3 (Sigmoid)	Node 4 (Sigmoid)	Node 5 (Sigmoid)
NRadius1: 2.012	NRadius1: 1.569	NRadius1: 0.969	NRadius1: 0.760	NRadius1: 1.137
NTexture1: 2.477	NTexture1: 1.785	NTexture1: 1.420	NTexture1: 1.085	NTexture1: 1.680
NPerimeter1: 3.416	NPerimeter1: 2.176	NPerimeter1: 2.440	NPerimeter1: 1.404	NPerimeter1: 1.959
NArea1: 0.947	NArea1: 0.399	NArea1: 0.379	NArea1: -0.002	NArea1: 0.269
NSmoothness1: 0.255	NSmoothness1: 1.176	NSmoothness1: -0.104	NSmoothness1: 0.657	NSmoothness1: 1.413
NCompactness1: 0.534	NCompactness1: 0.712	NCompactness1: 0.755	NCompactness1: 0.763	NCompactness1: 1.052
NConcavity1: 4.429	NConcavity1: 3.432	NConcavity1: 3.208	NConcavity1: 2.290	NConcavity1: 3.666
NConcave_points1: 1.325	NConcave_points1: 1.129	NConcave_points1: 1.650	NConcave_points1: 1.099	NConcave_points1: 1.496
NSymmetry1: 1.338	NSymmetry1: 0.810	NSymmetry1: 0.509	NSymmetry1: 0.225	NSymmetry1: 0.500
NFractal_dimension1: -5.372	NFractal_dimension1: -3.449	NFractal_dimension1: -3.271	NFractal_dimension1: -2.011	NFractal_dimension1: -3.182
NRadius2: -0.200	NRadius2: -0.230	NRadius2: -0.229	NRadius2: -0.146	NRadius2: -0.186
NTexture2: -1.637	NTexture2: -0.249	NTexture2: -1.646	NTexture2: -0.394	NTexture2: -0.493
NPerimeter2: 1.792	NPerimeter2: 1.046	NPerimeter2: 1.633	NPerimeter2: 0.954	NPerimeter2: 1.453
NArea2: -1.602	NArea2: -0.991	NArea2: -0.854	NArea2: -0.382	NArea2: -0.925
NSmoothness2: 0.109	NSmoothness2: 0.614	NSmoothness2: -0.120	NSmoothness2: 0.384	NSmoothness2: 0.598
NCompactness2: -2.806	NCompactness2: -0.684	NCompactness2: -2.535	NCompactness2: -0.777	NCompactness2: -0.926
NConcavity2: 0.390	NConcavity2: 1.186	NConcavity2: -0.326	NConcavity2: 0.341	NConcavity2: 0.914
NConcave_points2: -1.323	NConcave_points2: -0.250	NConcave_points2: -1.622	NConcave_points2: -0.583	NConcave_points2: -0.565
NSymmetry2: -0.336	NSymmetry2: -0.517	NSymmetry2: -0.182	NSymmetry2: -0.261	NSymmetry2: -0.245
NFractal_dimension2: -3.523	NFractal_dimension2: -1.430	NFractal_dimension2: -2.809	NFractal_dimension2: -1.207	NFractal_dimension2: -1.598
NRadius3: 2.938	NRadius3: 1.441	NRadius3: 2.530	NRadius3: 1.314	NRadius3: 1.516
NTexture3: 5.303	NTexture3: 2.944	NTexture3: 3.954	NTexture3: 2.079	NTexture3: 3.100
NPerimeter3: 2.220	NPerimeter3: 1.457	NPerimeter3: 1.598	NPerimeter3: 1.110	NPerimeter3: 1.423
NArea3: 3.895	NArea3: 2.711	NArea3: 3.016	NArea3: 1.926	NArea3: 2.791
NSmoothness3: 1.840	NSmoothness3: 0.569	NSmoothness3: 2.312	NSmoothness3: 1.003	NSmoothness3: 0.871
NCompactness3: -0.186	NCompactness3: -0.718	NCompactness3: 0.403	NCompactness3: -0.200	NCompactness3: -0.737
NConcavity3: 2.771	NConcavity3: 1.424	NConcavity3: 2.480	NConcavity3: 1.268	NConcavity3: 1.712
NConcave_points3: 4.041	NConcave_points3: 2.526	NConcave_points3: 3.341	NConcave_points3: 1.965	NConcave_points3: 2.856
NSymmetry3: 2.052	NSymmetry3: -0.429	NSymmetry3: 1.984	NSymmetry3: -0.055	NSymmetry3: -0.343
NFractal_dimension3: -0.304	NFractal_dimension3: -1.985	NFractal_dimension3: 0.003	NFractal_dimension3: -1.098	NFractal_dimension3: -2.123
Bias: 4.781	Bias: 2.603	Bias: 3.292	Bias: 1.725	Bias: 2.952

Node 6 (Sigmoid)	Node 7 (Sigmoid)	Node 8 (Sigmoid)	Node 9 (Sigmoid)	Node 10 (Sigmoid)
NRadius1: 2.332	NRadius1: 1.208	NRadius1: 0.748	NRadius1: -1.736	NRadius1: 0.596
NTexture1: 2.262	NTexture1: 1.556	NTexture1: 1.229	NTexture1: -4.300	NTexture1: 1.286
NPerimeter1: 3.221	NPerimeter1: 1.632	NPerimeter1: 1.489	NPerimeter1: 3.301	NPerimeter1: 1.679
NArea1: 0.576	NArea1: 0.344	NArea1: 0.097	NArea1: -4.195	NArea1: -0.057
NSmoothness1: 1.042	NSmoothness1: 1.339	NSmoothness1: 0.953	NSmoothness1: -3.240	NSmoothness1: 1.010
NCompactness1: 0.419	NCompactness1: 0.809	NCompactness1: 0.841	NCompactness1: 3.812	NCompactness1: 1.201
NConcavity1: 4.117	NConcavity1: 3.243	NConcavity1: 2.730	NConcavity1: 1.257	NConcavity1: 3.198
NConcave_points1: 1.165	NConcave_points1: 1.183	NConcave_points1: 1.264	NConcave_points1: 5.717	NConcave_points1: 1.630
NSymmetry1: 1.320	NSymmetry1: 0.579	NSymmetry1: 0.312	NSymmetry1: -0.035	NSymmetry1: 0.078
NFractal_dimension1: -4.704	NFractal_dimension1: -2.803	NFractal_dimension1: -2.386	NFractal_dimension1: -0.745	NFractal_dimension1: -2.475
NRadius2: -0.271	NRadius2: -0.174	NRadius2: -0.181	NRadius2: 0.568	NRadius2: -0.137
NTexture2: -0.587	NTexture2: -0.156	NTexture2: -0.426	NTexture2: -4.311	NTexture2: -0.725
NPerimeter2: 1.074	NPerimeter2: 1.069	NPerimeter2: 1.217	NPerimeter2: 3.867	NPerimeter2: 1.605
NArea2: -1.249	NArea2: -0.876	NArea2: -0.536	NArea2: 1.827	NArea2: -0.644
NSmoothness2: 0.481	NSmoothness2: 0.644	NSmoothness2: 0.442	NSmoothness2: -1.347	NSmoothness2: 0.454
NCompactness2: -1.375	NCompactness2: -0.513	NCompactness2: -0.880	NCompactness2: -4.285	NCompactness2: -1.277
NConcavity2: 1.443	NConcavity2: 1.105	NConcavity2: 0.495	NConcavity2: -5.104	NConcavity2: 0.312
NConcave_points2: -0.431	NConcave_points2: -0.275	NConcave_points2: -0.606	NConcave_points2: -3.666	NConcave_points2: -0.860
NSymmetry2: -1.033	NSymmetry2: -0.278	NSymmetry2: -0.230	NSymmetry2: -0.536	NSymmetry2: -0.181
NFractal_dimension2: -2.271	NFractal_dimension2: -1.218	NFractal_dimension2: -1.364	NFractal_dimension2: -4.676	NFractal_dimension2: -1.763
NRadius3: 2.059	NRadius3: 1.188	NRadius3: 1.399	NRadius3: 7.881	NRadius3: 1.768
NTexture3: 3.984	NTexture3: 2.641	NTexture3: 2.417	NTexture3: 2.591	NTexture3: 2.942
NPerimeter3: 1.922	NPerimeter3: 1.288	NPerimeter3: 1.218	NPerimeter3: 0.257	NPerimeter3: 1.403
NArea3: 3.524	NArea3: 2.397	NArea3: 2.193	NArea3: 3.282	NArea3: 2.614
NSmoothness3: 0.704	NSmoothness3: 0.372	NSmoothness3: 0.962	NSmoothness3: 6.262	NSmoothness3: 1.421
NCompactness3: -0.648	NCompactness3: -0.763	NCompactness3: -0.395	NCompactness3: 1.882	NCompactness3: -0.312
NConcavity3: 1.759	NConcavity3: 1.392	NConcavity3: 1.444	NConcavity3: 1.850	NConcavity3: 1.824
NConcave_points3: 3.200	NConcave_points3: 2.313	NConcave_points3: 2.266	NConcave_points3: 3.925	NConcave_points3: 2.847
NSymmetry3: 0.008	NSymmetry3: -0.546	NSymmetry3: -0.087	NSymmetry3: 1.465	NSymmetry3: -0.064
NFractal_dimension3: -1.382	NFractal_dimension3: -2.112	NFractal_dimension3: -1.306	NFractal_dimension3: 8.986	NFractal_dimension3: -1.407
Bias: 3.503	Bias: 2.436	Bias: 2.165	Bias: 2.249	Bias: 2.670

Figure 4.47 The Result of Hidden Layer 1 of ANN from RapidMiner Studio 7.4-1

Node 11 (Sigmoid)	Node 12 (Sigmoid)	Node 13 (Sigmoid)	Node 14 (Sigmoid)	Node 15 (Sigmoid)
NRadius1: 0.274	NRadius1: 0.841	NRadius1: 0.831	NRadius1: 0.328	NRadius1: 0.250
NTexture1: 0.628	NTexture1: 1.199	NTexture1: 1.290	NTexture1: 0.479	NTexture1: 0.690
NPerimeter1: 1.300	NPerimeter1: 1.426	NPerimeter1: 1.667	NPerimeter1: 0.981	NPerimeter1: 1.367
NArea1: -0.349	NArea1: 0.083	NArea1: 0.132	NArea1: -0.236	NArea1: -0.423
NSmoothness1: 0.170	NSmoothness1: 0.937	NSmoothness1: 0.858	NSmoothness1: 0.211	NSmoothness1: 0.310
NCompactness1: 1.110	NCompactness1: 0.757	NCompactness1: 0.801	NCompactness1: 0.736	NCompactness1: 1.188
NConcavity1: 2.195	NConcavity1: 2.566	NConcavity1: 2.711	NConcavity1: 1.484	NConcavity1: 2.331
NConcave_points1: 1.475	NConcave_points1: 1.162	NConcave_points1: 1.224	NConcave_points1: 1.026	NConcave_points1: 1.509
NSymmetry1: -0.174	NSymmetry1: 0.282	NSymmetry1: 0.304	NSymmetry1: -0.138	NSymmetry1: -0.228
NFractal_dimension1: -1.516	NFractal_dimension1: -2.122	NFractal_dimension1: -2.515	NFractal_dimension1: -0.984	NFractal_dimension1: -1.587
NRadius2: -0.053	NRadius2: -0.174	NRadius2: -0.246	NRadius2: -0.018	NRadius2: -0.098
NTexture2: -0.851	NTexture2: -0.393	NTexture2: -0.432	NTexture2: -0.411	NTexture2: -0.779
NPerimeter2: 1.332	NPerimeter2: 1.037	NPerimeter2: 1.205	NPerimeter2: 0.840	NPerimeter2: 1.464
NArea2: -0.243	NArea2: -0.553	NArea2: -0.599	NArea2: -0.063	NArea2: -0.269
NSmoothness2: 0.322	NSmoothness2: 0.380	NSmoothness2: 0.464	NSmoothness2: 0.274	NSmoothness2: 0.311
NCompactness2: -1.042	NCompactness2: -0.923	NCompactness2: -0.890	NCompactness2: -0.590	NCompactness2: -1.164
NConcavity2: -0.377	NConcavity2: 0.515	NConcavity2: 0.475	NConcavity2: -0.141	NConcavity2: -0.343
NConcave_points2: -0.997	NConcave_points2: -0.609	NConcave_points2: -0.551	NConcave_points2: -0.705	NConcave_points2: -1.082
NSymmetry2: -0.271	NSymmetry2: -0.277	NSymmetry2: -0.270	NSymmetry2: -0.311	NSymmetry2: -0.407
NFractal_dimension2: -1.463	NFractal_dimension2: -1.248	NFractal_dimension2: -1.350	NFractal_dimension2: -0.873	NFractal_dimension2: -1.554
NRadius3: 1.677	NRadius3: 1.329	NRadius3: 1.465	NRadius3: 0.976	NRadius3: 1.747
NTexture3: 2.120	NTexture3: 2.373	NTexture3: 2.396	NTexture3: 1.461	NTexture3: 2.277
NPerimeter3: 1.179	NPerimeter3: 1.210	NPerimeter3: 1.222	NPerimeter3: 0.910	NPerimeter3: 1.274
NArea3: 2.037	NArea3: 2.097	NArea3: 2.228	NArea3: 1.384	NArea3: 2.180
NSmoothness3: 1.756	NSmoothness3: 0.880	NSmoothness3: 1.079	NSmoothness3: 1.106	NSmoothness3: 1.739
NCompactness3: 0.218	NCompactness3: -0.329	NCompactness3: -0.353	NCompactness3: 0.204	NCompactness3: 0.276
NConcavity3: 1.478	NConcavity3: 1.347	NConcavity3: 1.414	NConcavity3: 0.978	NConcavity3: 1.527
NConcave_points3: 2.265	NConcave_points3: 2.116	NConcave_points3: 2.263	NConcave_points3: 1.516	NConcave_points3: 2.395
NSymmetry3: 0.117	NSymmetry3: -0.143	NSymmetry3: -0.047	NSymmetry3: -0.001	NSymmetry3: 0.034
NFractal_dimension3: -0.832	NFractal_dimension3: -1.218	NFractal_dimension3: -1.341	NFractal_dimension3: -0.730	NFractal_dimension3: -0.811
Bias: 1.788	Bias: 1.919	Bias: 2.095	Bias: 0.970	Bias: 1.892

Node 16 (Sigmoid)	Node 17 (Sigmoid)
NRadius1: -0.395	NRadius1: 0.289
NTexture1: -0.987	NTexture1: 0.728
NPerimeter1: 1.598	NPerimeter1: 1.258
NArea1: -1.855	NArea1: -0.296
NSmoothness1: -1.052	NSmoothness1: 0.498
NCompactness1: 3.034	NCompactness1: 1.067
NConcavity1: 2.155	NConcavity1: 2.241
NConcave_points1: 2.686	NConcave_points1: 1.418
NSymmetry1: -1.987	NSymmetry1: -0.180
NFractal_dimension1: 0.886	NFractal_dimension1: -1.539
NRadius2: 0.359	NRadius2: -0.094
NTexture2: -2.349	NTexture2: -0.652
NPerimeter2: 1.529	NPerimeter2: 1.342
NArea2: 0.259	NArea2: -0.292
NSmoothness2: -0.281	NSmoothness2: 0.339
NCompactness2: -1.890	NCompactness2: -0.952
NConcavity2: -2.100	NConcavity2: -0.185
NConcave_points2: -2.920	NConcave_points2: -0.981
NSymmetry2: -1.480	NSymmetry2: -0.259
NFractal_dimension2: -2.604	NFractal_dimension2: -1.378
NRadius3: 0.554	NRadius3: 1.557
NTexture3: 2.692	NTexture3: 2.071
NPerimeter3: 2.555	NPerimeter3: 1.171
NArea3: 2.381	NArea3: 2.024
NSmoothness3: 4.330	NSmoothness3: 1.528
NCompactness3: 2.786	NCompactness3: 0.090
NConcavity3: 2.234	NConcavity3: 1.407
NConcave_points3: 3.884	NConcave_points3: 2.192
NSymmetry3: 0.331	NSymmetry3: -0.046
NFractal_dimension3: -1.384	NFractal_dimension3: -0.930
Bias: 1.635	Bias: 1.756

Figure 4.48 The Result of Hidden Layer 1 of ANN from RapidMiner Studio 7.4-2

Hidden 2
========

Node 1 (Sigmoid)	Node 2 (Sigmoid)	Node 3 (Sigmoid)	Node 4 (Sigmoid)	Node 5 (Sigmoid)	Node 6 (Sigmoid)
Node 1: 1.134	Node 1: 0.860	Node 1: 1.325	Node 1: 0.482	Node 1: 1.040	Node 1: 0.990
Node 2: 1.029	Node 2: 0.755	Node 2: 1.203	Node 2: 0.324	Node 2: 0.966	Node 2: 0.951
Node 3: 1.004	Node 3: 0.833	Node 3: 1.216	Node 3: 0.350	Node 3: 0.898	Node 3: 0.915
Node 4: 0.770	Node 4: 0.550	Node 4: 0.950	Node 4: 0.246	Node 4: 0.718	Node 4: 0.632
Node 5: 1.097	Node 5: 0.841	Node 5: 1.263	Node 5: 0.334	Node 5: 0.982	Node 5: 0.896
Node 6: 1.246	Node 6: 0.909	Node 6: 1.455	Node 6: 0.375	Node 6: 1.052	Node 6: 1.068
Node 7: 1.002	Node 7: 0.746	Node 7: 1.205	Node 7: 0.316	Node 7: 0.910	Node 7: 0.831
Node 8: 0.911	Node 8: 0.672	Node 8: 1.049	Node 8: 0.253	Node 8: 0.748	Node 8: 0.733
Node 9: 2.170	Node 9: 1.645	Node 9: 2.598	Node 9: 0.645	Node 9: 2.003	Node 9: 1.928
Node 10: 0.949	Node 10: 0.774	Node 10: 1.164	Node 10: 0.289	Node 10: 0.820	Node 10: 0.856
Node 11: 0.737	Node 11: 0.588	Node 11: 0.927	Node 11: 0.260	Node 11: 0.685	Node 11: 0.664
Node 12: 0.784	Node 12: 0.595	Node 12: 1.028	Node 12: 0.292	Node 12: 0.705	Node 12: 0.729
Node 13: 0.903	Node 13: 0.668	Node 13: 1.045	Node 13: 0.294	Node 13: 0.828	Node 13: 0.715
Node 14: 0.591	Node 14: 0.473	Node 14: 0.715	Node 14: 0.143	Node 14: 0.543	Node 14: 0.484
Node 15: 0.751	Node 15: 0.571	Node 15: 0.940	Node 15: 0.273	Node 15: 0.681	Node 15: 0.668
Node 16: 0.929	Node 16: 0.675	Node 16: 1.056	Node 16: 0.283	Node 16: 0.864	Node 16: 0.830
Node 17: 0.784	Node 17: 0.575	Node 17: 0.863	Node 17: 0.210	Node 17: 0.703	Node 17: 0.649
Bias: -3.085	Bias: -2.615	Bias: -3.489	Bias: -1.845	Bias: -2.909	Bias: -2.840

Node 7 (Sigmoid)	Node 8 (Sigmoid)	Node 9 (Sigmoid)	Node 10 (Sigmoid)	Node 11 (Sigmoid)	Node 12 (Sigmoid)
Node 1: 0.714	Node 1: 0.919	Node 1: 1.045	Node 1: 1.014	Node 1: 1.295	Node 1: 0.739
Node 2: 0.646	Node 2: 0.808	Node 2: 0.972	Node 2: 0.899	Node 2: 1.255	Node 2: 0.569
Node 3: 0.698	Node 3: 0.766	Node 3: 0.892	Node 3: 0.885	Node 3: 1.180	Node 3: 0.571
Node 4: 0.448	Node 4: 0.596	Node 4: 0.694	Node 4: 0.709	Node 4: 0.906	Node 4: 0.416
Node 5: 0.683	Node 5: 0.786	Node 5: 0.962	Node 5: 0.886	Node 5: 1.248	Node 5: 0.611
Node 6: 0.742	Node 6: 0.963	Node 6: 1.144	Node 6: 1.016	Node 6: 1.447	Node 6: 0.768
Node 7: 0.676	Node 7: 0.773	Node 7: 0.915	Node 7: 0.841	Node 7: 1.201	Node 7: 0.529
Node 8: 0.561	Node 8: 0.620	Node 8: 0.770	Node 8: 0.750	Node 8: 1.023	Node 8: 0.539
Node 9: 1.397	Node 9: 1.644	Node 9: 1.956	Node 9: 1.922	Node 9: 2.598	Node 9: 1.252
Node 10: 0.631	Node 10: 0.707	Node 10: 0.901	Node 10: 0.802	Node 10: 1.108	Node 10: 0.573
Node 11: 0.440	Node 11: 0.585	Node 11: 0.735	Node 11: 0.607	Node 11: 0.882	Node 11: 0.409
Node 12: 0.530	Node 12: 0.596	Node 12: 0.720	Node 12: 0.768	Node 12: 1.015	Node 12: 0.463
Node 13: 0.526	Node 13: 0.631	Node 13: 0.778	Node 13: 0.710	Node 13: 1.074	Node 13: 0.527
Node 14: 0.376	Node 14: 0.413	Node 14: 0.508	Node 14: 0.476	Node 14: 0.712	Node 14: 0.314
Node 15: 0.455	Node 15: 0.639	Node 15: 0.704	Node 15: 0.630	Node 15: 0.943	Node 15: 0.473
Node 16: 0.611	Node 16: 0.694	Node 16: 0.818	Node 16: 0.839	Node 16: 1.109	Node 16: 0.525
Node 17: 0.446	Node 17: 0.586	Node 17: 0.720	Node 17: 0.634	Node 17: 0.896	Node 17: 0.421
Bias: -2.400	Bias: -2.621	Bias: -2.893	Bias: -2.830	Bias: -3.492	Bias: -2.304

Node 13 (Sigmoid)	Node 14 (Sigmoid)	Node 15 (Sigmoid)	Node 16 (Sigmoid)	Node 17 (Sigmoid)
Node 1: 0.857	Node 1: 0.969	Node 1: 0.885	Node 1: 0.881	Node 1: 0.857
Node 2: 0.762	Node 2: 0.838	Node 2: 0.815	Node 2: 0.822	Node 2: 0.770
Node 3: 0.690	Node 3: 0.865	Node 3: 0.864	Node 3: 0.791	Node 3: 0.756
Node 4: 0.550	Node 4: 0.651	Node 4: 0.645	Node 4: 0.576	Node 4: 0.522
Node 5: 0.743	Node 5: 0.852	Node 5: 0.905	Node 5: 0.781	Node 5: 0.790
Node 6: 0.817	Node 6: 0.957	Node 6: 0.938	Node 6: 0.951	Node 6: 0.869
Node 7: 0.661	Node 7: 0.791	Node 7: 0.819	Node 7: 0.695	Node 7: 0.715
Node 8: 0.602	Node 8: 0.693	Node 8: 0.716	Node 8: 0.652	Node 8: 0.645
Node 9: 1.583	Node 9: 1.790	Node 9: 1.794	Node 9: 1.630	Node 9: 1.610
Node 10: 0.676	Node 10: 0.788	Node 10: 0.783	Node 10: 0.757	Node 10: 0.667
Node 11: 0.557	Node 11: 0.615	Node 11: 0.616	Node 11: 0.524	Node 11: 0.588
Node 12: 0.599	Node 12: 0.686	Node 12: 0.689	Node 12: 0.573	Node 12: 0.584
Node 13: 0.625	Node 13: 0.715	Node 13: 0.686	Node 13: 0.609	Node 13: 0.612
Node 14: 0.384	Node 14: 0.436	Node 14: 0.431	Node 14: 0.447	Node 14: 0.441
Node 15: 0.516	Node 15: 0.650	Node 15: 0.633	Node 15: 0.618	Node 15: 0.607
Node 16: 0.678	Node 16: 0.733	Node 16: 0.791	Node 16: 0.729	Node 16: 0.644
Node 17: 0.535	Node 17: 0.594	Node 17: 0.631	Node 17: 0.579	Node 17: 0.534
Bias: -2.539	Bias: -2.724	Bias: -2.729	Bias: -2.613	Bias: -2.567

Figure 4.49 The Result of Hidden Layer 2 of ANN from RapidMiner Studio 7.4

Hidden 3
========

Node 1 (Sigmoid)	Node 2 (Sigmoid)	Node 3 (Sigmoid)	Node 4 (Sigmoid)	Node 5 (Sigmoid)	Node 6 (Sigmoid)
Node 1: -0.638	Node 1: -0.685	Node 1: -0.625	Node 1: -0.594	Node 1: -0.683	Node 1: -0.650
Node 2: -0.498	Node 2: -0.533	Node 2: -0.512	Node 2: -0.476	Node 2: -0.458	Node 2: -0.485
Node 3: -0.812	Node 3: -0.856	Node 3: -0.811	Node 3: -0.736	Node 3: -0.760	Node 3: -0.795
Node 4: -0.142	Node 4: -0.188	Node 4: -0.133	Node 4: -0.192	Node 4: -0.137	Node 4: -0.181
Node 5: -0.611	Node 5: -0.655	Node 5: -0.566	Node 5: -0.558	Node 5: -0.612	Node 5: -0.624
Node 6: -0.571	Node 6: -0.540	Node 6: -0.541	Node 6: -0.579	Node 6: -0.558	Node 6: -0.511
Node 7: -0.370	Node 7: -0.361	Node 7: -0.420	Node 7: -0.369	Node 7: -0.414	Node 7: -0.434
Node 8: -0.516	Node 8: -0.463	Node 8: -0.484	Node 8: -0.464	Node 8: -0.497	Node 8: -0.489
Node 9: -0.545	Node 9: -0.559	Node 9: -0.636	Node 9: -0.609	Node 9: -0.628	Node 9: -0.570
Node 10: -0.517	Node 10: -0.587	Node 10: -0.595	Node 10: -0.579	Node 10: -0.594	Node 10: -0.564
Node 11: -0.814	Node 11: -0.781	Node 11: -0.827	Node 11: -0.752	Node 11: -0.840	Node 11: -0.727
Node 12: -0.400	Node 12: -0.331	Node 12: -0.327	Node 12: -0.314	Node 12: -0.351	Node 12: -0.331
Node 13: -0.453	Node 13: -0.484	Node 13: -0.422	Node 13: -0.447	Node 13: -0.507	Node 13: -0.474
Node 14: -0.529	Node 14: -0.508	Node 14: -0.581	Node 14: -0.533	Node 14: -0.496	Node 14: -0.497
Node 15: -0.542	Node 15: -0.507	Node 15: -0.520	Node 15: -0.545	Node 15: -0.572	Node 15: -0.560
Node 16: -0.496	Node 16: -0.515	Node 16: -0.518	Node 16: -0.433	Node 16: -0.492	Node 16: -0.434
Node 17: -0.445	Node 17: -0.518	Node 17: -0.485	Node 17: -0.461	Node 17: -0.438	Node 17: -0.452
Bias: 1.620	Bias: 1.684	Bias: 1.661	Bias: 1.517	Bias: 1.675	Bias: 1.567

Node 7 (Sigmoid)	Node 8 (Sigmoid)	Node 9 (Sigmoid)	Node 10 (Sigmoid)	Node 11 (Sigmoid)	Node 12 (Sigmoid)
Node 1: -0.666	Node 1: -0.695	Node 1: -0.701	Node 1: -0.667	Node 1: -0.705	Node 1: -0.645
Node 2: -0.480	Node 2: -0.473	Node 2: -0.444	Node 2: -0.520	Node 2: -0.497	Node 2: -0.483
Node 3: -0.771	Node 3: -0.844	Node 3: -0.823	Node 3: -0.719	Node 3: -0.817	Node 3: -0.796
Node 4: -0.123	Node 4: -0.160	Node 4: -0.153	Node 4: -0.228	Node 4: -0.189	Node 4: -0.149
Node 5: -0.618	Node 5: -0.572	Node 5: -0.641	Node 5: -0.526	Node 5: -0.618	Node 5: -0.524
Node 6: -0.605	Node 6: -0.592	Node 6: -0.556	Node 6: -0.547	Node 6: -0.617	Node 6: -0.540
Node 7: -0.401	Node 7: -0.443	Node 7: -0.425	Node 7: -0.359	Node 7: -0.388	Node 7: -0.361
Node 8: -0.530	Node 8: -0.545	Node 8: -0.463	Node 8: -0.484	Node 8: -0.495	Node 8: -0.430
Node 9: -0.620	Node 9: -0.573	Node 9: -0.605	Node 9: -0.531	Node 9: -0.619	Node 9: -0.583
Node 10: -0.597	Node 10: -0.611	Node 10: -0.523	Node 10: -0.501	Node 10: -0.552	Node 10: -0.529
Node 11: -0.770	Node 11: -0.806	Node 11: -0.793	Node 11: -0.792	Node 11: -0.798	Node 11: -0.765
Node 12: -0.369	Node 12: -0.376	Node 12: -0.361	Node 12: -0.369	Node 12: -0.344	Node 12: -0.343
Node 13: -0.415	Node 13: -0.456	Node 13: -0.415	Node 13: -0.419	Node 13: -0.453	Node 13: -0.480
Node 14: -0.564	Node 14: -0.522	Node 14: -0.540	Node 14: -0.545	Node 14: -0.482	Node 14: -0.502
Node 15: -0.545	Node 15: -0.545	Node 15: -0.506	Node 15: -0.481	Node 15: -0.507	Node 15: -0.462
Node 16: -0.504	Node 16: -0.542	Node 16: -0.447	Node 16: -0.432	Node 16: -0.465	Node 16: -0.422
Node 17: -0.492	Node 17: -0.510	Node 17: -0.508	Node 17: -0.415	Node 17: -0.480	Node 17: -0.458
Bias: 1.686	Bias: 1.760	Bias: 1.622	Bias: 1.469	Bias: 1.668	Bias: 1.450

Node 13 (Sigmoid)	Node 14 (Sigmoid)	Node 15 (Sigmoid)	Node 16 (Sigmoid)	Node 17 (Sigmoid)
Node 1: -0.610	Node 1: -0.641	Node 1: -0.609	Node 1: -0.618	Node 1: -0.684
Node 2: -0.500	Node 2: -0.461	Node 2: -0.446	Node 2: -0.521	Node 2: -0.450
Node 3: -0.762	Node 3: -0.793	Node 3: -0.764	Node 3: -0.757	Node 3: -0.809
Node 4: -0.132	Node 4: -0.224	Node 4: -0.157	Node 4: -0.172	Node 4: -0.129
Node 5: -0.571	Node 5: -0.567	Node 5: -0.539	Node 5: -0.541	Node 5: -0.634
Node 6: -0.547	Node 6: -0.517	Node 6: -0.569	Node 6: -0.592	Node 6: -0.588
Node 7: -0.415	Node 7: -0.420	Node 7: -0.436	Node 7: -0.359	Node 7: -0.379
Node 8: -0.454	Node 8: -0.458	Node 8: -0.475	Node 8: -0.518	Node 8: -0.471
Node 9: -0.597	Node 9: -0.614	Node 9: -0.558	Node 9: -0.627	Node 9: -0.533
Node 10: -0.535	Node 10: -0.577	Node 10: -0.536	Node 10: -0.524	Node 10: -0.547
Node 11: -0.736	Node 11: -0.774	Node 11: -0.808	Node 11: -0.806	Node 11: -0.792
Node 12: -0.388	Node 12: -0.407	Node 12: -0.385	Node 12: -0.366	Node 12: -0.336
Node 13: -0.413	Node 13: -0.418	Node 13: -0.442	Node 13: -0.491	Node 13: -0.464
Node 14: -0.500	Node 14: -0.478	Node 14: -0.548	Node 14: -0.523	Node 14: -0.542
Node 15: -0.538	Node 15: -0.525	Node 15: -0.476	Node 15: -0.551	Node 15: -0.520
Node 16: -0.420	Node 16: -0.507	Node 16: -0.529	Node 16: -0.482	Node 16: -0.513
Node 17: -0.415	Node 17: -0.433	Node 17: -0.422	Node 17: -0.446	Node 17: -0.504
Bias: 1.473	Bias: 1.576	Bias: 1.535	Bias: 1.614	Bias: 1.620

Figure 4.50 The Result of Hidden layer 3 of ANN from RapidMiner Studio 7.4

Output
======

Class 'M' (Sigmoid)	Class 'B' (Sigmoid)
Node 1: -1.804	Node 1: 1.851
Node 2: -1.875	Node 2: 1.841
Node 3: -1.810	Node 3: 1.885
Node 4: -1.778	Node 4: 1.774
Node 5: -1.882	Node 5: 1.825
Node 6: -1.817	Node 6: 1.790
Node 7: -1.877	Node 7: 1.833
Node 8: -1.905	Node 8: 1.868
Node 9: -1.833	Node 9: 1.828
Node 10: -1.725	Node 10: 1.791
Node 11: -1.808	Node 11: 1.889
Node 12: -1.761	Node 12: 1.757
Node 13: -1.753	Node 13: 1.770
Node 14: -1.808	Node 14: 1.806
Node 15: -1.812	Node 15: 1.771
Node 16: -1.814	Node 16: 1.827
Node 17: -1.851	Node 17: 1.811
Threshold: 4.622	Threshold: -4.622

Figure 4.51 The Result of Output Layer of ANN from RapidMiner Studio 7.4

accuracy: 95.08% +/- 3.02% (mikro: 95.08%)

	true M	true B	class precision
pred. M	198	14	93.40%
pred. B	14	343	96.08%
class recall	93.40%	96.08%	

Figure 4.52 The Confusion Matrix of the ANN

4.10 The Performance of Support Vector Machine (SVM)

This classifier has conducted to evaluate the performance of 54 combinations which customize four factors of k-fold, kernel type, gamma, and C-value. The classification result has shown the highest percentage of accuracy with 96.84% which declared the error of RMSE with 0.195, and the classification lead time with 0.52 sec. Nonetheless, the classification result of parameter customizations can be seen in Figure 4.53-4.56 which shown the highest performance in the gray color as can be seen in Figure 76. Furthermore, the classification result of this method has shown in the RapidMiner

studio 7.4 which separate into two categories of confusion matrix and kernel model as can be seen in Figure 4.57-4.58 respectively.

No	k-fold	Kernel type	Gamma	C-Value	%Acc	RMSE	Lead time (sec)
1	5	RBF	0.0	0	64.68	0.432	0.49
2	5	RBF	0.0	50	95.25	0.253	0.46
3	5	RBF	0.0	100	95.07	0.236	0.50
4	5	RBF	0.1	0	95.42	0.261	0.56
5	5	RBF	0.1	50	95.96	0.189	0.52
6	5	RBF	0.1	100	95.78	0.187	0.50
7	5	RBF	0.2	0	95.42	0.247	0.52
8	5	RBF	0.2	50	95.78	0.196	0.49
9	5	RBF	0.2	100	95.78	0.199	0.47
10	5	Poly	0.0	0	95.07	0.249	0.46
11	5	Poly	0.0	50	62.74	0.494	0.53
12	5	Poly	0.0	100	62.74	0.493	0.56
13	5	Poly	0.1	0	95.07	0.249	0.47
14	5	Poly	0.1	50	95.07	0.224	0.38
15	5	Poly	0.1	100	96.13	0.212	0.40
16	5	Poly	0.2	0	95.07	0.249	0.49
17	**5**	**Poly**	**0.2**	**50**	**96.84**	**0.195**	**0.520**
18	5	Poly	0.2	100	95.61	0.198	0.43
19	5	Linear	NA	0	94.72	0.242	0.75
20	5	Linear	NA	50	94.90	0.184	0.54
21	5	Linear	NA	100	94.72	0.187	0.59
22	5	Sigmoid	0.0	0	94.90	0.242	0.45
23	5	Sigmoid	0.0	50	94.90	0.278	0.48
24	5	Sigmoid	0.0	100	95.25	0.253	0.51
25	5	Sigmoid	0.1	0	95.07	0.243	0.52
26	5	Sigmoid	0.1	50	95.60	0.196	0.43
27	5	Sigmoid	0.1	100	94.90	0.188	0.39
28	5	Sigmoid	0.2	0	94.02	0.254	0.52
29	5	Sigmoid	0.2	50	88.75	0.316	0.46
30	5	Sigmoid	0.2	100	88.39	0.328	0.50
31	10	RBF	0.0	0	65.03	0.431	0.76
32	10	RBF	0.0	50	95.78	0.252	0.56
33	10	RBF	0.0	100	94.91	0.236	0.56
34	10	RBF	0.1	0	95.78	0.256	0.63
35	10	RBF	0.1	50	96.14	0.185	0.56
36	10	RBF	0.1	100	95.61	0.188	0.55
37	10	RBF	0.2	0	95.26	0.243	0.63
38	10	RBF	0.2	50	95.79	0.196	0.60
39	10	RBF	0.2	100	95.26	0.199	0.65
40	10	Poly	0.0	0	95.25	0.245	0.54
41	10	Poly	0.0	50	62.74	0.494	0.66
42	10	Poly	0.0	100	62.74	0.494	0.76
43	10	Poly	0.1	0	95.25	0.245	0.46
44	10	Poly	0.1	50	95.79	0.218	0.46
45	10	Poly	0.1	100	95.96	0.206	0.52
46	10	Poly	0.2	0	95.25	0.245	0.42
47	10	Poly	0.2	50	96.49	0.192	0.63
48	10	Poly	0.2	100	95.96	0.196	0.59
49	10	Sigmoid	0.0	0	95.08	0.239	0.55
50	10	Sigmoid	0.0	50	95.43	0.276	0.75
51	10	Sigmoid	0.0	100	95.78	0.252	0.60
52	10	Sigmoid	0.1	0	95.43	0.240	0.51
53	10	Sigmoid	0.1	50	95.61	0.190	0.56
54	10	Sigmoid	0.1	100	94.73	0.190	0.46
55	10	Sigmoid	0.2	0	93.86	0.253	0.61
56	10	Sigmoid	0.2	50	88.76	0.318	0.63
57	10	Sigmoid	0.2	100	88.41	0.329	0.53
58	10	Linear	NA	0	94.56	0.189	0.56
59	10	Linear	NA	50	94.91	0.189	0.58
60	10	Linear	NA	100	94.91	0.189	0.63

Figure 4.53 The Performance Evaluation of SVM Technique

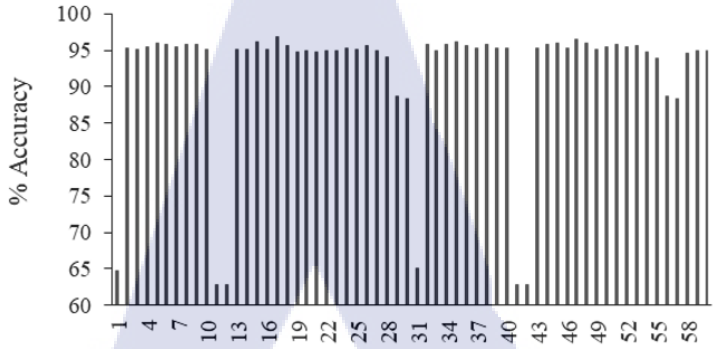


Figure 4.54 The Accuracy Result of the SVM

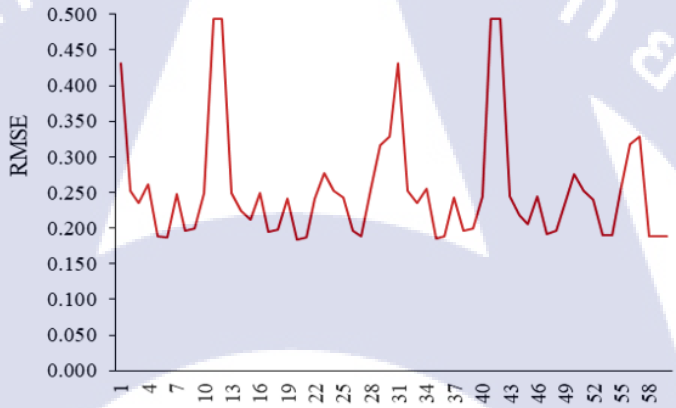


Figure 4.55 The RMSE Result of SVM

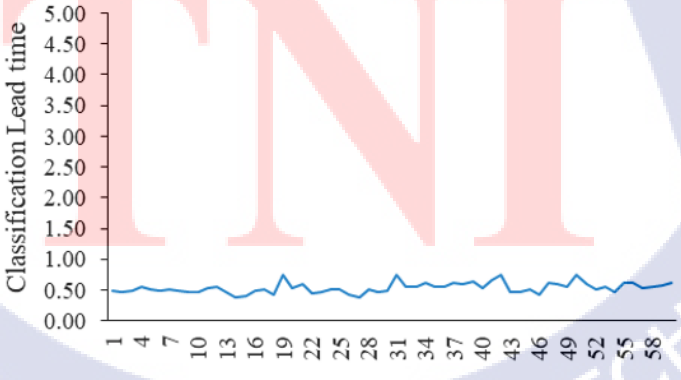


Figure 4.56 The Classification Lead Time of SVM

accuracy: 96.84% +/- 0.70% (mikro: 96.84%)

	true M	true B	class precision
pred. M	200	6	97.09%
pred. B	12	351	96.69%
class recall	94.34%	98.32%	

Figure 4.57 The Confusion Matrix of SVM

Kernel Model

Total number of Support Vectors: 76 Bias (offset): 3.161	
w[NRadius1] = 438.180	w[NCompactness2] = 195.483
w[NTexture1] = 436.390	w[NConcavity2] = 100.560
w[NPerimeter1] = 411.255	w[NConcave_points2] = 264.934
w[NArea1] = 270.478	w[NSymmetry2] = 182.505
w[NSmoothness1] = 418.936	w[NFractal_dimension2] = 88.563
w[NCompactness1] = 276.411	w[NRadius3] = 364.272
w[NConcavity1] = 215.758	w[NTexture3] = 526.663
w[NConcave_points1] = 269.599	w[NPerimeter3] = 352.232
w[NSymmetry1] = 414.878	w[NArea3] = 200.158
w[NFractal_dimension1] = 242.997	w[NSmoothness3] = 448.743
w[NRadius2] = 128.466	w[NCompactness3] = 238.906
w[NTexture2] = 218.693	w[NConcavity3] = 247.494
w[NPerimeter2] = 97.361	w[NConcave_points3] = 515.854
w[NArea2] = 56.176	w[NSymmetry3] = 292.349
w[NSmoothness2] = 164.955	w[NFractal_dimension3] = 204.776
number of classes: 2 number of support vectors for class M: 32 number of support vectors for class B: 44	

Figure 4.58 The Kernel Model of the SVM

4.11 The Performance Comparison

The research objective has compared the classification performance of each machine learning technique which predict the result of a breast cancer dataset. Nonetheless, the performance comparison of this research can be divided into three categories of accuracy rate, error rate, and classification lead time as follows.

4.11.1 Accuracy rate

The accuracy comparison of the classification result can be considered by the percentage of accuracy and the F-measure score that decided from the percentage of precision and recall. However, all accuracy rate can be calculated from the confusion matrix as can be seen in table 4.2.

Table 4.2 Confusion Matrix of Each Machine Learning Technique

DT	True M	True B	Class Precision
Predict M	195	12	94.20%
Predict B	17	345	95.30%
Class Recall	91.98%	96.64%	
NB	True M	True B	Class Precision
Predict M	186	18	91.18%
Predict B	26	339	92.88%
Class Recall	87.74%	94.96%	
ANN	True M	True B	Class Precision
Predict M	198	14	93.40%
Predict B	14	343	96.08%
Class Recall	93.40%	96.08%	
SVM	True M	True B	Class Precision
Predict M	200	6	97.09%
Predict B	12	351	96.69%
Class Recall	94.34%	98.32%	

Based on the confusion matrix result, this research can calculate and compare the accuracy rate of breast cancer data classification which can be seen in Figure 4.59-4.60.

Finally, the accuracy comparison shown the support vector machine technique is the highest performance which the accuracy percentage of 96.84%, the F-measure (M) score of 95.70%, and F-measure (B) score of 97.50% followed by ANN, Decision tree, and Naive Bayes with the accuracy percentage of 95.08%, 94.90%, and 92.26% respectively.

Machine learning Techniques	Accuracy	Precision (M)	Recall (M)	F-measure (M)	Precision (B)	Recall (B)	F-measure (B)
Decision Tree	**94.90%**	94.20%	91.98%	**93.08%**	95.30%	96.64%	**95.97%**
Naïve Bayes	**92.26%**	91.18%	87.74%	**89.43%**	92.88%	94.96%	**93.91%**
Artificial Neural Network	**95.08%**	93.40%	93.40%	**93.40%**	96.08%	96.08%	**96.08%**
Support Vector Machine	**96.84%**	97.09%	94.34%	**95.70%**	96.69%	98.32%	**97.50%**

Figure 4.59 The Accuracy Rate Comparison Result

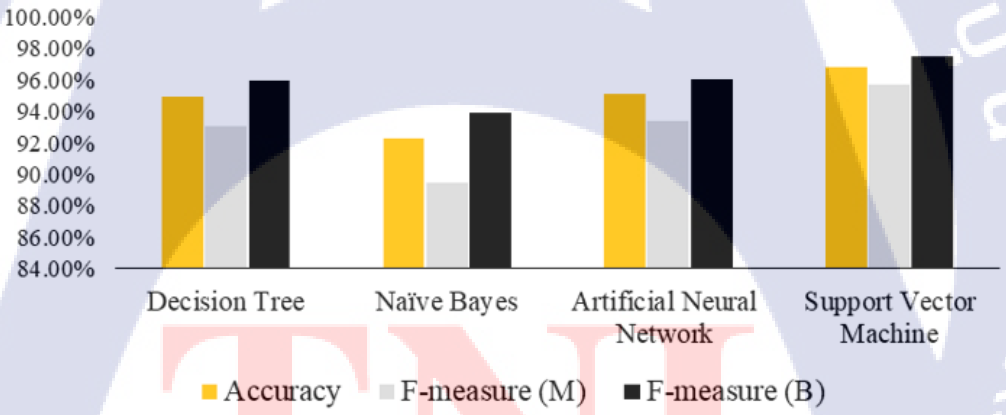


Figure 4.60 The Accuracy Rate Comparison Graph

4.11.2 Error rate

The error comparison of the classification result can be decided by the root mean square error that calculated from the confusion matrix. It can be seen in Figure 4.61-4.62. Nonetheless, the root means square error comparison shown that the lowest error

is an artificial neural network technique with 0.194 followed by SVM, Decision tree, and Naive Bayes with 0.195, 0.215, and 0.258 respectively.

Machine learning Techniques	Root Mean Square Error (RMSE)
Decision Tree	0.215
Naïve Bayes	0.258
Artificial Neural Network	0.194
Support Vector Machine	0.195

Figure 4.61 The RMSE Comparison Result

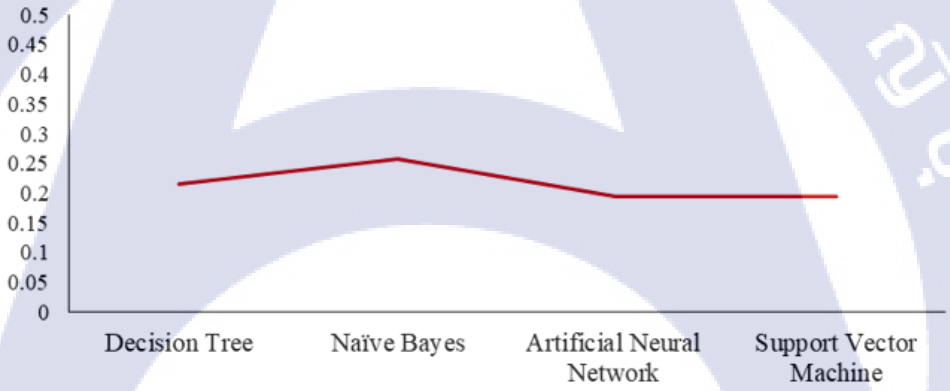


Figure 4.62 The RMSE Comparison Graph

4.11.3 Classification lead time comparison

The classification lead time of this research is conducted by count the using time that each classifier spends on the prediction process. Ultimately, this research found two techniques which are the shortest lead time of the prediction process of support vector machine and Naive Bayes with 0.52 sec followed by Decision tree and ANN with 1.53 sec and 38.89 sec respectively. It can be seen in Figure 4.63-4.64

Machine learning Techniques	Classification Lead Time (sec)
Decision Tree	1.53
Naïve Bayes	0.52
Artificial Neural Network	38.89
Support Vector Machine	0.52

Figure 4.63 The Classification Lead Time Comparison Result

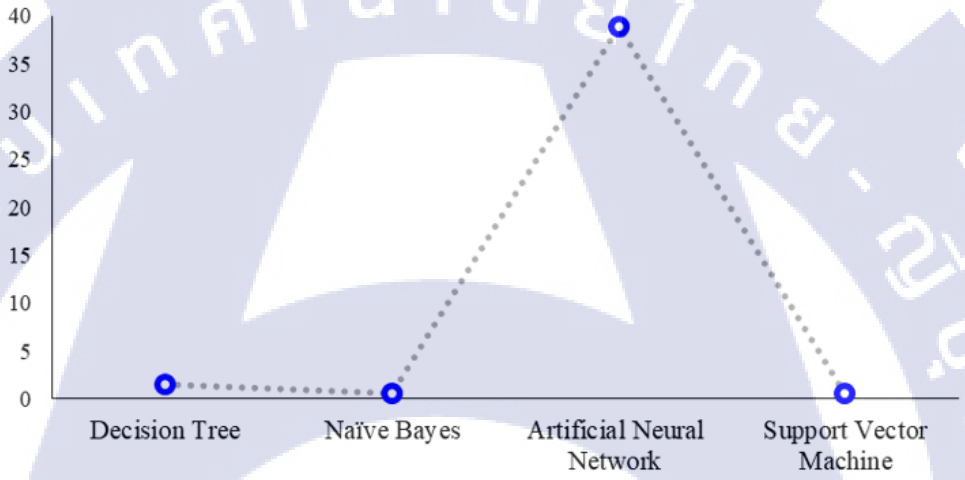


Figure 4.64 The Classification Lead Time Comparison Graph

Chapter 5
Conclusion

5.1 Conclusions

This research has proposed to compare the performance of machine learning techniques through data classification. Subsequently, the breast cancer dataset is selected from the UCI data repository which contains 30 attributes of breast cell nucleus feature and the classification type is the binary problem with 569 instances. Furthermore, this research has selected the machine learning algorithm to perform the data classification which can be divided into four techniques of the Decision Tree, Naive Bayes, Artificial Neural Network, and Support Vector Machine. Moreover, this research has randomly customized the parameters of each algorithm for finding the highest result which using the full combination method thus evaluate the performance by cross-validation technique. In addition, the performance evaluation of data classification is conducted via the RapidMiner studio 7.4 program. Nonetheless, the objective of this research is finding the appropriate machine learning that can provide the highest performance of data classification which can be compared into 3 performance of accuracy rate, error rate, and classification lead time as can be seen as follows.

First, this research has compared the accuracy result of the data classification of each machine learning technique which considers two performance of the percentage of accuracy and the percentage of F-measure. However, the highest accuracy rate is the support vector machine technique which shown the accuracy percentage of 96.84%, the F-measure(M) percentage of 95.70%, and the F-measure(B) percentage of 97.50% followed by the artificial neural network, decision tree, and Naive Bayes with the accuracy percentage of 95.08%, 94.90%, and 92.26% respectively. Second, this research has proposed to compare the error rate of each machine learning prediction which considers by the root mean square error (RMSE) method. The lowest of RMSE is the artificial neural technique which shown the RMSE with 0.194 followed by the support vector machine, decision tree, and Naive Bayes with 0.195, 0.215, and 0.258 respectively. Finally, the

classification lead time is conducted to compare in this research which found two technique that shortest of classification lead time of support vector machine and Naive Bayes with 0.52 second followed by the decision tree and artificial neural network with 1.53 second and 38.89 second respectively.

In conclusion, the highest performance of data classification is the support vector machine technique which shown the highest accuracy percentage of 96.84%, small RMSE of 0.195, and shortest lead time of 0.52 second. Nonetheless, the four parameters of this technique are customized by performing to 5-fold validation and using the polynomial kernel type including to specific the gamma-value of 0.2 and c-value of 50. Moreover, another technique that provides the good classification performance is the decision tree technique which shown the accuracy percentage of 94.90%, the RMSE of 0.215, and classification lead time of 1.53 second. Even, the accuracy percentage of the decision tree technique is smaller than an artificial neural network (ANN) but this is a few different of percentages which calculate to 0.18%. On the other hand, the ANN technique is shown the longest of classification lead time thus this is unsuitable when would implement in the real.

5.2 Suggestions and Recommendations

This part has provided some suggestions and recommendations to improve the classification performance which perceived by the research result as follows.

5.2.1 The dataset is the one important thing that relates to the data classification. If can prepare the suitable of the dataset it might improve the classification performance. Because the dataset of this research is different in term of the number of class labeled thus impact to different the F-measure percentage between the classes as shown in the research result.

5.2.2 If increase the volume of the dataset it can make more confidence in the classification result which is become the big data analytics scale.

5.2.3 In term of the machine learning technique, if searching the algorithms that provide the higher accuracy and shortest of classification lead time which flexible to use

with other datasets it will encourage the decision-making activities such as ensemble technique or deep learning technique.

5.3 Future Work

5.3.1 The researcher has proposed to increase the numerical data classification performance of the support vector machine technique (SVM) technique which would find the significance parameter that impact to the numerical data classification performance result.

5.3.2 The future work has proposed to utilize the data classification method to encourage decision making in real. The researchers expect to conduct the big data analytics by analyzing the continuous data that relate which internet of thing technologies. Finally, the data can be updated in real time via IOT and performed analytics to support any activities that improve the human life.

References

References

[1] K. Kripibul and P. Kripibul, *"Breast Cancer,"* [Online]. Available: http://haamor.com/th/ Breast Cancer [Accessed: November 1, 2017].

[2] N. Borecky and L. Wylie, *"Breast Fine Needle Aspiration (FNA),"* [Online]. Available: https://www.insideradiology.com.au/breast-fna. [Accessed: December 20, 2017].

[3] Center for Machine Learning and Intelligent Systems, *"UCI Machine Learning Repository,"* [Online]. Available: https://archive.ics.uci.edu/ml/index.php. [Accessed: October 15, 2017].

[4] T. Theeramunkong, *"Introduction to Concepts and Techniques in Data Mining and Application to Text Mining,"* Bangkok: Thammasat University Press and Tana Press Co., Ltd, November 2012.

[5] P. Gupta, "*Decision Trees in Machine Learning,*" [Online]. Available: https://towardsdatascience.com/decision-trees-in-machine-learning. [Accessed: December 15, 2017].

[6] J. Brownlee, *"Naive Bayes Tutorial for Machine Learning,"* [Online]. Available: https://machinelearningmastery.com/naive-bayes-tutorial-for-machine-learning. [Accessed: January 8, 2018].

[7] J. Brownlee, *"Support Vector Machines for Machine Learning,"* [Online]. Available: https://machinelearningmastery.com/support-vector-machines-for-machine-learning. [Accessed: January 1, 2018].

[8] E. Pacharawongsakda, *"Practical Data Mining with RapidMiner Studio 7,"* Bangkok: Asia Digital Press Co., Ltd, February 2017.

[9] T. T. Wong and N. Y. Yang, "Dependency analysis of accuracy estimates in k-fold cross validation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, issue 11, pp. 1-12, November 2017.

[10] A. M. Llenas et al., "Performance evaluation of machine learning based signal classification using statistical and multiscale entropy features," *In the Proceeding of International Conference on Wireless Communications and Networking Conference*, WCNC 2017, pp. 1-5, San Francisco, CA, USA, March 19-22, 2017.

[11] V. Shanmugarajeshwari and R. Lawrance, "Analysis of students' performance evaluation using classification techniques," *In the Proceeding of International Conference on Wireless Communications and Networking Conference*, WCNC 2016, pp.1-6, Kovilpatti, India, January 7-9, 2016.

[12] P. Kumar et al., "Analysis of various machine learning algorithms for enhanced opinion mining using twitter data streams," *In the Proceeding of International Conference on Micro-Electronics and Telecommunication Engineering*, ICMETE 2016, pp. 1-6, Ghaziabad, India, September 22-23, 2016.

[13] A. S. Rani and S. Jyothi, "Performance analysis of classification algorithms under different datasets," *In the Proceeding of International Conference on Computing for Sustainable Global Development*, INDIACom 2016, pp. 1584-1589, New Delhi, India, March 16-18, 2016.

[14] R. Duriqi et al., "Comparative analysis of classification algorithms on three different datasets using WEKA," *In the Proceeding of Mediterranean Conference on Embedded Computing*, MECO 2016, pp. 335-338, Montenegro, Monaco, June 12-16, 2016.

[15] G. Gaikwad and D. J. Joshi, "Multiclass mood classification on twitter using lexicon dictionary and machine learning algorithms," *In the Proceeding of International Conference on Inventive Computation Technologies*, ICICT 2016, pp 1-6, Coimbatore India, August 26-27, 2016.

[16] S. Z. Mishu and S. M. R. Uddin, "Performance analysis of supervised machine learning algorithms for text classification," *In the Proceeding of International Conference on Computer and Information Technology*, ICCIT 2016, pp. 409-413, Dhaka, Bangladesh, December 18-20, 2016.

[17] Z. Aysun et al., "Performance evaluation of classification algorithms by excluding the most relevant attributes for dipper/non-dipper pattern estimation in type-2 DM patients," *In the Proceeding of International Conference on Intelligent Systems Design and Applications*, ISDA 2015, pp 16-18, Marrakech, Morocco, December 14-16, 2015.

[18] T. Verma et al., "Tokenization and filtering process in RapidMiner," *International Journal of Applied Information Systems*, vol.7, no.2, pp. 16-18, April 2014.

[19] K. Choudhary et al., “Glaucoma detection using cross validation algorithm: a comparative evaluation on RapidMiner,” *In the Proceeding of IEEE Conference on Norbert Wiener in the 21st Century*, 21CW 2014, pp. 1-5, Boston, MA, USA, June 24-26, 2014.

Appendixes

Publications

A. An Appraisement of Human Happiness Level based on Air Quality through Fuzzy Logic Inference System: In the Proceeding of IEEE 15th International Conference on Industrial Informatics (INDIN), University of Applied Science Emden/Leer, Emden, Germany.

An Appraisement of Human Happiness Level based on Air Quality through Fuzzy Logic Inference System

Nattaphon Talmongkol, Noppadon Pongpisuttinun, and Wimol San-Um
Intelligent Electronics Systems Research Laboratory
Master of Engineering Technology Program
Faculty of Engineering, Thai-Nichi Institute of Technology, Bangkok, Thailand
Tel: (+66)-2-763-2600, E-mail Address: wimol@tni.ac.th

***Abstract*—**this paper proposes an appraisement of Human Happiness Level (*HHL*) based on Air Quality (*AQ*) through the use of Fuzzy Logic Inference System (*FIS*) system. Such a *FIS* has recently been realized as a potential alternative to that of conventional AQ assessment based on standard scales, which encounter practical difficulties on complicated measures of various hazardous attributes and levels of chemical compounds in air. Despite the fact that several fuzzy-based AQ models have been reported, no fully-developed AQ model that considers from physical attributes to human actualization, especially happiness, has been reported. In this paper, five parameters affecting on human health, involving $PM_{2.5}$, *CO*, temperature, humidity, and air pressure, have been considered as inputs for the proposed fuzzy-based AQ model. The linguistic variables for those five parameters were initially determined based on *US EPA* 2016 and *NOAA* national weather service, and the input membership functions were correspondingly created. The experimental results were performed using MATLAB Toolbox 2016a. The particular set of 109 rules was employed for decision processes, and the defuzzification provides output values for the designated output membership functions, providing six human happiness levels, involving happy, comfort, unhappy, depress, sad, and feel sick. This paper offers a perspective on a correlation between *AQ* and *HHL* through a FIS, ultimately leading to healthy and happy livings.

Keywords- Human Happiness Level; Fuzzy Logic Inference System; Air Quality; Healthy and Happy Livings.

I. Introduction

The assessment of air quality has recently attacked much intention for occupational health research as air quality is one of a major health factor, which affects physical, physiological and biochemical systems of human, leading to attention and work effectiveness, and resulting in longevity and happiness. A conventional air quality assessment approach based on standard scales has long been utilized, reflecting to health statuses in six categories, i.e. good, moderate, unhealthy for sensitive groups, unhealthy, very unhealthy, and hazardous. Nonetheless, such a conventional assessment approach cannot be realized utilized to classify health statuses distinctly due to some practical difficulties, involving (i) an inappropriate selection of gas sensors for particular circumstances, (ii) complexity of air measurement procedures and environments, (iii) calculation methods, and (iv) data interpretation and visualization.

Consequently, a fuzzy system, which refers to a knowledge-based logic concept that cannot be expressed explicitly as either "True" or "False", has recently been of much interest for air quality assessment as an alternative to solve a complexity of various factors and chemical compounds in air and environments. Recently, several studies relating to *AQI* assessment using fuzzy systems have been reported, and hazardous attributes for *AQI* evaluation in previous publications are summarized in Table 1 where PM_x is a particulate matter, *CO* is Carbon Monoxide, SO_x is Sulfur oxides, NO_x is Nitrogen oxides, O_3 is Ozone, *VOC*s are Volatile Organic Compounds, *BTEX* are chemicals of benzene, toluene, ethylbenzene and xylene, and Bioaerosols involve fungi, bacteria, viruses, and pollen. Initially, it can be considered that most recent publications consider five major factors, i.e. PM_x, *CO*, SO_x, NO_x, and O_3. Subsequently, Table 2 correspondingly describes the impacts of such five major attributes as significant factors of air pollution that resiliently affects human health.

TABLE I. Summary of hazardous attributes for AQI evaluation in previous publications

Ref.	PM_x	*CO*	SO_x	NO_x	O_3	*VOC*s	*BTEX*	Bioaerosols
[1]	/	/	/	/	/	×	×	×
[2]	/	/	/	/	/	×	×	×
[3]	/	/	/	/	/	/	×	/
[4]	/	/	/	/	/	×	×	×
[5]	/	/	/	/	/	/	/	×
[6]	/	/	/	/	/	×	×	×
[7]	/	/	/	/	/	×	×	×
[8]	/	×	×	×	×	/	×	×

Notes: /=Yes, ×=No

With reference to Table 1, advanced fuzzy-based *AQI* models have particularly been studied with additional factors. *A. K. Gorai et al.* [1] have studied a weighted fuzzy inference system through the use of outdoor air pollutant parameters in order to determine a Fuzzy Air Quality Health Index (*FAQHI*). Three parameters, including Location Sensitivity, Population Density, and Population Sensitivity, were involved into the fuzzy system, and the result suggests that *FAQHI* can potentially estimate the air quality and hence the health impacts. *H. Sarkheil and S. Rahbari* [2] have compared Mamdani Fuzzy Air Quality Index (*MFAQI*) and Takagi–Sugeno Fuzzy Air Quality Index (*TSFAQI*) via five air pollutant parameters. The result reveals that both indices provide an accuracy of approximately 95%, but *TSFAQI* overestimates the *AQI* whilst the *MFAQI* underestimates the *AQI*. *Javid et al.* [3] have developed the novel model called Fuzzy-based Indoor Air Quality Index (*FIAQI*) that can be used as a mutual tool with the *AQI* of *US EPA* in order to

Five major AQ parameters	Importance on air pollution and Effects on Human Health [9-12]
(i) Particulate matters	Particle sizes inhaled into body directly affects serious problems on heart and lungs. Generally, particle pollution can be classified into PM_{10} and $PM_{2.5}$, which means particles smaller than 10 μm. and 2.5 μm, respectively.
(ii) Carbon monoxide (*CO*)	*CO* is a colorless, odorless, non-irritating gas. Inhalation of low level causes a serious problem for those who have cardiovascular disease, and remarkably higher level of *CO* can be poisonous.
(iii) Sulfur dioxide (SO_2)	SO_2 is a toxic and pungent gas that irritates smell. Inhalation of SO_2 may result in an increase in respiratory symptoms, difficulty in breathing, and premature death.
(iv) Nitrogen dioxide (NO_2)	NO_2 is a reddish-brown toxic gas which is a prominent air pollutant. Inhalation of NO_2 could possibly worsen respiratory diseases such as emphysema or bronchitis.
(v) Ozone (O_3)	O_3 is a pale blue with distinctively pungent smell. Inhalation of *O3* could activate various health problems, involving chest pain, coughing, throat irritation, congestion, bronchitis, emphysema, and asthma.

TABLE III. TYPICAL POLLUTANT-SPECIFIC SUB-INDICES OF AIR QUALITY INDEX [13]

$PM_{2.5}$ (μg./m^3) [24 Hours]	*CO* (ppm.) [8 Hours]	SO_2 (ppb.) [1 Hours]	NO_2 (ppm.) [1 Hours]	O_3 (ppm.) [1 Hours]	*AQI* Ranges	*AQI* Category
0 – 12.0	0 – 4.4	0 - 35	0 - 53	-	Up to 50	**Good**
12.1 – 35.4	4.4 – 9.4	36 - 75	54 - 100	-	51 - 100	**Moderate**
35.5 – 55.4	9.5 – 12.4	76 - 185	101 - 360	0.125 - 0.164	101 - 150	**Unhealthy for Sensitive Groups**
55.5 – 150.4	12.5 – 15.4	186 - 304	361 - 649	0.165 - 0.204	151 - 200	**Unhealthy**
150.5 – 250.4	15.5 – 30.4	305 – 604 [24-hour]	650 - 1249	0.205 - 0.404	201 - 300	**Very Unhealthy**
250.5 – 500.4	30.5 – 50.4	605 – 1004 [24-hour]	1250 - 2049	0.405 - 0.604	301 - 500	**Hazardous**

estimate the impact of both ambient and indoor air pollutants on human health.

In the case where fuzzy-based *AQI* was realized in some specific countries and area, *M.A. Olvera-Garcia et al.* [4] has proposed fuzzy inferences combined with an analytic hierarchy process for creating the *AQI* for assessing the Mexico City atmospheric system. This model introduces unique rules for fuzzy system based on parameter behaviors and five classes of *AQI*, involving excellent, good, regular, bad, and dangerous were evaluated. Meanwhile, *M.H. Sowlat el al.* [5] has developed the *AQI* based on a fuzzy inference system call *FAQI* for Tehran, Iran, which particularly includes *BTEX*, and compared to *US EPA AQI*. The system suggested a balance via different weighting factors of air environment and inference rules in fuzzy system. According to *US EPA* 2016*D*, *Mintz et al.* [6] has suggested technical assistances for the reporting the AQI, which concentrates on six different categories with an emphasis on concentration breakpoints, health and cautionary statements.

In Pardubice micro-region, Czech Republic., *P. Hajek and V. Olej* [7] have designed a hierarchical fuzzy inference system for air pollution assessment. The study also suggests that different areas of the world are characterized by different climatic conditions influencing the effect of atmospheric pollutants on human health. Finally, *M. N. Assimakopoulos el al.* [8] has studied a fuzzy logic assessment system on indoor AQ of the underground trains in Athens, Greece. The system also considers human comfort, including temperature, relative humidity and Number of passengers, and it was concluded that the fuzzy logic can be used as a practical tool for optimum management of air pollutants indoor.

Despite the fact that several fuzzy-based AQ models have been reported, no fully-developed AQ model based on fuzzy system that completely considers from physical attributes to human actualization, especially happiness, has been reported. Regardless the use of fuzzy system, *X. Zhang et al.*, [9] have ranged happiness variable from 0 to 4 based on data collected from China Family Panel Studies (CFPS) and daily air quality data, involving six main pollutants and weather conditions. Meanwhile, *C.L. Ambrey et al.* [10] have employed a life satisfaction approach to evaluating air pollution in South East Queensland. Emotional stability has been classified in seven degrees.

In this paper, five air quality parameters affecting on human health, involving $PM_{2.5}$, CO, temperature, humidity, and air pressure, will be considered as causes of human satisfaction. Based upon a hypothesis that the state of human mind is related to air quality parameters, this paper presents a new indicator called Human Happiness Level (*HHL*). The fuzzy-based *AQ* model will be employed for *HHL* evaluation.

II. FUZZY LOGIC INFERENCE SYSTEM

The conception is to provide *HHL* assessment through a Fuzzy Logic Inference System (*FIS*), which refers to as an expert system with an approximate reasoning process, which allows transforming several input vectors to a single scalar output [4]. The FIS is generally based on Set Theory developed for modeling of nonlinear, uncertain and complex systems. Two types of FIS available in MATLAB toolbox

[11] are Mamdani and Takagi–Sugeno algorithms, but Mamdani is commonly utilized for complex system and decision processes. Typically, the FIS system comprises four processes, i.e. (i) Fuzzification, (ii) Rule inference, (iii) Rule composition, and (iv) Defuzzification. First, the fuzzification process is initiated by a transformation of crisp values into linguistic terms. Membership functions (μ) subsequently transform those linguistic terms into scores. In other word, a membership function for a fuzzy set A on the universe of a discourse x is defined as $\mu_A(x) \rightarrow [0,1]$, where each element of X is mapped into a range from 0 to 1, i.e.

$$A = \left\{ x, \mu_A(x) \middle| x \in X \right\} \quad (1).$$

This paper realized Trapezoidal membership for non-statistical-based inputs while Gaussian membership function was used for statistical-based inputs with a mean (m) and a standard deviation (σ). The Trapezoidal membership function is given by

$$\mu_A(x) = \begin{cases} 0, (x < a) or (x > d) \\ \dfrac{x-a}{b-a}, a \le x \le b \\ 1, b \le x \le c \\ \dfrac{d-x}{d-c}, c \le x \le d \end{cases} \quad (2)$$

where a, b, c, and d are a lower limit, an upper limit, a lower support limit, and an upper support limit, respectively. Note that $a<b<c<d$. On other hand, the Gaussian membership function can mathematically be expressed as follows;

$$\mu_A(x) = \exp\left(\frac{-(x-m)^2}{\sigma^2}\right) \quad (3)$$

In addition to membership function expressions, fuzzy operators are necessarily utilized for processing the results of membership functions. In the case where two fuzzy sets A and B are computed, two common operators are Union and Intersection defined as follows;

$$\mu_{A \cup B}(x) = \max\left\{\mu_A(x), \mu_B(x)\right\} \quad (4)$$

$$\mu_{A \cap B}(x) = \min\left\{\mu_A(x), \mu_B(x)\right\} \quad (5)$$

Second, rule inference applies linguistic parameters as for describing antecedents and consequents. In other words, the FIS utilizes an "IF–THEN" rule-based system, where IF is the antecedent and THEN is the consequent. It should be noted that the FIS robustness significantly depends upon the number of defined rules. Third, the rule composition refers to a process of performing the inference operation on the fuzzy rules in approximate reasoning. Last, the defuzzification refers to a transforming process for producing quantifiable results in Crisp values and corresponding membership degrees. Generally, defuzzification methods include Centroid, Bisector, Smallest-of-Maximum (SOM), Middle-of-Maximum (MOM), and Largest-of-Maximum (LOM) [12].

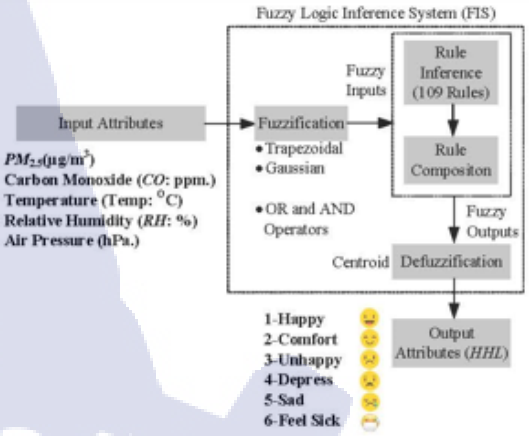


Figure 1. Research framework for *HHL* assessments based on air quality.

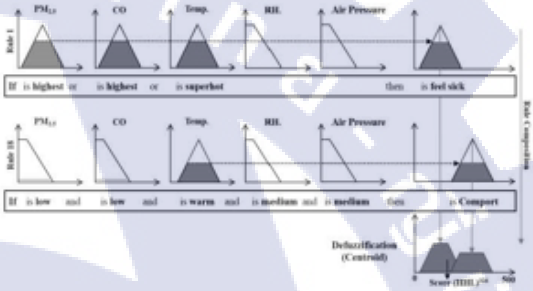


Figure 2. Fuzzy inference architecture for the *HHL* assessment with five parameters and two examples of rules 1 and 18 and all truncated functions are combined to determine the *HHL* using Centroid method.

TABLE IV. RANGES OF INPUT PARAMETERS

Parameters	Input Ranges	Units
$PM_{2.5}$	0-500	$\mu g/m^3$
CO	0-100,000	ppm.
Temp.	-40-120	°C
RH	0-100	%
Air Pressure	300-1,000	hPa

TABLE V. LINGUISTIC VARIABLES FOR EACH INPUT PARAMETER.

Parameters	Linguistic Variables					
$PM_{2.5}$ ($\mu g/m^3$)	Very Low	Low	Medium	High	Very High	Highest
CO (ppm.)	Very Low	Low	Medium	High	Very High	Highest
Temp. (°C)	Cool	Warm	Hot	Very hot	Super Hot	-
RH (%)	Low	Medium	High	Very High	-	-
Air Pressure (hPa)	Very Low	Low	Medium	High	-	-

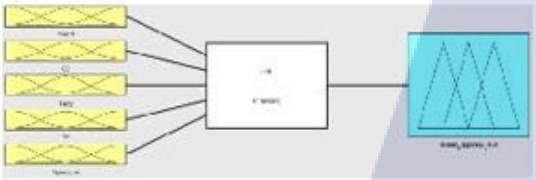


Figure 3. Overall fuzzy logic diagram of the proposed *HHL*.

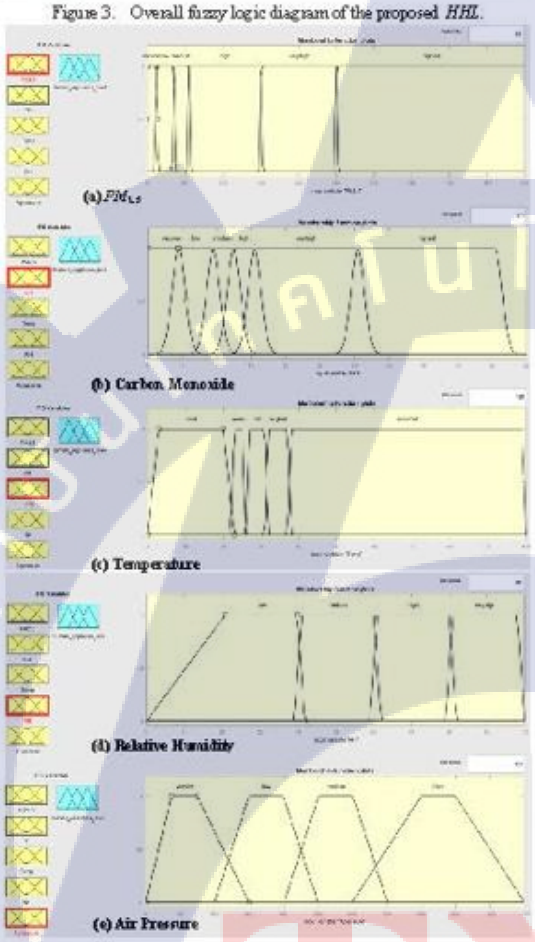


Figure 4. Membership function plots of each input parameters.

III. PROPOSED *FIS* FOR *HHL* ASSESSMENTS

Fig.1 illustrates research framework for *HHL* assessments based on air quality. It is seen in Fig.1 that there are five input attributes considered in this paper, including (i) Particulate matter ($PM_{2.5}$: $\mu g./m^3$), (ii) Carbon monoxide (*CO*: ppm.), (iii) Temperature (^0C), (iv) Relative Humidity (RH: %), and (v) Air pressure (hPa). Table 4 summarizes the ranges of the five input parameters. Table 5 determines linguistic variables for each input parameter. Such five input attributes are processed in the fuzzification stage through membership functions. As for illustration, Fig. 2 illustrates fuzzy inference architecture for the *HHL* assessment with five parameters and two examples of rules 1 and 18 and all truncated functions are combined to determine the *HHL* using Centroid Method. Trapezoidal member ship function was realized for temperature, relative humidity, and air pressure. On the other hand, Gaussian membership function was realized for PM_{25} and *CO* since these two parameters

TABLE VI. EXAMPLES OF RULES VERSUS LINGUISTIC VARIABLES, INDICATING HHL.

Rules	Linguistic Variables					*HHL*
	$PM_{2.5}$ ($\mu g/m^3$)	*CO* (ppm.)	Temp. (^0C)	*RH* (%)	Air Pressure (hPa)	
1	Highest	Highest	Super hot	-	-	**Feel Sick**
2	-	-	Very hot	Very High	-	**Feel Sick**
3	Very High	Very High	-	-	-	**Sad**
4	-	-	Very hot	High	-	**Sad**
5	High	High				**Depress**
6			Very hot	Medium	-	**Depress**
12	Low	Low	Hot	Low		**Unhappy**
13	-	-	-	-	Very Low	**Unhappy**
84	Very Low	Low	Cool	Very High	Medium	**Comfort**
85	Very Low	Low	Cool	Very High	High	**Comfort**
99	Very Low	Very Low	Cool	Low	Medium	**Happy**
100	Very Low	Very Low	Cool	Low	High	**Happy**

Note:
OR Operator
AND Operator

TABLE VII. PROPOSED DESCRIPTIONS AND EMOTIONAL INTERPRETATIONS OF HUMAN HAPPINESS LEVEL (*HHL*).

HHL	Descriptions and Emotional Interpretations [13-16]
1-Happy (1-50)	Self-Realization, Pleasure Attainment, Pain Avoidance, Living a Life of Virtue, Actualization of One's Inherent Potentials in the Pursuit of Complex and Meaningful Goals in both Individual and Society.
2-Comfort (51-100)	Pleasant State or Relaxed Feeling of a Human Being in Reaction to Environments.
3-Unhappy (101-150)	Not cheerful or Glad, Not Appropriate or Lucky.
4-Depress (151-200)	Irritable Mood, Loss of Pleasure, Disturbance in Falling Asleep, Poor Quality of Sleep, Feeling Tired, Changes in Weight.
5-Sad (201-300)	Emotional Pain, Feelings of Disadvantage, Loss, Despair, Grief, Helplessness, Disappointment and Sorrow, Suffering of Psychological or Non-Physical Origin.
6-Feel sick (301-500)	Feel ill, Feel Very Upset, and Affected with Disease or Ill Health.

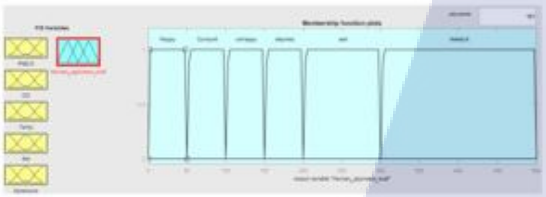


Figure 5. Output membership function plots of *HHL*.

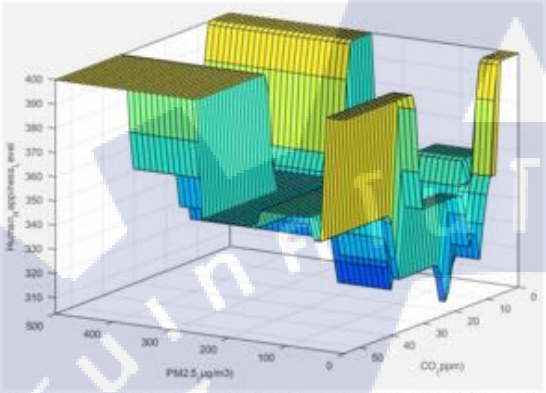


Figure 6. A 3-dimentional view of *HLL* between particulate matter versus Carbon monoxide.

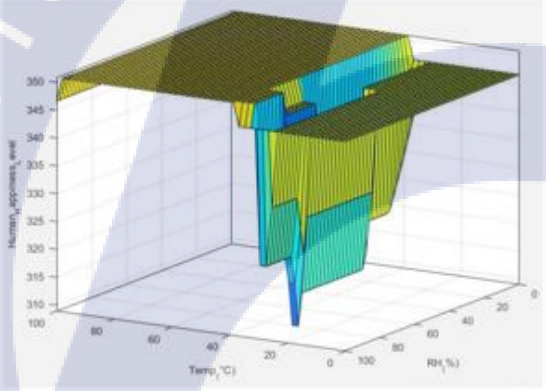


Figure 7. A 3-dimentional view of *HLL* between temperature versus relative humidity

have a range with average and standard deviation. Subsequently, the fuzzy inputs were operated based on 109 rules with OR and AND operators.

IV. Simulation Results

Simulations have been performed using MATLAB Toolbox 2016a. Table 6 demonstrates some examples the linguistic rules versus linguistic variables, indicating all six levels of *HHL*. Rule composition was performed and the fuzzy outputs were then proceeded in defuzzification stage through centroid method. Table 7 describes and interprets

Figure 8. The developed board for *HHL* assessments.

Figure 9. the preliminary design of the mobile application user interface with raw data of the proposed five parameters affecting on human health.

state of mind, emotions, and health statuses for each level of HHL, and the output attributes in terms of *HHL* are happy, comfort, unhappy, depress, sad, and feel sick. Fig.3 shows an overall fuzzy logic diagram of the proposed *HHL*. Fig.4 correspondingly shows membership function plots of each input parameters. Fig.5 finally shows an output membership function plots of *HHL*. Fig.6 illustrates a 3-dimentional view of *HHL* between particulate matters versus Carbon monoxide while Fig.7 shows a 3-dimentional view of *HHL* between temperatures versus relative humidity.

The circuit board for *HHL* assessment has been developed and shown in Fig. 8. The power supply is 5V and the Arduino was employed as a central processor. The HTS221 was used as a 16-bit humidity and temperature sensor. The gas sensor is PMS1003, which can detect $PM_{2.5}$ effectively. The MiCS-6814 is a robust MEMS sensor for the detection of pollution, especially for detecting CO in this paper. The pressure was measured by a digital pressure sensor BMP280. All sensors will transmit values via either a Bluetooth 4.0 module or an ESP8266 Wi-Fi module. Data will be stored in cloud sever. The future work is to implement the proposed

FIS for *HHL* assessment as a cloud computing and display the results on Android mobile application. Fig.9 demonstrates preliminary design of the mobile application user interface with raw data of the five parameters affecting on human health.

V. Conclusions

This paper has presented an appraisement of human happiness level based on air quality through the use of fuzzy logic inference system. Despite the fact that several fuzzy-based *AQ* models have been reported, no fully-developed *AQ* model that considers from physical attributes to ultimately human actualization, especially happiness, has been reported. In this paper, five parameters affecting on human health, involving $PM_{2.5}$, *CO*, temperature, humidity, and air pressure, have been considered as inputs for the fuzzy-based *AQ* model. The linguistic variables have been determined based on *US EPA* 2016 and *NOAA* national weather service. Simulation results were performed using MATLAB Toolbox 2016a. The particular set of 109 rules was employed with *OR* and *AND* operators. The Trapezoidal membership function was realized for temperature, relative humidity, and air pressure. The Gaussian membership function was realized for $PM_{2.5}$ and *CO*. A Centroid method was realized in defuzzification stage. Six human happiness levels have been suggested, involving happy, comfort, unhappy, depress, sad, and feel sick. This paper has thoroughly demonstrated membership plots, the 3-dimentional view of *HLL* between particulate matter versus Carbon monoxide, and the 3-dimentional view of *HLL* between temperatures versus relative humidity. The circuit board has been developed based on Arduino with various electronics sensors, including HTS221, PMS1003, MiCS-6814, BMP280, ESP8266, and a power supply module. Implementation on *FIS* on Android application is being developed for further use in real-world applications. In terms of human health aspects, *HHL*1 is important for those who are healthy and happy, and therefore any entities affecting people to feel sick should be diminished. It has been proven that *FIS* is useful to evaluate an air quality and appraise human happiness level, leading to healthy and happy livings of human kinds.

ACKNOWLEDGEMENTS

The authors are grateful to Thai-Nichi Institute of Technology and Piwat Air Co.Ltd. for financial supports. The authors are also would like to thank members of Intelligent Electronics Research (*IES*) Laboratory for kind cooperation and efforts on results discussions.

REFERENCES

[1] Amit Kumar Goraia, Kanchanb, Abhishek Upadhyayc, Francis Tulurid, Pramila Goyalc, Paul B. Tchounwoue, "An Innovative Approach for Determination of Air Quality Health Index", J. Science of the Total Environment, Vol.533, pp. 495-505, 2017.

[2] Sarkheil, Hamid, Rahbari, Shahrokh, "Development of Case Historical Logical Air Quality Indices via Fuzzy Mathematics (Mamdani and Takagi-Sugeno Systems) A Case Study for Shahre Rey Town", J. Environ Earth Sci, pp. 1-13, 2016.

[3] Allahbakhsh Javid, Amir Abbas Hamedian, Hamed Gharibi, Mohammad Hossein Sowlat, "Towards the Application of Fuzzy Logic for Developing a Novel Indoor Air Quality Index (FIAQI)", Iranian Journal of Public Health, Vol. 45, pp. 203-213, 2016.

[4] Miguel Ángel Olvera-García, José J. Carbajal-Hernández ,Luis P. Sánchez-Fernández, Ignacio Hernández-Bautista, "Air quality assessment using a weighted Fuzzy Inference System", J. Ecological Informatics, Vol. 33, pp. 57–74, 2016.

[5] Mohammad Hossein Sowlat, Hamed Gharibi, Masud Yunesian, Maryam Tayefeh Mahmoudi, Saeedeh Lotfi, "A novel, fuzzy-based air quality index (FAQI) for air quality assessment", J. Atmospheric Environment, Vol. 45, pp. 2050-2059, 2011.

[6] Rachel Mintz, Brent R. Young, William Y. Svrcek, "Fuzzy logic modeling of surface ozone concentrations", J. Computers and Chemical Engineering, Vol. 29, pp. 2049-2059, 2005.

[7] Petr Hájek, Vladimír Olej, "Air pollution assessment using hierarchical fuzzy inference systems", Scientific papers of the University of Pardubice, Series D, Faculty of Economics and Administration, Vol. 15 ,pp. 52-61, 2009.

[8] M.N. Assimakopoulos, A. Dounis, A. Spanou, M. Santamouris, "Indoor air quality in a metropolitan area metro using fuzzy logic assessment system", J. Science of the Total Environment, Vol. 449, pp. 461–469, 2013.

[9] Xin Zhang, Xiaobo Zhang, Xi Chen,"Valuing Air Quality Using Happiness Data: The Case of China" J. Ecological Economics, Vol. 137, pp. 29-36, 2017.

[10] Christopher L. Ambrey, Christopher M. Fleming, Andrew Yiu-Chung Chan, "Estimating the cost of air pollution in South East Queensland: An application of the life satisfaction non-market valuation approach" J. Ecological Economics, Vol. 97, pp. 172-181, 2014.

[11] Fuzzy Inference System Modeling - MATLAB & Simulink – MathWorks Manual.

[12] José Juan Carbajal-Hernández, Luis P. Sánchez-Fernández, Jesús A. Carrasco-Ochoa, José Fco. Martínez-Trinidad, "Assessment and prediction of air quality using fuzzy logic and autoregressive models", Atmospheric Environment 60, pp. 37-50, 2012.

[13] Antonella Delle Fave, Ingrid Brdar, Teresa Freire, Dianne Vella-Brodrick, Marié P. Wissing, "The eudaimonic and hedonic components of happiness: qualitative and quantitative findings", Social Indicator Research 100, pp. 185–207, 2011.

[14] R.M. Ryan and E.L. Deci "On happiness and human potentials: a review of research on hedonic and eudaimonic well-being", Annual Review on Psychology, Vol. 52, pp. 141-66, 2001.

[15] Peter Vink, "Comfort and Design: Principles and Good Practice", CRC Press, Taylor and Francis Group.

[16] Depression, UK National Institute for Health and Clinical Excellence (NICE), October, 2009.

B. Sentiment Analysis of Foreign Tourists to Bangkok using Data Mining through Online Social Network: In the Proceeding of IEEE 15th International Conference on Industrial Informatics (INDIN), University of Applied Science Emden/Leer, Emden, Germany.

Sentiment Analysis of Foreign Tourists to Bangkok using Data Mining through Online Social Network

Taweesak Kuhamanee*, Nattaphon Talmongkol*, Krit Chaisuriyakul*, Wimol San-Um*, Noppadon Pongpisuttinun*, Surapong Pongyupinpanich**

* Center of Excellence in Intelligent System Integrations
Thai-Nichi Institute of Technology, Bangkok, Thailand, 10250, Tel: (+66)-2-763-2600, E-mail Address: wimol@tni.ac.th
** Faculty of Engineering, Ramkhamhaeng University, Bangkok, 10240, Tel: (+66)-2-310-8570, Email: surapong@riees.org

***Abstract*—this paper presents an analysis of sentiment of foreign tourists to Bangkok, Thailand, using data mining approach through online social networks. The objective is to acquire information on sentiment of foreign tourists in order to improve and foster tourism industry of Bangkok. This paper has retrieved 10,000 datasets from Twitter in 2017. Such datasets were tokenized and filtered in order to obtain sentiment English words. Subsequently, the sentiment English words were purposely classified into five categories of visiting Bangkok, involving (i) Traveling, (ii) Business, (iii) Visiting Family, (iv) Education, and (v) Health and Treatments. It has revealed that the traveling purpose has the highest percentage of 71.93% followed by business and visiting family. Therefore, the sentiment of foreign tourists to traveling in Bangkok was analyzed through four approaches, i.e. (i) Decision Tree, (ii) Naïve Bayes, (iii) Support Vector Machine (SVM), and (iv) Artificial Neural Network (ANN), using RapidMiner Studio7.4. The results have shown that the foreign tourists visit in Bangkok mostly for nightlife activity, Thai culture, and shopping with percentages of 65.54%, 16.07%, and 13.61%, respectively, meanwhile temple and historical sites, Thai cuisine, and nature are not significant. The accuracy of sentiment analysis approaches of Decision Tree, Naïve Bayes, SVM, and ANN are 79.83%, 55.66%, 80.11%, and 80.33%, respectively. Based upon ANN approach that provides the highest accuracy, the positive sentiments were found to be a visit for nightlife activity, temple and historical sites, Thai cuisine, and nature. On the other hand, the negative sentiment was Thai culture while shopping is relatively neutral. This paper therefore suggests an acceleration of nightlife activity of Bangkok in order to foster tourism industry of Bangkok.**

Keywords-Sentiment Anlysis; Data Mining; Online Social Network; Foriegn Tourists;

I. Introduction

Thailand is a Southeast Asian country, and Bangkok is a capital city where a variety of places and activities have attracted foreign tourists to visit for several purposes such as business, traveling, health and treatments, visiting family, or even education. Recently, Thai government has launched an economic growth engine which includes a digital tourism economy. Therefore, information on the sentiment of foreign tourists will be useful for setting a tourism roadmap as a part of a digital tourism strategy. Although information on sentiment can be acquired by means of a poll through the use of traditional questionnaire and interview approaches, data analytics has been of much interested as one of potential alternative through Online Social Network (OSN) in smartphones or computers, involving, for instance Twitter, Facebook, Instagram, and Google Plus.

Based on a Digital, Social and Mobile in 2017 Report [1], active users of Twitter are 317 million people due to a function of rapid analytics and uncomplicated monitoring through all summarizing an opinion in 140 characters. Therefore, opinions from Twitter have been utilized as a major social media for sentiment analysis over other OSNs. Several studies regarding sentiment analysis have been reported in recent years. Shweta Rana *et al.* [2] have studied sentiment of movie reviewers for four movie types, i.e. action, adventure, drama and romantic, using Support Vector Machine (SVM) and Naïve Bayes approaches to compare the accuracy of sentiment by RapidMiner. Akshi Kumar and Ritu Rani [3] have proposed a Probabilistic Neural Network (PNN) with a self-adaptive approach to perform sentiment analysis. Two types of PNN were also introduced, i.e. PNNC and PNNS. The accuracy of PNNC is 95% while PNNS is 92%, suggesting that PNNC has a better performance than PNNS. Ankur Goel *et al.* [4] have implemented Naive Bayes using Sentiment140 for a high-speed training process, and also employed SentiWordNet in order to improve an accuracy of classification of tweets, resulting in an accuracy of 58.4%. Pierre Ficamos *et al.* [5] have retrieved data of Weibo for sentiment analysis using Naïve Bayes, and particularly suggested to rely on Part of Speech (POS) tags in order to extract unigrams and bigrams features. It can be consider that sentiment analysis can potentially be realized in various applications and purposes. In addition to sentiment analysis examples in [2-5], RapidMiner [6-8] has been recognized as one of advantageous data science software, which offers an integrated environment for machine learning, deep learning, text mining, and predictive analytics.

This work particularly presents sentiment analysis of foreign tourists to Bangkok, Thailand. The objective is to discover real purposes of visits as well as sentiment of foreign tourists in order to foster tourism industry of Bangkok. Datasets were retrieved from Twitter using RapidMiner Studio Version 7.4. The research procedure will initially classify a purpose of visit, and consequently emphasize on traveling in order to acquire information on sentiment, which may be positive, negative, and neutral.

II. Proposed Sentiment Analysis Approaches

A. Text Analytics

Text mining is a process of discovering information as forms, patterns, or trends, which are hidden in original text based on Statistics and Mathematics. In order to perform text mining for sentiment analystics, the theories related to an unstructural language are analysed due to the difference in nature of background of human learning skill and knowledgded. Therefore, the Natural Language Processing (NLP) [9] is importantly required for transforming unstructural to structural languages. Typically, NLP can be separated into three processes, i.e. (i) document summarization that reduces unnecessary texts, retaining the important points from original document, (ii) document classification is a process to assign classes or categories of documents in order to simplify further processes in terms of training, managing, and sorting, and (iii) document clustering is a grouping of documents with similar contents for fast topic extraction and information retrieval or filtering. The following sentiment analytics has been followed the three text mining process as described above in order to obtain an accuracy analytics results .

B. Sentiment Analytics Techniques

Generally, human attitudes and sentiments can be transformed into mathematical models by representing the numbers that indicates positive, negative, or neutral expression. Therefore, sentiment analytics based on OSNs has lately been one of attractive approaches in order to understand criticisms or perceptions on products or services. This research focuses on four techniques of machine learning techniques, involving (i) Decision Tree, (ii) Naïve Bayes, (iii) SVM, and (iv) ANN, for sentiment classification and accuracy evaluation as follows.

1) Naïve Bayes

Naïve Bayes technique is a family of probabilistic classifiers based on Bayes' theorem with independence assumptions among features [2,4,9]. The calculation of posterior probability is given by

$$p(\mathrm{b}|\mathrm{a}) = \frac{p(\mathrm{a}|\mathrm{b}) \times p(b)}{p(\mathrm{a})} \qquad (1)$$

where $p(\mathrm{b}|\mathrm{a})$ is the probability that class b occurs before class a, $p(\mathrm{a}|\mathrm{b})$ is the probability that class a occurs before class b, $p(\mathrm{a})$ is the probability of occurrence a, and $p(\mathrm{b})$ is the probability of occurrence b. Such a Naïve Bayes technique provides uncomplicated computation process as each distribution can be independently estimated as a one-dimensional distribution.

2) Decision Tree

A decision tree [10-11] is a decision support tool, which is a non-parametric supervised learning method commonly applied to classification and regression of multiple variable analyses. The decision tree has a tree-shaped diagram for representing a possible decision and consequences,

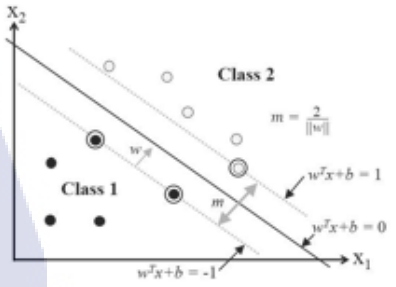


Figure 1. Demonstration of linear separating hyperplanes for the separable 2-dimentional case of SVM technique.

involving, for instance, chances, event outcomes, resource costs, and utilities. Typically, the decision tree can be constructed by Entropy (Ent) and Information Gain (IG). The Entropy is average number of binary questions which are in the form of infinitely trials to distinguish events, and can be calculated by

$$\mathrm{Ent}(c_i) = -p(c_i)\log_2 p(c_i) \qquad (2)$$

where $p(c_i)$ is a probability of dataset i=1,2,3..n. Generally, entropy is always nonnegative and is zero when one items c_i has a unity probability. The IG is the change in entropy from prior states to a state, and is based on the decrease in entropy after a dataset is split on an attribute. The IG can be found as follows;

$$\mathrm{IG} = \mathrm{Ent}(PR) - \{[p(c_1)\times \mathrm{Ent}(c_1)] + [p(c_2)\times \mathrm{Ent}(c_2)] + \cdots\} \qquad (3)$$

where Ent(PR) is an information entropy of overall datasets before splitting.

3) Support Vector Machine (SVM)

Support Vector Machine (SVM) [12-13] is principally a discriminative classifier that performs both regression and classification by constructing hyperplanes in a multi-dimensional space that separates cases of different classes. Generally, SVM offers effective in high dimensional spaces, and exploits less memory since a subset of training points in the decision function is realized. Moreover, SVM provides versatility in terms of Kernel functions types for any specific classification purposes. Fig.1 demonstrates linear separating hyperplanes for the separable 2-dimentional case. It can be seen from Fig.1 that the support vectors are highlighted with large circle. Intuitively, the decision boundary should be as far away from the data of both classes as possible. This property implies the maximization of the margin (m). With reference to Fig.1, given the training data $\{x_i,y_i\}$ for i=1,2,3,…,n, $x_i \in R^d$, $y_i \in \{-1,1\}$ where x_i is datum, representing by a vector with the d dimension and y is a binary class of -1 or +1, the support vector machine finds the best hyperplane which separate the positive from the negative example, i.e. a separating hyperplane. In principle, the points x on the hyperplane satisfy the formula $w^Tx+b=0$,

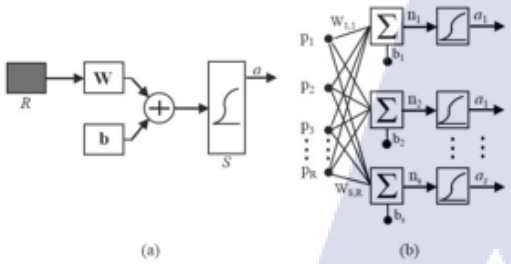


Figure 2. General ANN diagram; (a) a single neuron model, (b) a three-layer ANN structure.

where w is a normal vector, that is perpendicular to the hyperplane. $|b|/\|w\|$ is the perpendicular distance from the hyperplane to the origin, and $\|w\|$ is a Euclidean norm of w and b is a constant.

4) Artificial Neutral Network (ANN)

An artificial neuron network [14] is a computational system composed by highly interconnected processing elements based on the structure and functions of biological nervous systems. The ANN processes information through dynamic state response to external inputs and learning process. Fig.2 (a) and (b) show a single neuron model and a three-layer ANN, respectively. It is seen in Fig.2 (a) that the number of input element vectors R, which is weighted by a gain W, is combined with a bias b, and the combination result n is fed to an activation function f, which is a sigmoidal function for this case, providing the final result a. On the other hand, Fig.2 (b) illustrates a full diagram of the ANN composed by inputs, hidden, and outputs with a total of S layers. The generalized mathematical model of the ANN can be expressed as

$$a_S = f(W_{S,R} p_R + b_S) \qquad (4)$$

Generally, the ANN can be configured for several specific applications, such as pattern recognition, data classification, clustering, and prediction.

III. Proposed Sentiment Analysis

A. Research Methodology and Framework

The research methodology is an intelligent approach of sentiment data retrieval using RapidMiner Studio version 7.4. A number of 10,000 tweets in English language were randomly retrieved and summarized from those foreign tourists who express their opinions to Bangkok. The purposes of tourists have been classified into five categories, i.e. (i) traveling, (ii) business, (iii) visiting family, (iv) education, and (v) health and treatments.

Of particular interest in a category of traveling, remarkable places and activities in Bangkok are classified into nature, Thai culture, temple and historical sites, Thai cuisine, nightlife activity, and shopping. Based upon such remarkable places and activities mentioned above,

TABLE I. SOME EXAMPLES OF TWEETS OF EACH FOREIGN TOURIST.

No.	Tweets	Sentiment Analysis
1	*Bangkok buddha this was beautiful*	Positive
2	*A perfect stay at Mandarin Oriental Bangkok*	Positive
3	*Walking..the weather is amazing but is very hot*	Negative
4	*I don't understand the ir intent behind the hashtag trend # where is pappu?*	Negative
5	*Casting call for aspiring model siam center in Bangkok. Iamsiamish*	Neutral
6	*Bangkok thailand national museum. Last year I was at the National museum in Bangkok.*	Neutral

classifications of sentiment of those tourist activities can be divided into positive, negative and neutral, which are represented by a score of +1, -1, and 0, respectively. Table 1 provides some examples of sentiments of retrieved from each foreign tourist. Four machine learning techniques, involving Decision Tree, Naïve Bays, Artificial Neural Network (ANN), and Support Vector Machine (SVM), were employed for sentiment analysis. Additionally, the accuracy of the four machine learning techniques was compared in order to achieve the best analysis technique. Fig. 3 depicts the classification of visiting purposes of foreign tourists from tweets where the sentiment analysis realizes five major processes in consequences, i.e. data collection, text analysis and mining, and classification. In particular, five specific purposes were classified, including traveling, business, (iii) visiting family, education, and health and treatments.

Figure 3. Classification of visiting purposes of foreign tourists from tweets.

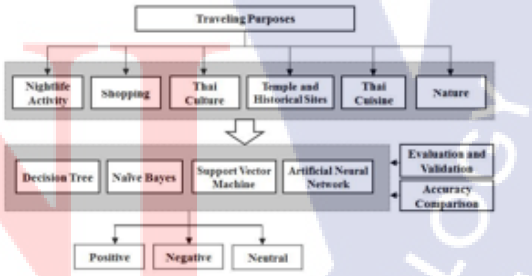


Figure 4. Research and methodology framework for Sentiment analysis of foreign tourist.

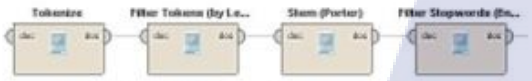


Figure 5. Block diagram of text processing operators.

B. Data Analytics Processes

Fig.6 shows the classification of visiting purposes of 10,000 tweets, which were randomly retrieved from foreign tourists. The retrieved data were subsequently transformed to be structural data prior to classification process. Fig. 4 shows research and methodology framework for sentiment analysis. Fig.5 demonstrates the block diagram of text processing operator. As can be seen from Fig.5, tokenization has initially been performed in order to transform original texts into phrases, words, symbols, or other meaningful elements generally called tokens. Such tokens are filtered by length, and subsequently converted into be stem words. Finally, stop words such as articles, e.g. a, an, the, were removed by a stop-word filtering operator.

In order to expose the importance of a word in a collection or corpus, a numerical statistic method called Term Frequency–Inverse Document Frequency (*tfidf*) [15] was realized in order to create a vector format for further segmentation process. Such a *tfidf* is a function of the number of times that term t occurs in document (d), and total number of datasets in the corpus (D). In other words, *tfidf* can mathematically be expressed as

$$tfidf_{(i,d,D)} = tf_{(i,d)} \times idf_{(i,D)} \quad (5)$$

where $tf_{(i,d)}$ is the number of times that a word occurs in each data set while $idf_{(i,D)}$ is a measure of an occurrence of a word that provides common or rare information across that of total number of documents. The $idf_{(i,D)}$ can be found through the logarithmically scaled inverse fraction as follows;

$$idf_{(i,D)} = \log\left|\frac{D}{\{d \in D, t \in d\}}\right| \quad (6)$$

Those significant words obtained from (5) were analyzed through four machine learning techniques mentioned earlier. The k-fold cross validation technique was employed for training and testing datasets. Typically, the dataset is divided into k subsets, and the holdout method is repeated k times. In this paper, the datasets were divided into 10 sets, where 9 sets were used for training and the remaining dataset was used for testing. The iteration process was performed 10 times. The measurement of sentiment classification accuracy in percentage can be calculated using

$$Accuracy(\%) = \left|\frac{True(Sentiment)}{True(Sentiment) + False(Sentiment)}\right| \times 100 \quad (7)$$

where *True*(*sentiment*) is a summation of true sentiment prediction for the classes of Positive *T*(*Pos*), Negative *T*(*Neg*), and Neutral *T*(*Neu*), i.e. *True*(*sentiment*) = *T*(*Pos*)+*T*(*Neg*)+*T*(*Neu*). Meanwhile, *False*(*sentiment*) is a summation of false sentiment prediction for the classes of Positive *F*(*Pos*) , Negative *F*(*Neg*) , and Neutral *F*(*Neu*), i.e. *False*(*sentiment*) = *F*(*Pos*)+*F*(*Neg*)+*F*(*Neu*). The numerical values for the measurement of sentiment classification accuracy can be obtained from the confusion matrix as described in Table 2.

TABLE II. CONFUSION MATRIX FOR THE PREDICTION ACCURACY.

Predictions	Sentiment Classes		
	True Neutral	True Positive	True Negative
Neu.	*T*(*Neu*)	*F*(*Neu*)	*F*(*Neu*)
Pos.	*F*(*Pos*)	*T*(*Pos*)	*F*(Pos)
Neg.	*F*(*Neg*)	*F*(*Neg*)	*T*(*Neg*)

Figure 6. Proportion of visiting purposes to Bangkok.

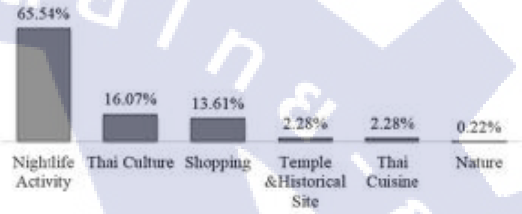


Figure 7. Proportion in percentage of purposes for traveling in Bangkok.

IV. EXPERMENTAL RESULTS

The experimental results will be described in four aspects, involving (i) visiting purposes, (ii) traveling types, (ii) numerical values of sentiment classes, and (iv) comparison of accuracy of the four analysis techniques.

A. Results of Visiting Purposes

Fig. 6 reveals the proportion of the visiting purposes of foreign tourists to Bangkok. It is seen in Fig.6 that the largest portion is for traveling with a percentage of 71.93% followed by business, visiting family, education, and health and treatments with percentages of 17.93%, 5.30%, 3.99% and 0.85%, respectively. Fig. 7 shows the proportion in percentage of purposes for traveling in Bangkok. The highest percentage is 65.54% of nightlife activity followed by Thai culture, shopping, temple and historical site, Thai cuisine and nature with percentage is 16.07%, 13.61%, 2.28%, 2.28% and 0.22% respectively. As the objective of this paper is to

TABLE III. THE CONFUSION MATRIX OF TECHNIQUES AND PREDICTIONS ON SENTIMENT CLASSES.

Techniques and Predictions		Sentiment Classes		
		True Neu.	True Pos.	True Neg.
(i) Decision Tree	Neu.	**4122**	1216	194
	Pos.	26	**562**	6
	Neg.	2	4	**1046**
(ii) Naïve Bayes	Neu.	**1800**	76	10
	Pos.	838	**990**	31
	Neg.	1512	716	**1205**
(iii) Support Vector Machine	Neu.	**3948**	1020	174
	Pos.	197	**756**	26
	Neg.	5	6	**1046**
(iv) Artificial Neural Network	Neu.	**4025**	1084	184
	Pos.	124	**695**	16
	Neg.	1	3	**1046**

TABLE IV. COMPARISONS OF SENTIMENT CLASSIFICATION ACCURACY.

Sentiment Classification Techniques	Decision Tree	Naïve Bayes	Artificial Neural Network	Support Vector Machine
Accuracy (%)	79.83	55.66	80.33	80.11

acquire information on sentiment of foreign tourists in order to improve tourism in Bangkok, Fig. 7 finally suggests that traveling should be fostered in order to promote the tourism in Bangkok.

B. Accuracy Analysis

The accuracy was analyzed using RapidMiner Studio Version 7.4. Table 3 summarizes predictions of each technique on sentiment classes in terms of true neutral, true negative, and true positive. It can be considered based on 7,193 tweets of the traveling type, ANN provides the highest number of 4,025 for true neutral prediction whilst Naïve Bayes provides the highest number of 990 for true positive prediction. For true positive, the number of 1,046 was equally predicted by decision Tree, SVM, and ANN.

Table 4 shows the comparison of sentiment classification accuracy for each technique, the best performance is given by ANN with an accuracy of 80.33% followed by SVM, decision tree, and Naïve Bayes with an accuracy of 80.11%, 79.83% and 55.66%, respectively. Based upon the prediction results in Table 4, ANN technique was selected for analyzing the sentiment of foreign tourists in traveling category. Fig.8 summarizes the sentiment results of foreign tourists to Bangkok. It is shown in Fig.8 that the positive sentiment of nightlife activity, temple and historical sites, Thai cuisine, and nature have scores of 0.29, 0.26, 0.39, and 0.37 respectively. On the other hand, the negative sentiment of Thai culture with a score of -0.36, while shopping is relatively neutral with a score of 0.03.

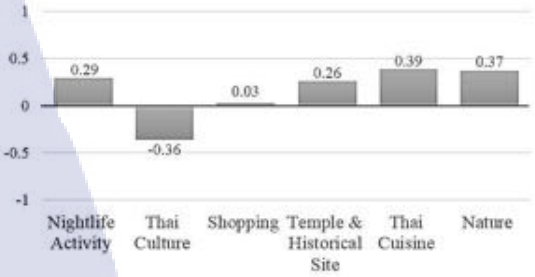


Figure 8. Sentiment of foreign tourists.

V. CONCLUSION

This paper has presented an analysis of sentiment of foreign tourists to Bangkok, Thailand, based on 10,000 datasets from Twitter in 2017. The purposes of visits were initially classified into five categories of visiting Bangkok, involving (i) Traveling, (ii) Business, (iii) Visiting Family, (iv) Education, and (v) Health and Treatments. The results have revealed that the traveling purpose has the highest percentage of 71.93% followed by business and visiting family. Therefore, the sentiment of foreign tourists to traveling in Bangkok was analyzed through four intelligent approaches, i.e. Decision Tree, Naïve Bayes, SVM, and ANN, using RapidMiner Studio7.4. The results have shown that the foreign tourists visit in Bangkok mostly for nightlife activity, Thai culture, and shopping with percentages of 65.54%, 16.07%, and 13.61%, respectively, meanwhile temple and historical sites, Thai cuisine, and nature are not significant. The accuracy of sentiment analysis approaches of decision tree, Naïve Bayes, SVM, and ANN are 79.83%, 55.66%, 80.11%, and 80.33%, respectively. ANN technique was selected for analyzing the sentiment of foreign tourists in traveling category. Classifications of sentiment of those tourist activities can be divided into positive, negative and neutral, which are represented by a score of +1, -1, and 0, respectively. The positive sentiments of nightlife activity, temple and historical sites, Thai cuisine, and nature have scores of 0.29, 0.26, 0.39, and 0.37 respectively. On the other hand, the negative sentiment of Thai culture with a score of -0.36, while shopping is relatively neutral with a score of 0.03. This paper therefore suggests an acceleration of nightlife activity of Bangkok in order to foster tourism industry of Bangkok.

ACKNOWLEDGEMENTS

The authors are grateful to Thai-Nichi Institute of Technology, Thailand, for financial supports. The authors are also would like to thank members of Intelligent Electronics Research Laboratory for kind cooperation and efforts on results discussions.

REFERENCES

[1] Online: http://wearesocial.com/blog/2017/, accessed on 1 April 2017.

[2] Shweta Rana and Archana Singh, "Comparative Analysis of Sentiment Orientation Using SVM and Naïve Bayes Techniques", In the Proceeding of the 2nd International Conference on Next

Generation Computing Technologies (NGCT-2016), pp. 106-111, 2016.

[3] Akshi Kumar and Ritu Rani, "Sentiment Analysis Using Neural Network", In the Proceeding of the 2nd International Conference on Next Generation Computing Technologies (NGCT-2016), pp. 262-267, 2016.

[4] Ankur Goel, Jyoti Gautam and Sitesh Kumar, "Real Time Sentiment Analysis of Tweets Using Naive Bayes", In the Proceeding of the 2nd International Conference on Next Generation Computing Technologies (NGCT-2016), pp. 257-261, 2016.

[5] Pierre Ficamo, Yan Liu, and Weiyi Chen, "A Naive Bayes and Maximum Entropy approach to Sentiment Analysis: Capturing Domain-Specific Data in Weibo", In the Proceeding of the 4th International Conference on Big Data and Smart Computing (BigComp), pp. 336-339, 2017.

[6] Tanu Verma, Renu and Deepti Gaur, "Tokenization and Filtering Process in RapidMiner", International Journal of Applied Information Systems (IJAIS), Foundation of Computer Science FCS, Vol 7, No. 2, pp.16-18, 2014

[7] Kavita Choudhary Prateek Maheshwari and Sonia Wadhwa, "Glaucoma Detection using Cross Validation Algorithm: A comparitive evaluation on Rapidminer", In the Proceeding of IEEE Conference on Norbert Wiener in the 21st Century (21CW), 2014.

[8] Nopparoot Kitcharoen et al, "RapidMiner Framework for Manufacturing Data Analysis on the Cloud", In the Proceeding of 10th International Joint Conference on Computer Science and Software Engineering (JCSSE), pp.149-154, 2013.

[9] Mohit Mertiya and Ashima Singh, "Combining Naive Bayes and Adjective Analysis for Sentiment Detection on Twitter", In the Proceeding of the International Conference on Inventive Computation Technologies (ICICT), 2016.

[10] Jongchan LEE and Taeseon YOON, "Analysis of Relation between Aging and Telomere using Datamining – Apriori, Decision Tree, and Support Vector Machine(SVM)", In the Proceeding of 19th International Conference on Advanced Communication Technology (ICACT), pp. 685-689, 2017.

[11] Yurong Zhong, "The analysis of cases based on decision tree", In the Proceeding of 7th IEEE International Conference on Software Engineering and Service Science (ICSESS), pp. 142-147, 2016.

[12] Eujin Yang, Bokyung Gu and Taeseon Yoon, "Intensified Analysis and Comparison of 5 Flacicirus with the use of Decision Tree and Support Vector Machine (SVM)", In the Proceeding of 19th International Conference on Advanced Communication Technology (ICACT), pp. 526-529, 2017.

[13] Thanaruk Theeramunkong, "Introduction to Concepts and Techniques in Data Mining and Application to Text Mining", Thammasat University Press and Tana Press Co.,Ltd, Nov 2012.

[14] G.Vinodhini and RM.Chandrasekaran "Sentiment Classification Using Principal Component Analysis Based Neural Network Model", In the Proceeding of International Conference on Information Communication and Embedded Systems (ICICES2014), 2014.

[15] P. Kalarani and S. Selva Brunda, "An efficient approach for ensemble of SVM and ANN for sentiment classification", In the Proceeding of International Conference on Advances in Computer Applications (ICACA), pp. 99-103, 2016

IEEE 15th International Conference
on Industrial Informatics INDIN´2017
– The Undergoing Industrial Informatics R-Evolution –
24th to 26th July, 2017
Emden, Germany

Certificate of Attendance

This is to certify that

Nattaphon Talmongkol

has attended the 15th International Conference

on Industrial Informatics 24-26 July 2017, Emden, Germany.

Luis Gomes
General Chair

Kamal Al-Haddad
President
Industrial Electronics Society

Armando Walter Colombo
General Chair

IEEE
Advancing Technology for Humanity

ICE

I²AR

FCT
FACULDADE DE CIÊNCIAS E TECNOLOGIA

University of Applied Sciences
HOCHSCHULE EMDEN·LEER

C. Fuzzy Logic Inference System for Appraisement of Human Happiness Level Based on Temperature, Relative Humidity, and Particulate Matter: In the Proceeding of IEEE International Conference on Embedded Systems and Intelligent Technology, "Smart Embedded Systems for Industry 4.0" Thailand.

Fuzzy Logic Inference System for Appraisement of Human Happiness Level Based on Temperature, Relative Humidity, and Particulate Matter

Varuth Tanomvorsin, Nattaphon Talmongkol, Taweesak Kuhamanee, Noppadon Pongpisuttinun, Wimol San-Um
Center of Excellence in Intelligent System Integrations
Thai-Nichi Institute of Technology, Bangkok, Thailand, 10250, Tel: (+66)-2-763-2600, E-mail: ta.varuth_st@tni.ac.th

***Abstract*— The increasing utilization of Fuzzy Logic Inference System (FIS) has been recently realized as a potential alternative tool for complicated Air Quality (AQ) assessment associated with appraisement of Human Happiness Level (HHL). Since there is still an inadequacy of further experiment on accuracy comparison between predefined FIS and actual satisfaction for HHL appraisement, this paper proposes a particular FIS for an appraisement of HHL based on temperature, relative humidity, and particulate matter in air together with an accuracy evaluation through the use of actual survey from 22 diversified participants to compare the result. In this research study, temperature, humidity, and particulate matter are defined as main representatives for physical attributes or parameters affecting on human health and satisfaction in order to consider human actualization, especially happiness; the mentioned FIS is then created corresponding to those of specific parameters in form of linguistic variables based on *US EPA* 2016 and *NOAA* national weather service. Eventually, the concluded results are presented as a percentage accuracy of number of matches, on temperature, relative humidity, and HHL, occurred from both simulation using MATLAB Toolbox 2016a and actual survey per total number of matches and mismatches for each level of happiness including happy and comfort, unhappy, depress, sad, and feel sick, respectively.**

Keywords- Human Happiness Level; Fuzzy Logic Inference System; Healthy and Happy Livings.

I. INTRODUCTION

Typically, the Fuzzy Logic Inference System (FIS) has been extensively exploited to cope with conventional Air Quality (AQ) assessment based on standard scales, which encounter some practical difficulties on measure of complex levels of chemical compounds and various hazardous attributes in air, due to its advantageous concept of knowledge-based logic that cannot be evidently expressed as either "True" or "False" but as a linguistic form instead.

In contrast of conventional AQ assessment, the main physical attributes, indicating human happiness and being used in this research, include only temperature (Temp.), relative humidity (RH.), and particulate matter (PM2.5) in air, which are the major environmental factors conveniently perceived by common people in order to conduct a compared survey on satisfactory conditions emphasized on working environment. The happiness levels are subcategorized into 5 classes comprising of happy and comfort, unhappy, depress, sad, and feel sick, respectively, derived from NWS heat index, issued by US EPA 2016 and NOAA national weather service,

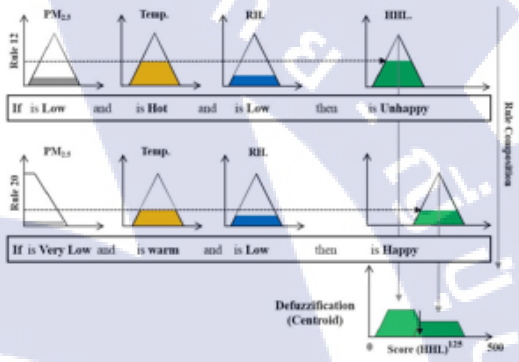


Figure 1. Fuzzy inference architecture for the HHL assessment with two parameters, two examples of rules 12 and 20, and all truncated functions combined for determining the HHL using Centroid method.

illustrating the significant relationship between air temperature and relative humidity causing a likelihood of heat disorders. Due to the additional aim to evaluate how accurate the particular FIS for HHL is compared to the actual outcome, the survey involving personal satisfaction on environmental factors along with level of happiness has been practically conducted as a comparator aside from only performing a conventional experimental result on an appraisement of AQ based HHL using FIS.

II. FUZZY LOGIC INFERENCE SYSTEM FOR HHL

To provide HHL assessment, FIS has been deployed as an expert system for modeling of nonlinear, uncertain, and complex systems generally based on Set theory, with an approximate reasoning process to allow transforming several input vectors to a single scalar output [2]. Table 1 summarizes the ranges of the three main parameters involving in this research study, and Table 2 defines linguistic variables for each parameter. Fig. 1 illustrates the fuzzy inference architecture for the HHL assessment with three parameters and two example of rules 12 and 20 through the use of Centroid Method combining all truncated functions in order to determine the HHL. For temperature and relative humidity, Trapezoidal membership function was realized, and Gaussian membership function for particulate matter. Two available types of FIS in MATLAB toolbox [11] are Mamdani and Takagi–Sugeno algorithms, though Mamdani algorithm has gained a greater widespread acceptance and is well suited for

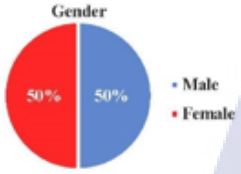


Figure 2. Proportion of participants' genders.

Figure 3. Proportion of participants' age range.

Figure 4. Proportion of participants' clothing type.

complex system and decision processes with human inputs. By using MATLAB Toolbox 2016a with two available types of FIS, the simulations have been performed according to formulated linguistic rules versus linguistic variables demonstrated on Table 3, which also indicates all five levels of human happiness.

III. Actual Survey - Comparator

The actual survey was conducted as a comparator through the gathering opinions from 22 participants pairing each satisfaction or happiness level corresponding to each interval of both temperature and relative humidity level, and also providing their personal level of awareness on particulate matter affecting their working environments. The participants are 22 sampled individuals with common experiences on physical environment and diversity of attributes including gender, age range, and clothing type, which are depicted in details on Fig. 2, Fig. 3, and Fig. 4, accordingly. The temperature intervals were consecutively categorized into level 1-5 as 0-22, 23-25, 26-31, 32-37, and above 37 in Celsius unit. Likewise, the relative humidity intervals were as well categorized into level 1-4 as 21-40%, 41-60%, 61-80%, and 81-100%. The intervals of both temperature and humidity were specified conform to the formerly mentioned FIS in order to make a proper verification on matching accuracy afterward. All participants were required to match each level

TABLE I. Ranges of Input Parameters

Parameters	Input Ranges	Units
Temp.	0-37	^{o}C
RH.	21-100	%
$PM_{2.5}$	0-500	$\mu g/m^3$

TABLE II. Linguistic Variables for Each Input Parameter.

Parameters	Linguistic Variables					
Temp. (^{o}C)	Cool	Warm	Hot	Very Hot	Super Hot	-
RH. (%)	Low	Medium	High	Very High	-	-
$PM_{2.5}$ ($\mu g/m^3$)	Very Low	Low	Medium	High	Very High	-

TABLE III. Proposed Descriptions and Emotional Interpretations of Human Happiness Level (*HHL*).

HHL	Descriptions and Emotional Interpretations [9-12]
Happy and Comfort	Self-Realization, Living a Life of Virtue, Pleasure Attainment, Pain Avoidance, Pleasant State or Relaxed Feeling of a Human Being in Reaction to Environments, Actualization of One's Inherent Potentials in the Pursuit of Complex and Meaningful Goals in both Individual and Society.
Unhappy	Not cheerful nor Glad, and also Not Appropriate nor Lucky.
Depress	Irritable Mood, Loss of Pleasure, Feeling Tired, Changes in Weight, Disturbance to Falling Asleep, Poor Quality of Sleep.
Sad	Emotional Pain, Feelings of Disadvantage, Loss, Despair, Grief, Helplessness, Disappointment and Sorrow, Suffering of Psychological or Non-Physical Origin.
Feel sick	Feel ill, Feel Very Upset, and Affected by Disease or Ill Health.

of both temperature and humidity with each one of 5 levels of happiness suited for working environment based on their own opinion. Then the number of matching levels from both simulation and survey compared to the total number of entire matches and mismatches shall be computed to obtain a particular percentage accuracy for each compared level of happiness. Additionally, the awareness level of particulate matter are rated as very high, high, medium, low, and very low due to their personal concern.

IV. Results of Comparison

According to the survey result demonstrated in Table 4, the accuracy around the middle range of HHL, i.e. unhappy, depress, and sad, tend to have considerably lower accuracy when being compare with the other two distinct levels of happiness, i.e. happy/comfort and feel sick. In this case, the output has obviously shown that it is more sensible for human, participants, to possibly classify each levels of happiness if they are not too similar or ambiguous to each other, e.g. comfort against feel sick. On the contrary, the simulation from FIS has more exceeding capability appropriated for such delicate classification on HHL through the aid of predefined rules. Besides, the particulate matter awareness level demonstrated on Fig. 5 also has shown that most of the participants did have a very high concern on particulate matter

TABLE IV. PROPOSED THE RESULT OF ACCURACY PERCENTAGE OF MATCHING BETWEEN NUMBER OF TRUE AND TOTAL NUMBER OF BOTH TRUE AND FALSE FROM SURVEY RESULT COMPARED TO THE SIMULATION.

Human Happiness Level (HHL)	Survey Result		Accuracy (%)
	Number of True	Number of True + False	
Happy/Comfort	28	35	80%
Unhappy	11	16	69%
Depress	3	11	27%
Sad	1	9	11%
Feel Sick	13	14	93%
Total	56	85	66%

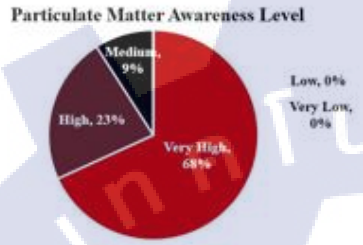


Figure 5. Result for each particulate matter awareness level.

affecting their working environment, and not even a single one has low or very low concern on particulate matter.

V. CONCLUSIONS

This paper has presented the use of Fuzzy Logic Inference System (FIS) for an appraisement of Human Happiness Level (HHL) based on temperature, relative humidity, and particulate matter and also has provided a brief evaluation on how accurate the FIS is when being compared with both the simulation and the actual satisfaction through related survey on the appraisement of HHL. The final result has shown that the FIS tend to have a high conformance to the actual happiness level based on participants' opinions when the HHL has more evident distinction. In addition, the result of actual HHL considered by participants may depends on variety of other factors such as ability of each individual's physical sensory, familiarity to local climate, and influences from cumulative experience. Moreover, the paper has also provide the definition on HHL consisting of five levels interpreted from human emotions.

The further research study should be extendedly focused on a significant effects on the number of people stayed within a fixed space and discover its hidden linkages to the HHL, enhancing from previous conventional fuzzy logic inference systems solely considering on only air quality assessment. The example of promising application could be introduced as an extraordinarily smart air conditioner control with genuine accordance to human actualization, especially happiness.

VI. ACKNOWLEDGEMENTS

The authors are grateful to Thai-Nichi Institute of Technology and Piwat Air Co., Ltd. for financial supports. The authors are also would like to thank members of Intelligent Electronics Research (IES) Laboratory for kind cooperation and facilitation.

VII. REFERENCES

[1] Allahbakhsh Javid, Amir Abbas Hamedian, Hamed Gharibi, Mohammad Hossein Sowlat, "Towards the Application of Fuzzy Logic for Developing a Novel Indoor Air Quality Index (FIAQI)", Iranian Journal of Public Health, Vol. 45, pp. 203-213, 2016.

[2] Miguel Ángel Olvera-García, José J. Carbajal-Hernández ,Luis P. Sánchez-Fernández, Ignacio Hernández-Bautista, "Air quality assessment using a weighted Fuzzy Inference System", J. Ecological Informatics, Vol. 33, pp. 57–74, 2016.

[3] Mohammad Hossein Sowlat, Hamed Gharibi, Masud Yunesian, Maryam Tayefeh Mahmoudi, Saeedeh Lotfi, "A novel, fuzzy-based air quality index (FAQI) for air quality assessment", J. Atmospheric Environment, Vol. 45, pp. 2050-2059, 2011.

[4] M.N. Assimakopoulos, A. Dounis, A. Spanou, M. Santamouris, "Indoor air quality in a metropolitan area metro using fuzzy logic assessment system", J. Science of the Total Environment, Vol. 449, pp. 461–469, 2013.

[5] Xin Zhang, Xiaobo Zhang, Xi Chen,"Valuing Air Quality Using Happiness Data: The Case of China" J. Ecological Economics, Vol. 137, pp. 29-36, 2017.

[6] Christopher L. Ambrey, Christopher M. Fleming, Andrew Yiu-Chung Chan, "Estimating the cost of air pollution in South East Queensland: An application of the life satisfaction non-market valuation approach" J. Ecological Economics, Vol. 97, pp. 172-181, 2014.

[7] Fuzzy Inference System Modeling, MATLAB and Simulink, MathWorks manual.

[8] José Juan Carbajal-Hernández, Luis P. Sánchez-Fernández, Jesús A. Carrasco-Ochoa, José Fco. Martínez-Trinidad, "Assessment and prediction of air quality using fuzzy logic and autoregressive models", Atmospheric Environment 60, pp. 37-50, 2012.

[9] Antonella Delle Fave, Ingrid Brdar, Teresa Freire, Dianne Vella-Brodrick, Marié P. Wissing, "The eudaimonic and hedonic components of happiness: qualitative and quantitative findings", Social Indicator Research 100, pp. 185–207, 2011.

[10] R.M. Ryan and E.L. Deci "On happiness and human potentials: a review of research on hedonic and eudaimonic well-being", Annual Review on Psychology, Vol. 52, pp. 141-66, 2001.

[11] P. Vink, M.P. Looze, L.F.M. Kuijt-Evers "Comfort and Design: Principles and Good Practice", J. Applied Ergonomics, Vol. 43, pp. 271–276, 2012.

[12] Depression, UK National Institute for Health and Clinical Excellence (NICE), October, 2009.

D. A Case Study using Recurrence Plot Analysis to Analyze the Characteristic Features of a Heart Disease: In the proceeding of IEEE Management and Innovation Technology International Conference (MITiCon), Thailand.

A Case Study using Recurrence Plot Analysis to Analyze the Characteristic Features of a Heart Disease

Wimol San-Um, Nattaphon Talmongkol, Varuth Tanomvorsin
Center of Excellence in Intelligent System Integrations
Thai-Nichi Institute of Technology, Bangkok, Thailand, 10250, Tel: (+66)-2-763-2600, E-mail: wimol@tni.ac.th

Abstract— **The recurrence plot analysis (RPA) has been widely utilized to extract characteristic features from conventional ECG signals with the transformed recurrence plot (RP) expressed in concurrent time domain instead, which is normally focused only on the interpretation of the plot itself rather than the existing characteristics caused by the original signal. Therefore, this research paper proposes a case study of using RPA to analyze the extracted features of a 15-lead ECG signal with particular anomaly, a bundle branch block (BBB), which is initially obtained from the MIT-BIH database. The dataset of ECG signals representing BBB have been carefully selected and properly performed in MATLAB software to illustrate all of the 15 unique patterns with the intention of verifying the shared features among those patterns representing BBB. Besides, the purpose is to emphasis on the example use of extracted characteristic patterns as to qualitatively identify the type of a heart disease compared visually with the normal RP generating from a normal ECG signal as a reference for even further research on analysis of heart diseases classification.**

Keywords- ECG; Heart Disease; Bundle Branch Block; Recurrence Plot; RPA; MIT-BIH.

I. INTRODUCTION

The recurrence plot analysis (RPA) has been widely used as an advanced visualization technique to visually analyze the nonlinear time-series dynamics in time domain. Normally, this approach is expected to be easily identified for the analysis of nonlinear signals, e.g. ECG signals, even in the situation where several signals possessing the close similarity that are inconvenient to identify immediately. Although the recurrence plot has been widely used especially with the ECG signal analysis, only few researches are essentially focused on the occurred features from recurrence plot (RP) pattern indicating the original type of major heart diseases. In this case, a dataset of bundle branch block (BBB) from the MIT-BIH database has been selected as a representative of a heart disease transformed into the form of RP for further analysis. The dataset of ECG signals comprise of 15 signals representing each lead that measures the patient. Even though the position of all the leads are one of the major indicators to identify the type of heart disease, this research paper emphasizes more on the appearance of fifteen characteristic patterns rather than considering intensively on the significant association between each pattern and each lead with different angles around the chest and limb.

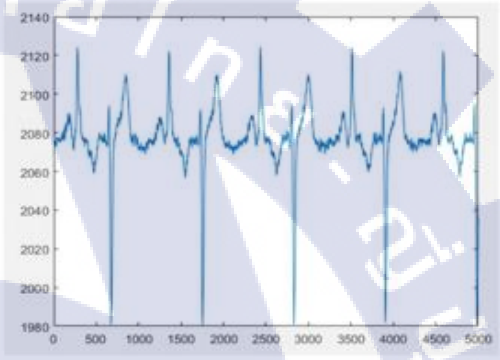


Figure 1. An ECG signal of a normal person

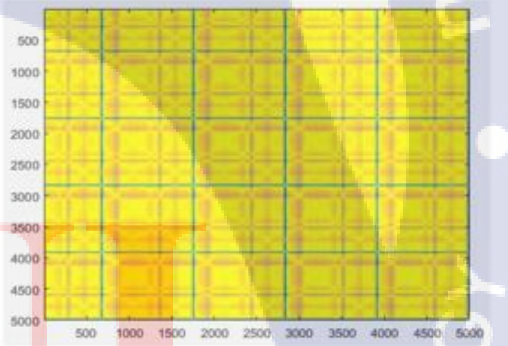


Figure 2. The related recurrence plot of normal ECG

II. EXPERIMENTS AND RESULTS

From the MIT-BIH database, both sets of normal ECG and ECG of bundle branch block were selected and processed to recurrence plot with MATLAB. Fig. 1 depicts the ECG signal of a normal person, and Fig. 2 also depicts its related recurrence plot. It has shown that, in case of normal ECG, the pattern of RP has been uniformly distributed among the entire plot with approximate symmetry. The normal ECG may varies for each individual person, but the patterns should possess the same features such as regularly distributed of lines and stripe, with uniform distances over the area. The blue colors indicates that the original signal value is very far

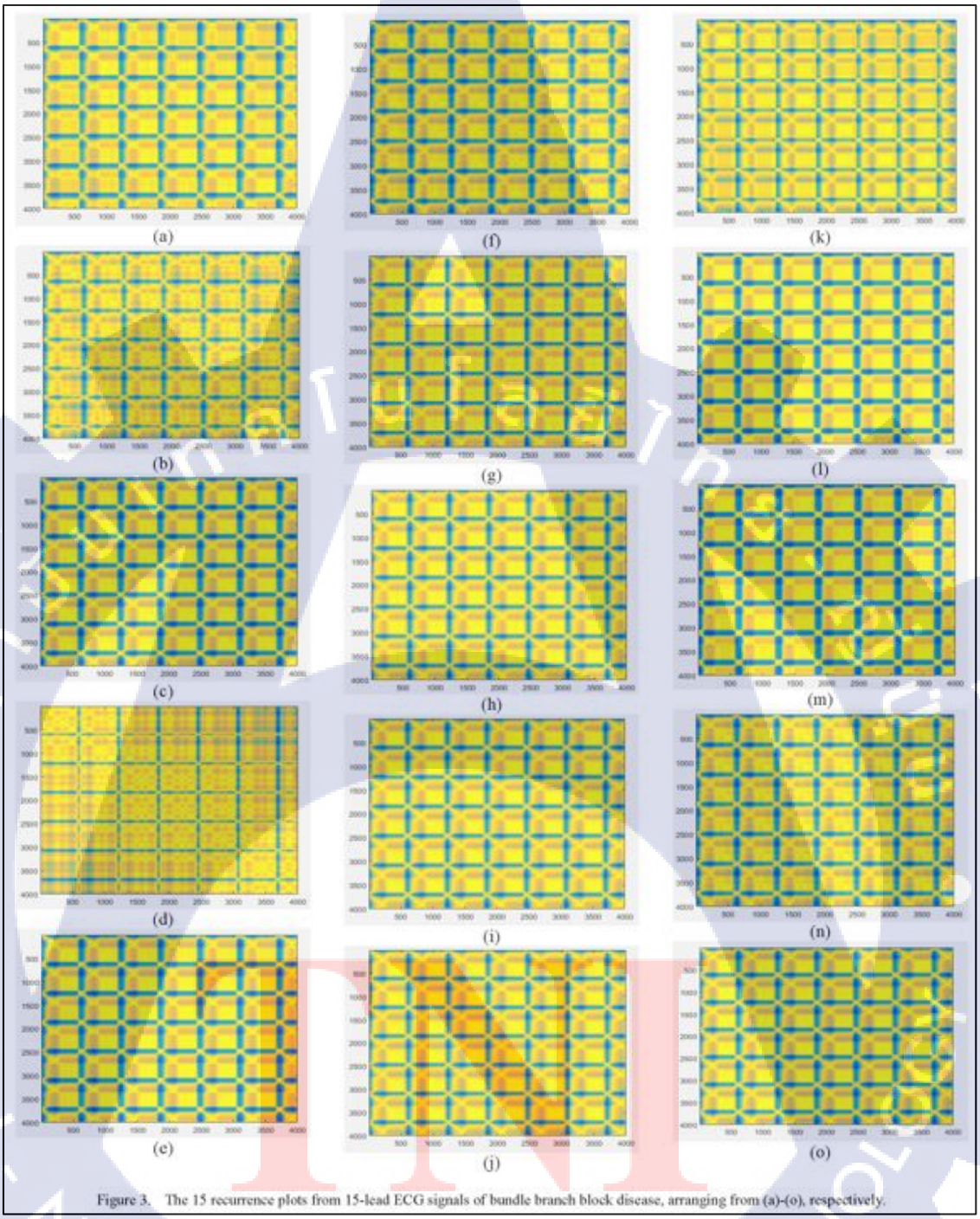


Figure 3. The 15 recurrence plots from 15-lead ECG signals of bundle branch block disease, arranging from (a)-(o), respectively.

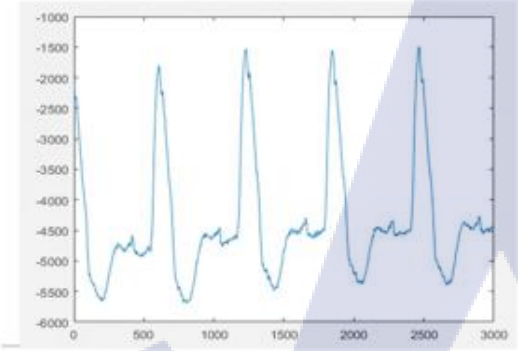


Figure 4. An ECG signal of bundle branch block disease.

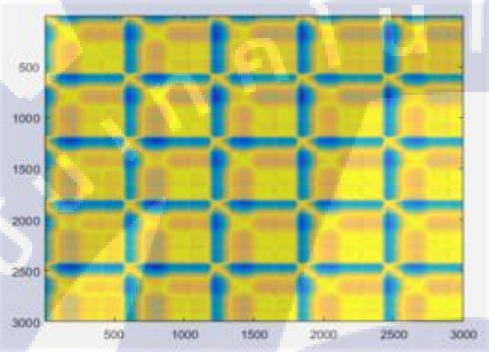


Figure 5. The related recurrence plot of bundle branch block disease.

different compared to others, the yellow colors indicates that the original signal approximately has the same value as others and the orange ones show that the signal values are moderately different compared to the others in respective time.

After a brief overview of both ECG and RP of normal case, an ECG representing anomaly of heart disease will then be performed as a countermeasure to the previous normal ECG. The ECG dataset of 15 leads of BBB for a patient has been obtained from MIT-BIH database as mentioned earlier. To display all possible RPs from every leads, those 15 sets of signal data were well transformed into all 15 respective recurrence plots as shown in Fig. 3, depicting all of the 15 related RPs ranging from (a) to (o), according to each lead. As illustrated, each of 15 RPs have their own unique patterns according to each lead as well, but a thing they have in common is . Additionally, the value scale of both each ECG and PR maybe varied due to the position of lead placing to measure the electrical signals.

In Fig. 4, a sample of ECG in case of BBB has well plotted and presented despite of the different time scale with the normal one. The aim is just to provide a clear figure of ECG with specific sinus rhythm. In this case, the distinction between the normal and the anomaly still could be easily distinguish due to their apparent appearance, but there are also many cases where, in ECG scenario, the signals might be quite difficult to identify and classify whether there is any issue of heart disease occurring. In contrary, with the aid of recurrence plot analysis, the result of matching characteristic pattern as depicted in Fig. 5 has apparently shown that the plot has some unique patterns appearing, which could be easily and immediately identified for any occurring anomalies.

III. Conclusion

Comparing to the recurrence plot of normal ECG, the recurrence plot on the case of bundle branch block disease tends to have thicker stripes in both blue and yellow color since the parameter Q, R, and S for ECG of BBB would be wider than usual. The particular cause is that BBB belongs to a condition where there is an obstruction resulting in the delay of impulses traveling along the pathway to stimulate the heartbeat. In addition, the entire fifteen leads might not be necessary to reveal and expose the same features, depending on the positions of the leads.

As aforementioned, this conceptual idea of case study is still only focused on the extracting of distinctive features of a heart disease in time-domain patterns, comparing visually to the RP of normal case in term of revealed pattern. The further research could be conducted by considering characteristic features from other major heart diseases in RP form as well, and the concept could even be dramatically leverage through the use of image processing techniques or deep learning techniques in order to properly perform precise classification of each heart diseases through RPA.

IV. Acknowledgments

The authors are grateful to Thai-Nichi Institute of Technology for facility supports and also would like to thank members of Intelligent Electronics Research (IES) Laboratory for kind cooperation and facilitation.

V. References

[1] Alexandru Serbanescu, Florin-Marian Birleanu, Angela Digulescu, "Overview of our recent results in signal analysis using recurrence plots", COMPUTERS and ARTIFICIAL INTELLIGENCE - ECAI-2013, pp. 1-6, 2013.

[2] Elif Tuba Celik, Bogdan Hurezeanu, Angela Digulescu, Madalina Mazilu, "FETAL ECG REPRESENTATION USING RECURRENCE PLOT ANALYSIS",2012 6th International Conference on Application of Information and Communication Technologies (AICT), pp. 1-4, 2012.

[3] Philip I. Terrill, Stephen J. Wilson, Sadasivam Suresh, David M. Cooper, "Characterising infant inter-breath interval patterns during active and quiet sleep using recurrence plot analysis", 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 6284-6287, 2009.

[4] David T. Mewett, Karen J. Reynolds, Homer Nazeran, "RECURRENCE PLOT FEATURES: AN EXAMPLE USING ECG", Fifth International Symposium on Signal Processing and its Applications, vol. 1, pp. 175-178, 1999.

[5] F Censi, G Calcagnini, S Cerutti, "Recurirence Plot Analysis of the Coupling between Respiration and Cardiovascular Variability Signals", Computers in Cardiology, pp. 211-214, 1997.