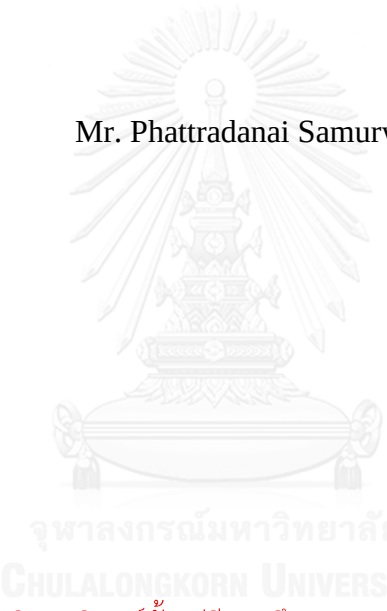
TWO-STAGE PREDICTIVE MODEL FOR THAI STOCK RETURN PREDICTION

Mr. Phattradanai Samurwong



A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Financial Engineering
Department of Banking and Finance
Faculty of Commerce and Accountancy
Chulalongkorn University
Academic Year 2015

แบบจำลองทำนายสองขั้นสำหรับการทำนายผลตอบแทนหุ้นไทย

นายภัทรดนัย เสมอวงษ์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมการเงิน ภาควิชาการธนาคารและการเงิน
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2558

Thesis Title	TWO-STAGE PREDICTIVE MODEL FOR THAI STOCK RETURN PREDICTION
By	Mr. Phattradanai Samurwong
Field of Study	Financial Engineering
Thesis Advisor	Assistant Professor Krung Sinapiromsaran, Ph.D.
Thesis Co-Advisor	Assistant Professor Thaisiri Watewai, Ph.D.

Accepted by the Faculty of Commerce and Accountancy, Chulalongkorn University in Partial Fulfillment of the Requirements for the Master's Degree

.....Dean of the Faculty of Commerce and Accountancy
(Associate Professor Pasu Decharin, Ph.D.)

THESIS COMMITTEE

.....Chairman
(Anant Chiarawongse, Ph.D.)

.....Thesis Advisor
(Assistant Professor Krung Sinapiromsaran, Ph.D.)

.....Thesis Co-Advisor
(Assistant Professor Thaisiri Watewai, Ph.D.)

.....Examiner
(Assistant Professor Sira Suchintabandid, Ph.D.)

.....External Examiner
(Kridsda Nimmanunta, Ph.D.)

ภัทรดนัย เสมอวงษ์ : แบบจำลองทำนายสองขั้นสำหรับการทำนายผลตอบแทนหุ้นไทย (TWO-STAGE PREDICTIVE MODEL FOR THAI STOCK RETURN PREDICTION) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: ผศ. ดร.กรุง สินอภิรมย์สราญ, อ.ที่ปรึกษาวิทยานิพนธ์ร่วม: ผศ. ดร.ไทยศิริ เวทไว, 50 หน้า.

วิทยานิพนธ์นี้ประกอบจากสองส่วนที่มุ่งเน้นไปที่การทำนายผลตอบแทนของดัชนี SET50 ส่วนแรกของวิทยานิพนธ์ใช้ตัวแบบเครื่องสนับสนุนเวกเตอร์ในการทำนายทิศทางของผลตอบแทน SET50 รายวัน ในขณะที่ส่วนที่สองของวิทยานิพนธ์ใช้ผลตอบแทนของสัญญาซื้อขายดัชนี Hangseng ล่วงหน้า สัญญาซื้อขายดัชนี Dow Jones ล่วงหน้าและการผสมผสานของดัชนีทั้งสอง เป็นตัวแปรเข้าในตัวแบบสองขั้นที่ประกอบไปด้วยตัวแบบเครื่องสนับสนุนเวกเตอร์สำหรับทำนายทิศทางในขั้นแรกและตัวแบบโครงข่ายประสาทเทียมสำหรับการทำนายผลตอบแทนราย 5 นาทีของดัชนี SET50 ในขั้นที่สอง

ส่วนแรกของวิทยานิพนธ์นำเสนอกระบวนการเตรียมข้อมูลและเทคนิคการทำนายทิศทาง ตัวแบบเครื่องสนับสนุนถูกเลือกใช้เพราะเป็นที่ยอมรับสำหรับการทำนายทิศทางของผลตอบแทน ผลตอบแทนของดัชนีในวันก่อนหน้าและผลต่างของดัชนีก่อนหน้าถูกนำมาใช้เป็นตัวแปรนำเข้าเพื่อเพิ่มสารสนเทศให้กับเครื่องสนับสนุนเวกเตอร์ ผลการทดลองแสดงให้เห็นว่าจำนวนผลต่างของลำดับของผลต่างของผลตอบแทนและจำนวนผลตอบแทนถดถอยไม่มีนัยสำคัญทางสถิติต่อการลดหรือเพิ่มของความแม่นยำ อย่างไรก็ตามการตั้งค่าที่ดีที่สุดจากตัวแบบเครื่องสนับสนุนเวกเตอร์แสดงถึงความแม่นยำที่ดีกว่าตัวแบบอื่น ๆ เช่น ตัวแบบโครงข่ายประสาทของ Rodriguez et al. (2000)

ส่วนที่สองของวิทยานิพนธ์นำเสนอตัวแบบสองขั้นที่ใช้ทำนายผลตอบแทนดัชนี SET50 รายห้านาที ตัวแปรที่ใช้ประกอบไปด้วย ผลตอบแทนสัญญาซื้อขายล่วงหน้าของดัชนี Hangseng และดัชนี Dow Jones และการผสมผสานของดัชนีทั้งสอง ผลตอบแทนที่มีขนาดเล็กจะถูกคัดออกตามค่าเปอร์เซ็นไทล์ต่างๆ ตั้งแต่ระดับ 0 ถึง 95 เปอร์เซ็นไทล์ ตัวแบบสองขั้นใช้ตัวแบบเครื่องสนับสนุนเวกเตอร์สร้างทิศทางของผลตอบแทนของดัชนีรายห้านาทีเพื่อเพิ่มตัวแปรนำเข้าในขั้นที่สองซึ่งใช้ตัวแบบโครงข่ายประสาทในการทำนายผลตอบแทนของดัชนีรายห้านาที จากผลการทดสอบพบว่าตัวแบบสองขั้นมีความแม่นยำที่ดีกว่าตัวแบบโครงข่ายประสาทเพียงอย่างเดียว นอกจากนั้นการศึกษาพบว่าการนำผลตอบแทนที่มีขนาดเล็กออกช่วยให้การทำนายทิศทางของผลตอบแทนในอนาคตมีแนวโน้มแม่นยำมากขึ้น

ภาควิชา การธนาคารและการเงิน	ลายมือชื่อนิสิต .....
สาขาวิชา วิศวกรรมการเงิน	ลายมือชื่อ อ.ที่ปรึกษาหลัก .....
ปีการศึกษา 2558	ลายมือชื่อ อ.ที่ปรึกษาร่วม .....

5582081126 : MAJOR FINANCIAL ENGINEERING
KEYWORDS: STOCK PREDICTION / ABNORMAL RETURN, STOCKS, PROFITS, BEHAVIORAL FINANCE / TWO-STAGE MODEL / SUPPORT VECTOR MACHINE / NEURAL NETWORK / SET50 INDEX / TWO-STAGE MODEL

PHATTRADANAI SAMURWONG: TWO-STAGE PREDICTIVE MODEL FOR THAI STOCK RETURN PREDICTION. ADVISOR: ASST. PROF. KRUNG SINAPIROMSARAN, Ph.D., CO-ADVISOR: ASST. PROF. THAISIRI WATEWAI, Ph.D., 50 pp.

This thesis consists of two parts focusing on predicting the SET50 index return. The first part uses a support vector machine model to predict the daily directions of SET50 index returns while the second part uses the index future returns from the Hangseng, the Dow Jones and their combinations as inputs of a two-stage model which is composed of a support vector machine model for the directional prediction in the first stage and a neural network model for the value prediction in the second stage to predict five-minute SET50 index returns.

Firstly, in this thesis, the combined data preprocessing and classification technique for the stock return direction is proposed. The support vector machine is selected due to its popularity for the stock return direction prediction. To give additional information for the support vector machine, the higher order differences and the higher order lags are fed as the additional inputs. Our experiments show that the predictive accuracy with respect to the number of difference orders and the number of lags are statistically insignificant. Nonetheless, the best setting of the support vector machine model shows the accuracy improvement over the other models such as the neural network model by Fernandez-Rodriguez et al. (2000).

Secondly, the two-stage model for the five-minute stock index returns is proposed. The inputs are the Hangseng and the Dow Jones index future returns and their combinations. Returns with small magnitude are filtered out at different thresholds ranging from 0 percentile to 95 percentile. Our two-stage model uses a support vector machine model to generate the direction of the five-minute stock index return and attaches it as the input of the second stage which is a neural network to predict the value of the five-minute stock index. Our two-stage model outperforms a single neural network model in terms of accuracy. The magnitude of the stock returns used in the model affects the predictive performance. Dropping low percentile ranks of the stock index returns improves the predictive accuracy.

Department:	Banking and Finance	Student's Signature	.....
Field of Study:	Financial Engineering	Advisor's Signature	.....
Academic Year:	2015	Co-Advisor's Signature	.....

ACKNOWLEDGEMENTS

Firstly, I would like to thank Assistant Professor Krung Sinapironsaran who is the adviser of this thesis. Also, I would like to thank Assistant Professor Thaisiri Watewai for giving advice during this study and also being the co-adviser of this thesis. I sincerely thank Assistant Professor Sira Suchintabandid, Anant Chiarawongse and Kridsa Nimmanunta for their invaluable comments.

Moreover, I would like to thank the Financial Engineering program, the Department of Banking and Finance, and Chulalongkorn University for the technical and financial support. Also, I thank Chayanapat Chiangraksa for her administrative support and lastly, Yotsanan Simakorn for revising the early version of this thesis.

CONTENTS

Page

LIST OF TABLE

LIST OF FIGURE

CHAPTER I
INTRODUCTION

1.1 Background and Problem Review

Nowadays, stock prediction is one of the most interesting topics in financial engineering. There are many methodologies to forecast the time series data using the time series models, machine learning and mixed approaches. Although there are rigorous assumptions and economical intuition behind the stock prediction using the time series models, the predictability of this approach underperforms in terms of predictive accuracy compared with the machine learning approaches (Doeksen, et al. (2005)). Therefore, many of quantitative trading firms are relying on the use of machine learning models with the lack of rationally explanation among inputs or intuitive understanding of a relationship among them.

One of the outstanding advantages of the data mining approach is that this approach works well on the non-linear pattern recognition. Thus, it can better capture a complexity of the data than the time series models. There are many learning models in literatures. For examples, Fernandez-Rodrıguez, et al. (2000) used a neural network model to predict the Madrid stock exchange general index; Bekaert and Harvey (2000) and Halliday (2004) used their neural network models to predict the New York stock exchange index; Lendasse, et al. (2000) used a radial basis function neural network to predict the Belgium stock exchange 20 index; Tay and Cao (2001) used a support vector machine to predict the S&P index; Kim (2003) used a support vector machine to predict the Korea stock exchange index; Doeksen, et al. (2005) used a Mamdani fuzzy system and a Takagi-Sugeno fuzzy system to predict the New York stock exchange index; Hsu, et al. (2009) used a support vector machine (SVM) + Self Organizing Map (SOM) model to predict the Hangseng index. One of the most popular algorithms in stock prediction is a neural network (NN) algorithm. The NN is widely used because of its accuracy in the prediction (Klimasauskas (1993)). There are a lot of studies that focus on the neural network prediction algorithms in stock markets. One of the most popularly used NN algorithms in stock market prediction is a NN back-propagation algorithm with the sigmoid learning function which can be a representative of the standard NN

because of its popularity. There is also another prediction algorithm that is widely used in the field which is a support vector machine.

A support vector machine (SVM) has recently become a popular algorithm due to its accuracy in the directional prediction from Tay and Cao (2001) and Wu (2011). It determines the best hyperplane that distinguishes data into groups which can be useful in the stock prediction. The stock index directional prediction has been one of the major tasks since the movement of the stock index is very noisy due to the trader's irrationality and many unpredictable factors. The multi-stage models were invented by many authors to capture the complexity of the stock prices and indices movement in order to predict the return of the stock prices. For instances, Hsu, et al. (2009) used a two-stage model from a self-organizing map and a support vector regression model to predict stock price. Their results showed that the prediction result was significantly better than an individual support vector regression model. Hsieh, et al. (2011) applied a two-stage model formed by the artificial bee colony algorithm and a recurrent neural network to predict the stock price that showed improvement. Using multiple models is now the most popular technique in predicting the stock price. Therefore, in this research, a two-stage model will be developed.

From the above observations, the SVM is one of the most used classification models. It is also used in the first part of this thesis and used as the first stage model in the proposed two-stage model to predict the direction of the stock index return. The SVM is designed to distinguish between the positive (the increase) and negative (decrease) of the stock index from the data provided.

In the second part of this thesis, the neural network (NN) is used to predict the stock index return. The directional prediction resulting from the first stage model will be used as additional input in the second stage. The NN functions as a predictor of the stock index return in both magnitude and directions. The two-stage model shows the performance improvement of the predictability compared to a single NN.

Moreover, the predictability of the model can be improved by generating additional variables or filtering the input data. In the first part of this thesis, the lags of returns and the higher difference orders of returns will be chosen as inputs. Using only a single time series as inputs for SVM to predict the stock return gives a low predictability for some financial indicators or some other signals and economic

indicators. Therefore, in this thesis, the appropriate inputs for the support vector machine model and the two-stage model is proposed to in-cooperate the appropriate number of inputs. In order to find the appropriate inputs for the prediction, all combinations of the lags and difference orders are considered. The setting with the best accuracy will be selected for the first and the second part of this thesis.

In this thesis, the SET50 index of Thai stock market is selected since it is a major index in Southeast Asia. Thai market is attractive to various investors since it is an emerging with a high potential growth. Thai stock index external effect mostly relies on major countries, so it is correlated to the major indices such as the Hangseng and the Dow Jones.

The SET50 index is a combination of weighted stock prices in Thai market that consists of the top 50 stocks in terms of market capitalization. The SET50 index return is used as the target in the two-stage model since the SET50 index represents the high liquidity stocks with the high market capitalization.

According to Chaigusin, et al. (2008), the intraday movement of the Thai stock index tends to correlate with the movement of the major foreign indices such as the Hangseng index, the Nikkei index etc. However the Hangseng index is not active in the afternoon session of Thai stock exchange, the inputs to predict intraday Thai stock index are the Hangseng and the Dow Jones index futures, which are 24-hour active, instead of the Hangseng index and the Dow Jones index.

This thesis consists of two parts. The first part is the study of the daily index return prediction using a SVM and the study of the effect of the number of lags and the number of difference orders of the inputs to the predictive accuracy. The second part is to build a predictive model for the five-minute SET50 index return using a two-stage model and the study of the effect of the low percentile rank data filtering to the predictive accuracy.

According to Schulmeister (2009), the daily trading is less profitable since the data and information change rapidly and trading methodologies rely on newer technologies, thus the trading profitability tends to work on higher frequencies such as 30-mintues-prices. In this thesis, the five-minute SET50 index returns are predicted. Moreover, Tolvi (2002) showed that the outlier detection was useful and informative

for stock prediction. The effects of prediction using large returns in this thesis are also shown by filtering low percentile rank data.

One of the most debated topics in the financial market is the efficiency of the markets (Malkiel and Fama (1970)). The market efficiency has long been argued (Malkiel (2003)) whether the market is absolutely efficient in the information that has reflected in the markets or the market is not efficient that investors have a room for making a profit. There are some evidences showing that the markets are extremely efficient. For example, Schwert (2003) reviewed some of the well-known studies about the effects of the size, value, weekend and momentum. The review showed that these effects disappeared after the publication of the related papers, thus it was implied the market efficiency. Also, some of the studies on individual investors (Barber and Odean (1999), Barber and Odean (2000), Benartzi and Thaler (2001)) showed that most individuals were irrational and behaved foolishly, since the market was efficient. However, some researchers criticized that the market was not efficient.

For example, Watts and Zimmerman (1978) concluded that there were significant abnormal returns from the effect of quarterly earnings announcement due to the fact that market was not efficient. Also, some studies showed that there was intraday inefficiency in the markets. Busse and Green (2002) showed that the price response to the morning call and midday call on CNBC news usually took time from 2 to 15 minutes. Epps (1979) concluded that price reflection to news was more than 10 minutes on average.

Our thesis which focuses on the intraday and daily market predictions will also be one of the evidence that shows the efficient of intraday and daily market.

1.2 Objective of the Study

This thesis has four major objectives. The first and second objectives are the results from the first part of the thesis, which focuses on using the multiple difference orders and the lags as inputs for SVM in the daily SET50 index directional prediction. The third and fourth objectives are the results from the second part of thesis, which focuses on using filtered percentile rank in the two-stage SVM-NN model for the five-minute SET50 index returns prediction.

The first objective is to focus on the improvement of the directional predictive accuracy by using the multiple lags and the difference orders in the data preprocessing step in the first stage model, SVM.

Either a neural network or a support vector machine has its own advantages. The neural network has its ability to capture the complexity of a non-linear relation. The support vector machine is good at classifying an imbalance data into two groups effectively. The second objective is to compare the accuracy of the support vector machine model with the neural network model. In the first part of this thesis, the neural network model is used as a benchmark against the support vector machine model.

In the second part of the thesis, we will try to extract the strength of each model in order to predict the stock more accurately by combining those models and creating a two-stage model in order to improve the predictability compared to an individual model. Thus, the third objective is to improve the predictability, compared with the individual neural network model, in terms of the accuracy and the mean absolute error by using the two-stage SVM-NN model. Since the prediction result of the two-stage model are the returns, which can be analyzed in two components. Those components are the direction of returns and the value of returns. Hence, the benchmarks of the two-stage model will also consist of two components which are the accuracy percentage and the mean absolute error (MAE). The accuracy percentage can be calculated by

$$Accuracy = \frac{Total\ \ number\ of\ \ correct\ \ direction\ \ prediction}{Total\ \ number\ of\ \ direction\ \ prediction} \times 100 \qquad (1.1)$$

And the mean absolute error can be calculated by

$$Mean\ \ Absolute\ \ Error = \frac{\sum \left| \Pr edicted\ \ \ Value - Actual\ \ value \right|}{Total\ \ Number\ of\ \ \Pr ediction} \times 100 \qquad (1.2)$$

Also, in the second part of the thesis, studying the effect of using a percentile rank filtering technique to improve the accuracy is the fourth objective.

The byproduct of this study will be the interpretation of the market efficiency in the multiple time frequencies. After applying the model, the result can be one of the evidence to support that the inefficiency of Thai market exists. The prediction accuracy of the thesis will be used as preliminary evidence showing whether there is a friction in Thai market by compared the accuracy of our model and of the random directions. If

the prediction accuracy of the model is far more than 50%, it may suggest that Thai market is somewhat predictable and the market is not perfectly efficient.

1.3 Contributions

Nowadays a machine learning model is popularly used and studied to predict stock returns. There are only a few works in Thai market that apply a neural network and a support vector machine (Inthachot, et al. (2015), Sirijunyapong, et al. (2014)). This thesis is an additional application of the SVM and the neural network model to predict the Thai stock index. The input variable in the previous study can also be applied with this study which focuses on the prediction of Thai stock index. By applying the inputs from the previous study, the inputs will be further modified by using the multiple difference orders, a technique that hasn't been applied elsewhere in this context, in the machine learning for the stock prediction.

Normally, the SVM and the neural network model are used independently to predict the stock prices (Trippi and Turban (1992), Kim (2003)). Recent studies, such as Hsu, et al. (2009), proposed using the combination of the models in order to improve the predictive accuracy. This study applies the combination of a SVM and a neural network model, which is called a two-stage model, for the Thai stock index. The combination of the SVM and the neural network has never been used before in predicting the stock index, thus the application and the predictability improvement of the multi-stage model from previous studies may provide an insight to the construction of a two-stage model.

As mentioned in the introduction section, the SVM has its strength in the predictive accuracy and the NN has its strength in capturing non-linear relationship. The combination of these models by using their strengths can be an insight to build the SVM-NN model and their strengths can also be expected to improve the accuracy of the prediction. As a result of combining the model, the accuracy improvement of the model will be shown by compared it to the individual neural network model.

Choosing the inputs for the prediction is one of the major concerns. Various inputs can increase the predictive power of a model. Many researchers used lags to create the momentum strategy (Lewellen (2002)). In addition to the lags, the first part

of this thesis will include difference orders for the alternative inputs. An additional technique proposed in the second part of this thesis is filtering noises according to their percentile ranks. Normally when no special event occurs, stocks are traded with low volatility. However, if a shocking news related to the stock market occurs, the price movement will fluctuate with higher volatility (Campbell and Hentschel (1992)). Thus, our new data preprocessing technique using percentile rank is to filter the noises of the input. The returns with small size will be removed. The returns with the larger size are likely to include the information that impact the subsequent stock returns. For the short-term prediction, the volatility is easier to predict. The idea of filtering out the input noises can also be later used in some other market predictions in order to improve the predictive power. If the upcoming news is predictable, then this can also be one of the evidence showing that the market is not efficient.

1.4 Hypothesis Development

Hypothesis 1: Input features from multiple lags and multiple difference orders and input filter will improve the predictive accuracy

One of the most popular trading strategies is a momentum strategy which uses the lag of returns as the input. However, the effects of the prior lags may still exist and these effects can affect the future price movement. This research extends the idea of using multiple lag orders. Also, the multiple difference orders are included as additional inputs. The implication of using the multiple lags is the existence of the previous lag with the momentum effect. Investors may look for the multiple returns from previous days to predict the return on the next day. This thesis also considers the rate of changes of the return and the higher order of rate of changes of the return as inputs. This idea has intuition from Taylor series which use the higher order differentiation to estimate the value of the function at a specific point. Prediction using the higher order difference is expected to improve the accuracy. Also, in the second part of this study, the small magnitudes of the input samples are not useful for the prediction, especially in high frequencies. Thus, discarding out the small magnitude samples will improve the predictive power.

Hypothesis 2: The two-stage model will outperform the individual neural network model in terms of the predictability

Currently machine learning models have played a big role in the stock market prediction since they such as a neural network can beat a linear model in terms of the accuracy of the prediction (Cao, et al. (2005)). The multi-stage model is developed for improving the accuracy (Hsu, et al. (2009)). This study extends the study on multi-stage models by using the SVM and the neural network model. The combination of these models is expected to increase the accuracy of the prediction. The base model for comparison is the neural network model due to its popular use.

CHAPTER II
LITERATURE REVIEW

The main contributions of this thesis are the use of the multiple lags and the difference orders to improve the accuracy of the SET50 index prediction, the use of the two-stage SVM-NN model to improve the prediction from NN, and the use of a percentile rank filter technique to improve the prediction accuracy of five-minute returns. Firstly, the feature extraction generating variables from lagging and differencing are applied. Secondly, after features are extracted, they are assigned as the inputs to the SVM. In this chapter, the SVM will be reviewed. Secondly, the values of small magnitudes are filtering out before the input is assigned to the SVM. The output from the SVM is then passed to the second stage model which is a neural network. Therefore, the neural network is also reviewed in this chapter.

2.1 Feature Extraction from a Time Series Data

There are many ways to improve the predictability of the predictive algorithm. One of those is the feature extraction. One can assume that prediction of a long-term stock using the NN with all signal inputs is better than the NN with the modified inputs. This assumption is proved to be wrong from many studies that used the indicators and the signals as the inputs of the NN rather than the unmodified time series of the stock prices or indices as the inputs (Trippi and Turban (1992), Tay and Cao (2001)). From this fact, the data that we will use as the inputs of the NN and the SVM will be modified by the methods presented from many studies, such as the modification and the feature extraction of data. But in order to improve the predictive accuracy, the input must associate with the output. Also, the time frequency of the prediction should relate to the input. For example, the tick data can be used to predict the short-term stock price but it is not appropriate to use the tick data to predict the quarterly stock price. Moreover, in the short-term prediction for the Thai stock market, the technical indicators can also be one of the best features due to the fact that technical indicators for the emerging market can provide the economically significant returns (Bessembinder and Chan (1995)).

According to Chaigusin, et al. (2008), Thai stock index is externally effected mostly by major countries. It is correlated with the major indices such as the Hangseng index, the Dow Jones index, the Nikkei index, the gold price, the minimum loan rate and the foreign exchange rate of Thai baht. These factors are good input candidates for the SET index prediction.

Rouwenhorst (1999) had shown that the momentum effect existed in the emerging markets. Cakici, et al. (2013) also show that there was a momentum in the emerging markets except for Eastern Europe. It implied that the previous return has a power to predict the future return. Thus, the previous returns could be considered as a useful input.

Moreover, Tolvi (2002) showed that the outlier are useful and informative for the stock prediction. Thus, in the second part of this thesis, the returns with small size are filtered out using the percentile rank of absolute returns. The remaining return samples are used as the prediction inputs.

2.2 The Neural Network

The neural network was firstly invented by McCulloch and Pitts (1943), from an idea of replicating the biological neural network that creates intelligence. The neural network structure comprises nodes organizing in the layers and the connections between nodes and the layers called arcs. There are weights assigned to all arcs connecting between nodes. Normally, there is no internal connection among nodes within the same layer, thus there are only weights between the nodes that are connecting to the different layers. A special bias node can be added in each layer with the value of one. It has arcs connecting to all nodes in the layer. Different number of nodes and number of layers in the neural network will provide different models. The number of nodes and number of layers can be determined by the method from Yao, et al. (1999). The function of the neural network is to predict or classify instances. Initially, the weights are randomly generated in the neural network. Then it uses the training dataset as the inputs with the target variable and then processes it to determine the target value. The neural network learns by adjusting weights so that it generates the best result.

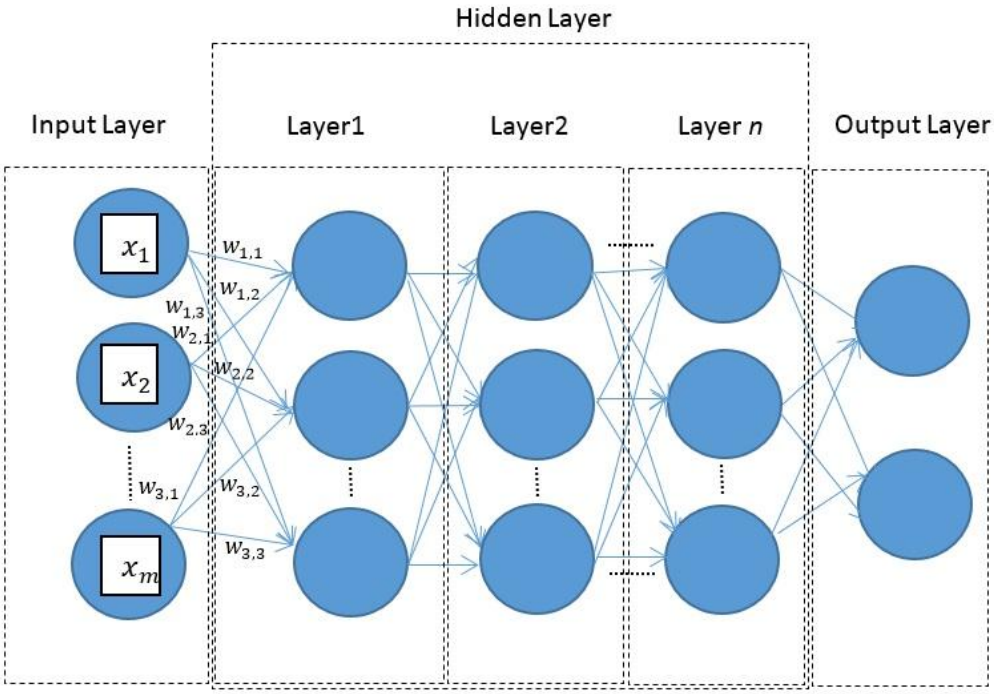


Figure 1 A structure of a neural network

The output value y_j is calculated by

$$y_j = f(\sum_{i=1}^{n} w_{i,j} x_i + \theta_i) \quad (2.1)$$

where y_j is the output value for a unit *j*

x_i is the input value at node *i* from the previous layer

$w_{i,j}$ is the weight attached between node *i* and node *j*

the function $f(\,)$ is the activation function

θ_i is the bias input at node *i*

n is the total number of nodes in the previous layer

There are also some choices that are needed to be determined to get the best performance of the neural network such as a learning algorithm, and an activation function. The learning algorithm is an algorithm that trains the neural network from the training dataset. The most used training algorithm is the back-propagating algorithm (Trippi and Turban (1992)) which adjusts the weights by the feedback errors from the output layer to the input layer. The activation function acts as a transfer function that converts the processed input value to the output on a specific range. Nowadays, most of the financial NN prediction algorithms use the back-propagation training algorithm

with the sigmoid activation function (Klimasauskas (1993)). The sigmoid activation function can be written as

$$f(z) = \frac{1}{1+e^{-z}} \tag{2.2}$$

In case of the sigmoid function, y_j can be calculated using

$$y_j = \frac{1}{1+e^{-(\sum_{i=1}^{m} w_{i,j}x_i+\theta_i)}} \tag{2.3}$$

After we calculate the output value at the output layer (y_{output}), the squared error (ε^2) between the output value at the output layer and the desired value (d) can be calculated by

$$\varepsilon^2 = (d - y_{output})^2 \tag{2.4}$$

The squared error is used in the back-propagation algorithm to update the weights in the neural network. It minimizes the squared error. In case of the sigmoid activation function, the adjusting weight can be calculated using

$$\Delta w_{i,j} = \alpha\varepsilon y_j(1 - y_j)x_i \tag{2.5}$$

where

$\Delta w_{i,j}$ is the change of the weight between node *i* and node *j*;

α is the learning rate constant;

x_i is the input value from node *i* (at the previous layer).

The weight in each node and each layer will be adjusted according to this algorithm respectively until all weights in the 1st layer are adjusted.

The inputs of the NN can be any time series data, however, recently, most of the inputs for the NN are derived from financial experts such as the technical indicators, the fundamental indicators and some others economic factors (Kwon and Moon (2007)).

2.3 The Support Vector Machine

A support vector machine is an algorithm that discriminates the input data into two desired groups by searching for the best hyperplane in n-dimension space. There are many types of hyperplanes such as a linear, a polynomial and a radial. By using the linear hyperplane as a classifier in the support vector machine, the hyperplane can be written as

$$g(\boldsymbol{t_k}) = \boldsymbol{w} \cdot \boldsymbol{t_k} + b \tag{2.5}$$

where $g(\boldsymbol{t_k})$ is the hyperplane function, $\boldsymbol{w}$ is the hyperplane's normal vector, $\boldsymbol{t_k}$ is the input vector at the *k*th sample and *b* is the offset. For example, let $q_k \in \{+1, -1\}$ where *q* denotes its class label. If we set the hyperplane function $g(\boldsymbol{t_k}) = 0$, then we will get $g(\boldsymbol{t_k}) > 0 \; for \; q_k = +1$ and $g(\boldsymbol{t_k}) < 0 \; for \; q_k = -1$. The hyperplane $g(\boldsymbol{t}) = 0$ separates the input data into two desired groups.

In order to determine the best hyperplane to separate instances into classes, the weight ($\boldsymbol{w}$) and the offset (b) must be optimized as the following convex optimization problem subject to the hyperplane condition by Smola and Schölkopf (2004).

$$minimize \;\; \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{k=1}^{l}\xi_k$$

$$subject\ to \;\; q_k(\boldsymbol{w}\cdot\boldsymbol{t_k} - b) \geq 1 - \xi_k \qquad (2.6)$$

where $\boldsymbol{w}$ is the hyperplane's normal vector, $\boldsymbol{t_k}$ is the input vector, b is the offset, q_k is the predefined label which $q_k \in \{+1, -1\}$, ξ_i is the slack variables and C is the complexity constant parameter.

The SVM can also be used in the stock market prediction (Huang, et al. (2005)). The results of the prediction outperform the NN-BP in some cases. The SVM is considered to be an excellent predictor, since the SVM model is simpler and capture the pattern by solving a quadratic optimization problem rather than solving a non-linear optimization, which is more complex as in the NN (Yao, et al. (1999)).

The popularity of the NN with BP in financial forecasting comes from the fact that the NN-BP provides better accuracy than other algorithms such as the SVM in most cases (Wu (2011)). However, in some cases, the SVM provides a better prediction, especially in terms of the direction according to Hsu, et al. (2009). Moreover, a two-stage model can improve the predictability in terms of the accuracy (Hsu, et al. (2009)). Therefore, a two-stage SVM-NN model that uses SVM to predict the return direction of the stock returns in the first stage then uses the NN with BP to predict the returns in the second stage is expected to improve the predictability in terms of accuracy.

2.4 Wilcoxon Signed-rank Test

Wilcoxon (1945) invented the Wilcoxon signed-rank test which is one of the nonparametric hypothesis tests. It is mainly used for compared two related samples, or repeated measurements on a single sample. The assumptions when using the Wilcoxon signed-rank test are

1. data are from the same population and come in pairs
2. population is not assumed to be normally distributed
3. each pair is chosen independently and randomly
4. the measured data are in an ordinal scale

The hypotheses of the test are

H_0: $the\ median\ difference\ between\ the\ pairs\ is\ zero$

H_1: $the\ median\ difference\ between\ the\ pairs\ is\ not\ zero$

W is the test statistic of the Wilcoxon signed-rank test. In case of the non-tie pairs are more than 10 pairs, W converges to the normal distribution.

The null hypothesis is rejected if $|Z| > Z_{critical}$.

The Z value of the Wilcoxon signed-rank test can be calculated by

$$Z = \frac{W}{\sigma_w} \tag{2.7}$$

where W is the sum of signed ranks. W and σ_w can be calculated from

$$W = \sum_{i=1}^{N_r}[sgn(x_{2,i} - x_{1,i})R_i] \tag{2.8}$$

$$\sigma_w = \sqrt{\frac{N_r(N_r+1)(2N_r+1)}{6}} \tag{2.9}$$

N_r is the number of sample pairs excluding the tie pairs

sgn is the sign function

$x_{2,i} - x_{1,i}$ is the difference value of the sample pair

R_i is the rank of pair *i* where the pair with least absolute difference value has rank equal to one.

CHAPTER III
METHODOLOGIES

In this chapter, there are two major parts. The first part describes about a method of the daily SET50 directional prediction using the support vector machine. The second part consists of the method in the two-stage model.

3.1 Method for Daily Directional Returns Prediction using the SVM

A. Data Preprocessing

1. Inputs for the support vector machine and the neural network

Inputs for the time series of the stock market index returns may come from the previous stock market index returns, the commodities price return and the exchange rate returns. This thesis will concentrate on the log return which is chosen as a representative of a return. The following table shows the first order difference inputs for the directional classification.

Table I First order inputs for directional classification

Inputs	**Calculation**
$R_{SET50_{t-1}}$	$\ln(SET50closed_{t-1}/SET50closed_{t-2})$
R_{HSI_t}	$\ln(HSIopened_t/\ HSIclosed_{t-1})$
$R_{Dow_{t-1}}$	$\ln(DJclosed_{t-1}/\ DJclosed_{t-2})$
$R_{STI_{t-1}}$	$\ln(STIopened_t/\ STIclosed_{t-1})$
$R_{FTSE_{t-1}}$	$\ln(FTSEclosed_{t-1}/FTSEclosed_{t-2})$
$R_{Gas_{t-1}}$	$\ln(Gasclosed_{t-1}/Gasclosed_{t-2}\ )$
$R_{Gold_{t-1}}$	$\ln(Goldclosed_{t-1}/Goldclosed_{t-2})$
$R_{Oil_{t-1}}$	$\ln(Oilclosed_{t-1}/Oilclosed_{t-2})$
$R_{THBUSD_{t-1}}$	$\ln(THBUSDclosed_{t-1}/THBUSDclosed_{t-2})$

- $R_{SET50_{t-1}}$ refers to the SET50 index daily return at time t-1 (where t refer to today, t-1 refer to yesterday and t-2 refer to the day before yesterday) and $SET50closed_{t-1}$ refer to the yesterday closing price of the SET50 index The

calculation of the Hangseng index return (R_{HSI_t}) is based on the close-to-open Hangseng index return due to time zone.

- R_{Dow} refers to the Dow Jones index daily return $DJclosed$ refers to the Dow Jones closing price
- R_{FTSE}refers to the FTSE index daily return and $FTSEclosed$ refers to the FTSE index closing price
- R_{Gas} refers to the Natural Gas price return (Usd/ MMBTU) and $Gasclosed$refers to the Natural Gas Henry Hub closing price
- R_{Oil}refers to the Crude Oil-WTI Spot Cushing price return (Usd/BBL) and $Oilclosed$refers to the Crude Oil-WTI Spot Cushing closing price
- R_{Gold} refers to the Gold Bullion LBM price return (Usd/Troy Ounce) and $Goldclosed$ refers tothe Gold Bullion LBM closing price
- R_{THBUSD} refers to the Thai Baht to US Dollar exchange rate return and $THBUSDclosed$ refers to the Thai Baht to US Dollar closing exchange rate price

2. Second order and higher order differences

The second order inputs are generated from the difference of the adjacent first order input returns; for instances, the second order difference of the SET50 price index at time *t*-1 is calculated by

$$\Delta^2 R_{SET50_{t-1}} = \Delta R_{SET50_{t-1}} - \Delta R_{SET50_{t-2}} \quad (3.1)$$

By differencing this repeatedly, the higher order difference inputs are generated. For example, in the case of the SET50 price index, the third order differences and the *n*th order differences can be calculated as in the equation (3.2) and (3.3), respectively. The third order differences SET50 index return

$$\Delta^3 R_{SET50_{t-1}} = \Delta^2 R_{SET50_{t-1}} - \Delta^2 R_{SET50_{t-2}} \quad (3.2)$$

The n^{th} order differences SET50 index return

$$\Delta^{\mathrm{n}} R_{SET50_{t-1}} = \Delta^{\mathrm{n}-1} R_{SET50_{t-1}} - \Delta^{\mathrm{n}-1} R_{SET50_{t-2}} \quad (3.3)$$

3. Label preprocessing

The label of the directional prediction of the SET50 index will be specified as two classes which are “Up” and “Down”. According to these labels, the first value is “Up” which means the return of the SET50 index has exceeded a threshold which is set to 0% in this thesis. Note that if

$ln(\frac{SET50closed_t}{SET50closed_{t-1}}) > 0$ then the return at time t will be classified as “Up”. Also if

$ln(\frac{SET50closed_t}{SET50closed_{t-1}}) \leq 0$ then the return at time t will be classified as “Down”.

Table II Output Label in Classification

Classification	**Criteria**
Up	$\ln\left(\frac{SET50closed_t}{SET50closed_{t-1}}\right) > 0\%$
Down	$\ln\left(\frac{SET50closed_t}{SET50closed_{t-1}}\right) \leq 0\%$

4. Data time frame, data filtering and data normalizing

The time frame for building and testing the model is set from 1/1/2002 to 10/6/2013. All datasets are related to the stock market index or they are factors that usually affect the Thai stock market (Khumyoo (2000)).

The non-trading day data are dropped from both the input data and the output data, thus there will be no prediction on the non-trading day.

Moreover, after all non-trading day data are dropped, all inputs are normalized, and these normalized data are fed to the SVM and the NN.

B. Directional Classification Using the Support Vector Machine

A dataset for the support vector machine is split into two subsets which are a training set (in-sample) and a testing set (out-of-sample). According to the time frame, a total number of samples are 3,489.The first 80% of the dataset will be used as an in-sample set which contains the first 2,791 samples, and the number of samples in the testing set is 20% of the data which contains 698 samples.

The in-sample set is transformed and fed to the support vector machine to build the directional predictive model. After the model is constructed, it will be applied to the out-of-sample set and the accuracy percentage of the out-of-sample set will be measured.

C. Directional Classification Using the Neural Network

The training and testing sets of the neural network are the same as in the first part. Thus the training set of 2791 samples will be fed to the neural network, while the parameter settings are set according to Hagan, et al. (1996); see Table III for the detailed parameters. In this thesis, the number of hidden layers in the neural network is set to 1. The number of nodes is varying. It depends on the number of attributes and the number of classes. The other parameters which are a training cycle, a learning rate, momentum and an error epsilon are set to 500, 0.3, 0.2 and 0.00001 respectively. After the neural network model is trained, the model will be applied to the out-of-sample set to evaluate its accuracy. The accuracy of the NN are compared with the SVM model. Our implementation is designed as in Figure 2 and the implementation of the SVM model in Rapidminer software is shown in Figure 3.

Table III The neural network parameter settings

NN	Parameters
Hidden layers	1
Hidden nodes	(number of attributes + number of classes)/2 +1
Training cycles	500
Learning rate	0.3
Momentum	0.2
Error epsilon	0.00001

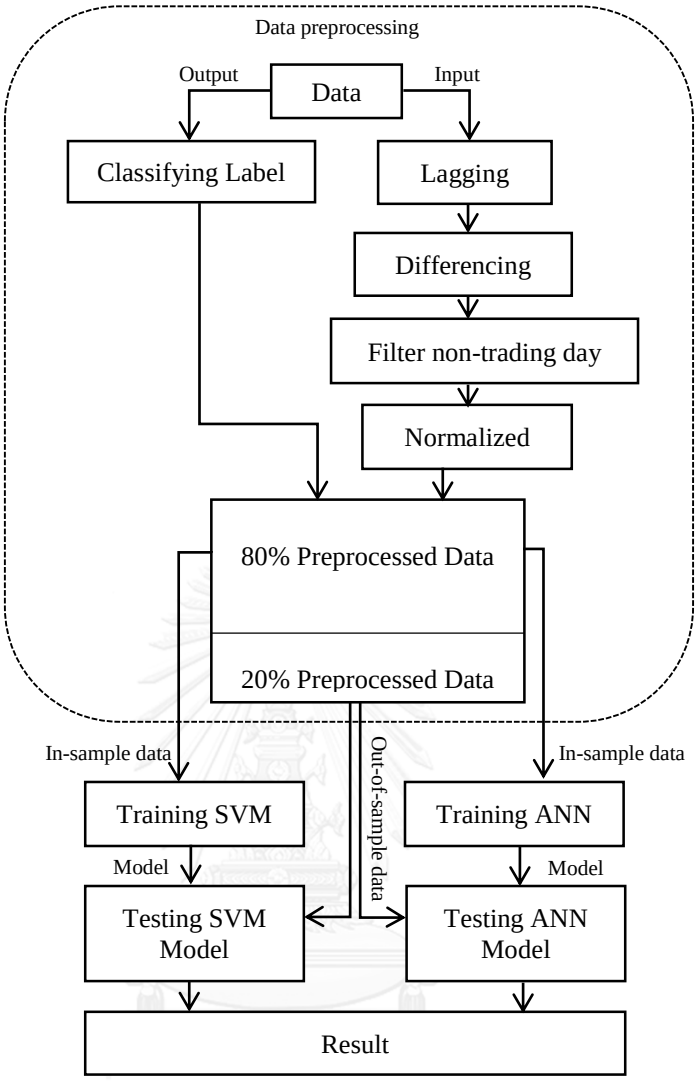


Figure 2 Overall method of creating daily directional prediction model

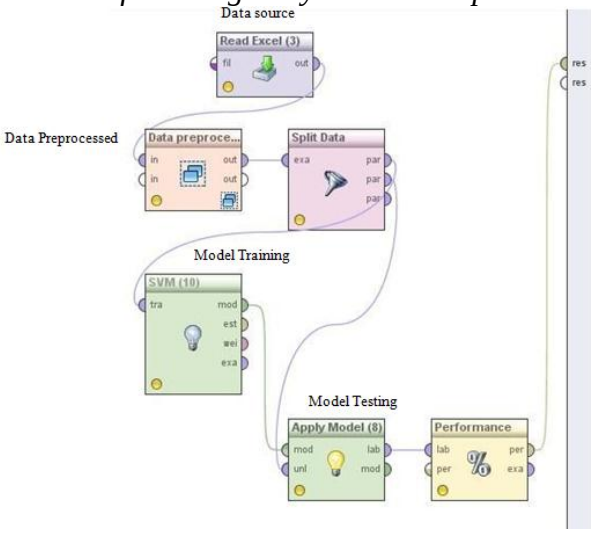


Figure 3 Overall of creating daily directional prediction using SVM model in Rapidminer

3.2 Method for Five-Minute Return Prediction Using the Two-stage Model

A. Attributes for the First Stage of the Two-stage Model

The first part of this thesis builds the SVM model to predict the direction of the SET50 index. Normally the SET50 index is affected by the major indices such as the oil and the gas prices, THB/USD exchange rate (Chaigusin, et al. (2008)). In this thesis, only the Hangseng and the Dow Jones indices are selected as attributes.

The Hangseng and the Dow Jones indices returns are collected at the different time intervals (5, 10, 15, and 30 minutes) as the inputs used to predict the 5minute returns of the SET50 index. Table IV illustrates the attributes for the directional prediction.

All attributes are not used simultaneously. First, we use the model with only one attribute. Thus, there are C(8, 1) = 8 predictive models where C(8,1) is the number of combinations when choosing 1 object from 8 distinct objects. Then, the combinations of two attributes are selected. There are C(8, 2) = 28 predictive models. Lastly, the combinations of three attributes are selected, thus there are C(8, 3) = 56 predictive models. Before a training set with the chosen attributes is fed to the model, the values of each attribute will be filtered using the percentile rank.

In this thesis, we also try to determine the effect of the size of the stock returns at various percentile ranks on the prediction performance. First, we determine the size of each attribute return by taking the absolute of the return and rank them. Then, if the size of the inputs and the stock index futures return fails to reach above the chosen percentile rank, then those will be filtered out. In this thesis, we use the following percentile ranks: 0%, 10%, 20%, 30%, 40%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, and 95%. The predictive performance will be discussed in the next chapter.

Table IV Attributes for the directional classification

Input	Formula
$R_{HSIF_{5Min}}$	$\ln\left(\frac{HSIF_{t-1}}{HSIF_{t-2}}\right)$
$R_{HSIF_{10Min}}$	$\ln\left(\frac{HSIF_{t-1}}{HSIF_{t-3}}\right)$
$R_{HSIF_{15Min}}$	$\ln\left(\frac{HSIF_{t-1}}{HSIF_{t-4}}\right)$
$R_{HSIF_{30Min}}$	$\ln\left(\frac{HSIF_{t-1}}{HSIF_{t-7}}\right)$
$R_{DJF_{5Min}}$	$\ln\left(\frac{DJF_{t-1}}{DJF_{t-2}}\right)$
$R_{DJF_{10Min}}$	$\ln\left(\frac{DJF_{t-1}}{DJF_{t-3}}\right)$
$R_{DJF_{15Min}}$	$\ln\left(\frac{DJF_{t-1}}{DJF_{t-4}}\right)$
$R_{DJF_{30Min}}$	$\ln\left(\frac{DJF_{t-1}}{DJF_{t-7}}\right)$

From Table IV, $R_{HSIF_{5Min}}$ denotes the five-minute Hangseng index future returns at 5 minutes before the current time;

$HSIF_{t-1}$ refers to the Hangseng index future closed price at 5 minutes before the current time;

$HSIF_{t-2}$ refers to the Hangseng index future closed price at 10 minutes before the current time;

$R_{HSIF_{10Min}}$ denotes the ten-minute Hangseng index future returns at 5 minutes before the current time;

$HSIF_{t-1}$ refers to the Hangseng index future closed price at 5 minutes before the current time;

$HSIF_{t-3}$ refers to the Hangseng index future closed price at 15 minutes before the current time;

$R_{HSIF_{15Min}}$ denotes the fifteen-minute Hangseng index future returns at 5 minutes before the current time;

$R_{HSIF_{30Min}}$ denotes the thirty-minute Hangseng index future returns at 5 minutes before the current time;

$R_{DJF_{5Min}}$ denotes the five-minute Dow Jones index future returns at 5 minutes before the current time;

DJF_{t-1} refers to the Dow Jones index future closed price at 5 minutes before the current time;

DJF_{t-2} refers to the Dow Jones index future closed price at 10 minutes before the current time;

$R_{DJF_{10Min}}$ denotes the ten-minute Dow Jones index future returns at 5 minutes before the current time;

DJF_{t-1} refers to the Dow Jones index future closed price at 5 minutes before current time;

DJF_{t-3} refers to the Dow Jones index future closed price at 15 minutes before the current time;

$R_{DJF_{15Min}}$ denotes the fifteen-minute Dow Jones index future returns at 5 minutes before the current time;

$R_{DJF_{30Min}}$ denotes the thirty-minute Dow Jones index future returns at 5 minutes before current time;

B. Attributes for the Second -stage of the Two-stage Model

The input attributes for the second-stage model are almost the same as the inputs of the first stage model, see Table IV. The combination of the Hangseng and the Dow Jones futures input attributes in the second stage are the same as the first stage. An additional input attribute comes from the predictive result of the first stage model which predicts the direction of the stock return.

C. Label Preprocessing

Since there are two stages of the predictions, there are also two stages for label preprocessing. The label for the first stage is the direction of the SET50 index five-minute return which can be calculated by $\ln\left(\frac{\mathrm{SET50_t}}{\mathrm{SET50_{t-1}}}\right)$ where $\mathrm{SET50_t}$ is the current SET50 index price and $\mathrm{SET50_{t-1}}$ is the SET50 index price at five minutes before the current time. Thus, the result of the classification will either be "1" for "up" or "-1" for "down" (see Table V).

Table V Label classification for the first stage model

Classification	**Criteria**
1	$\ln\left(\frac{SET50_t}{SET50_{t-1}}\right) > 0\%$
-1	$\ln\left(\frac{SET50_t}{SET50_{t-1}}\right) \leq 0\%$

The label in the second stage is the return of the five-minute SET50 index, which is calculated by the same formula, $ln\left(\frac{SET50_t}{SET50_{t-1}}\right)$. However, there is no need to label the data in the second stage since the input of the neural network must be numeric.

D. Data Period Used, Additional Filtering and Normalizing

In this thesis, we use the data period from 24/6/2013 to 18/11/2015. The number of five-minute SET50 return data points in this period is more than 10,313 data points. However, the non-trading time data, at the open time data and at the closed time data are filtered out. After the non-trading day data are dropped, we filter out the input based on the absolute of the five-minute SET50 return using the percentile rank. The filtering level for the absolute of the five-minute SET50 returns are set at 10, 20, 30, 40, 50, 55, 60, 65, 70, 75, 80, 85, 90 and 95 percentile ranks. Then all inputs are normalized before passing to the two-stage model.

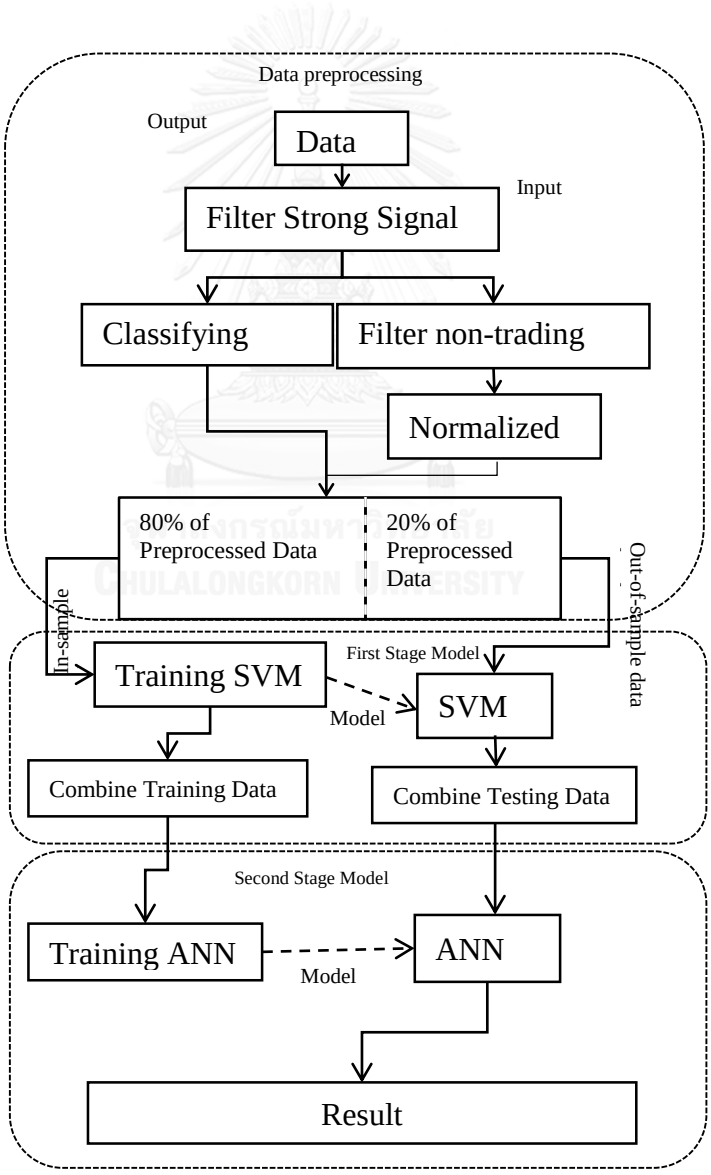


Figure 4 Flow diagram of creating the two-stage model

E. The Processes of the Two-stage Model Prediction

After the data preprocessing, the preprocessed data will be split into the training set and the testing set. The training set contains 80% of the preprocessed data and the testing set contains 20% of the preprocessed data. The training data will be fed into the SVM in the first stage. The result of the directional prediction from the trained SVM model will be later combined with the training preprocessed data as an additional attribute then all attributes are fed to train the second-stage model, the neural network. During the test phase, the testing data set will be passed through the trained SVM model to get the direction before merging with its input, and being fed to feed to the neural network model in the second stage.

Finally, the two-stage model predicts the five-minute SET50 index return as a result. The flow chart of the whole process is shown in Figure 4. Also, in Figure 5, the two-stage model using Rapidminer implementation is shown. The out-of-sample performance of the two-stage prediction will later be discussed in the next section.

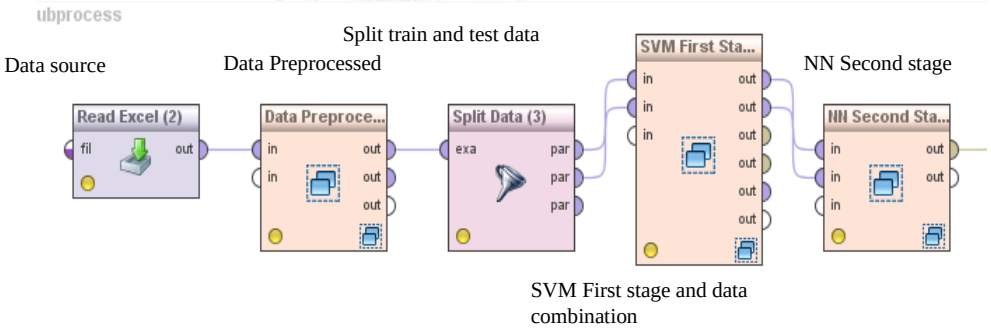


Figure 5 The two-stage model using Rapidminer

Table VI Parameters setting for the support vector machine

SVM	Kernel Type	Kernel cache	C	Convergence Epsilon	Max iteration	L positive	L negative	ε	ε plus	ε minus
Parameters	Dot	200	1	0.001	100000	1	1	0	0	0

CHAPTER IV
EXPERIMENT AND RESULT

4.1 Experimental Result for Daily Directional Prediction

A. The Support Vector Machine Parameter Settings

In this paper, the standard SVM is used due to its popularity and simplicity in applying the model in financial prediction (Pérez-Cruz, et al. (2003)). All of the parameters in SVM are set according to Table IV. The dot kernel type is used. Convergence epsilon is set to 0.001. The maximum number of iterations for the support vector machine setting is 100,000. The parameter C, which represents the tolerance of the misclassification, is set to 1.

B. Accuracy Percentage

To estimate the accuracy of the algorithm, the number of times of the correct predictions is counted against the number of all out-of-sample instances. Thus, the accuracy percentage can be determined by

$$Accuracy\ \ Percentage = \frac{Amount\ of\ \ Correct\ \text{Pr}ediction}{Total\ \ Number\ of\ \ Data} \times 100 \qquad (4.1)$$

Table VII Accuracy percentage of the support vector machine classification according to the number of lags and difference order of inputs

	Number of lags								
Difference orders		1	2	3	4	5	6	7	8
	1	64.68%	64.07%	64.83%	65.44%	65.90%	64.37%	64.37%	65.44%
	2	64.53%	65.75%	66.06%	66.06%	64.83%	64.68%	66.06%	65.29%
	3	66.21%	66.21%	66.06%	64.83%	64.83%	64.68%	65.14%	64.68%
	4	66.36%	67.77%	65.90%	64.98%	65.14%	64.98%	64.83%	64.83%
	5	66.67%	65.75%	65.29%	65.29%	64.83%	64.68%	64.83%	64.07%
	6	66.67%	65.14%	64.53%	64.98%	64.22%	64.07%	64.07%	64.68%
	7	65.75%	64.68%	65.29%	64.22%	63.15%	64.07%	64.68%	64.37%
	8	65.29%	65.14%	65.90%	64.37%	64.22%	63.76%	64.68%	65.29%

C. Directional Classification Result Using the Support Vector Machine

After testing the model with the 698 out-of-sample instances using the higher order difference inputs and the higher order lag inputs, the results are summarized in Table VII. Table VII shows the accuracy percentage for each model with a particular number of lags and a particular number of difference orders of the inputs. The model with the higher difference and the lag orders also includes the difference and lags of lower orders. For example, the model with "Difference orders 4" means the model will include the 1st, 2nd, 3rd and 4th difference orders of every input variables and the model with "lag 3" includes one lag, two lags and three lags of the input attributes.

D. Results Analysis

Figure 6 plots the accuracy of the model as a function of the difference order for lag 1 to lag 8. Figure 7 shows the averaged accuracy across all the lags as a function of the difference order. Figure 8 plots the accuracy of the model as a function of the lag for difference order 1 to difference order 8. Figure 9 shows the averaged accuracy across all the difference orders as a function of the lags. According to the result of the prediction of the SET index returns using the support vector machine shown in Figure 7 and Figure 9, the prediction accuracy tends to increase when the number of difference orders and the number of lags increase, but the prediction accuracy tends to decrease after the number of difference orders and the number of lags become too large. For instance, according to Figure 6, where the number of lag is 2, the predictive accuracy is more likely to increase from 64.07% to 67.77% when the difference orders increases from 1 to 4. The accuracy reaches its peak at 67.77% when using 4 difference orders but after the difference orders goes beyond 4 the accuracy is more likely to decrease. From Figure 7, the average accuracy percentage across all lags shows the tendency of increasing accuracy when the difference orders are less than 4.

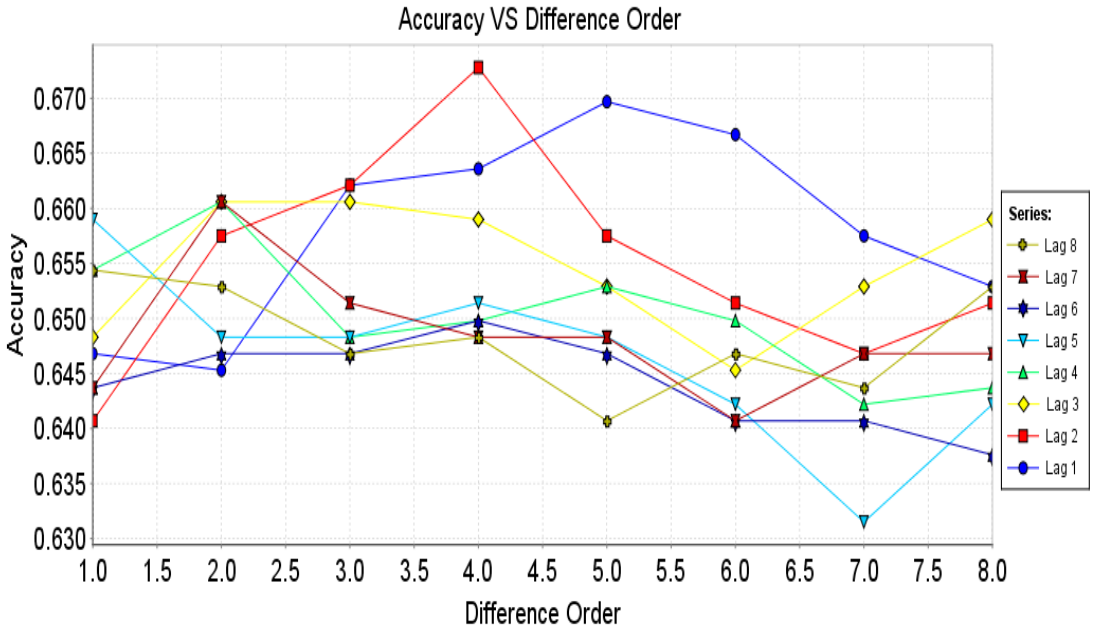


Figure 6 Accuracy percentage versus difference order of inputs at each lag

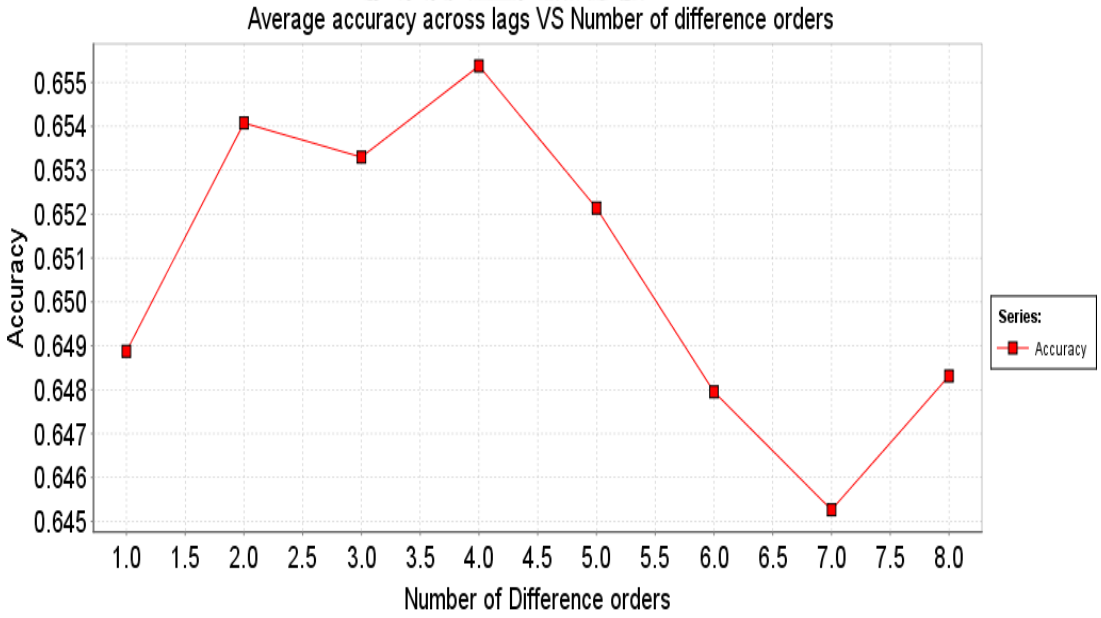


Figure 7 Average accuracy percentage along the number of lags versus the difference orders of inputs

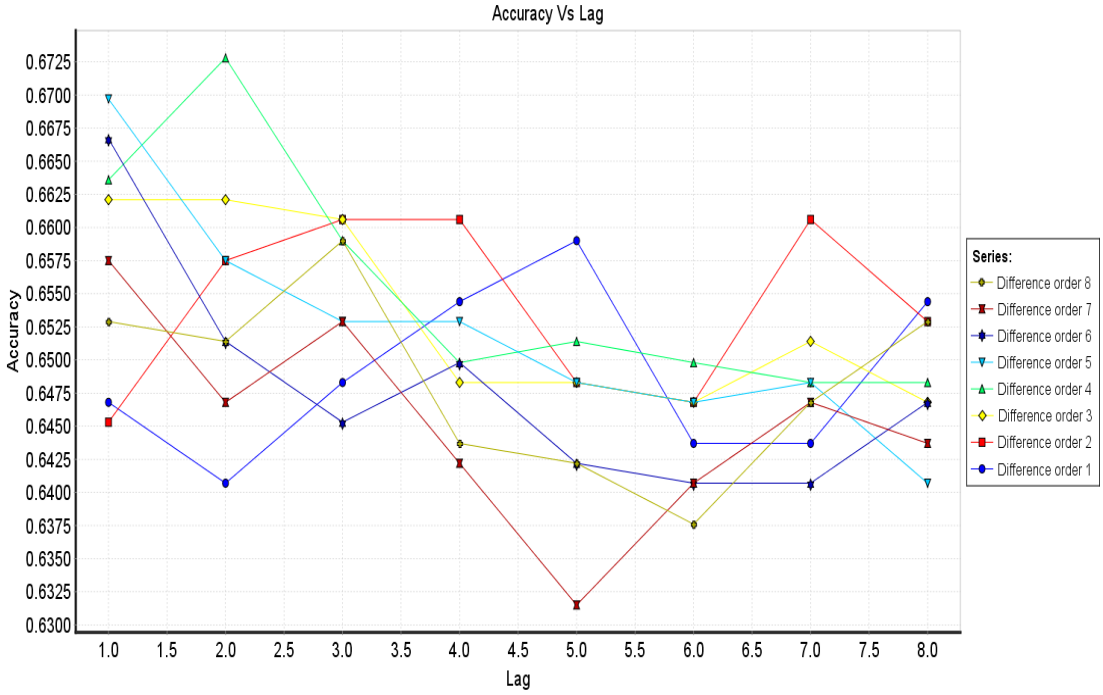


Figure 8 Accuracy percentage versus the number of lags at each the difference order

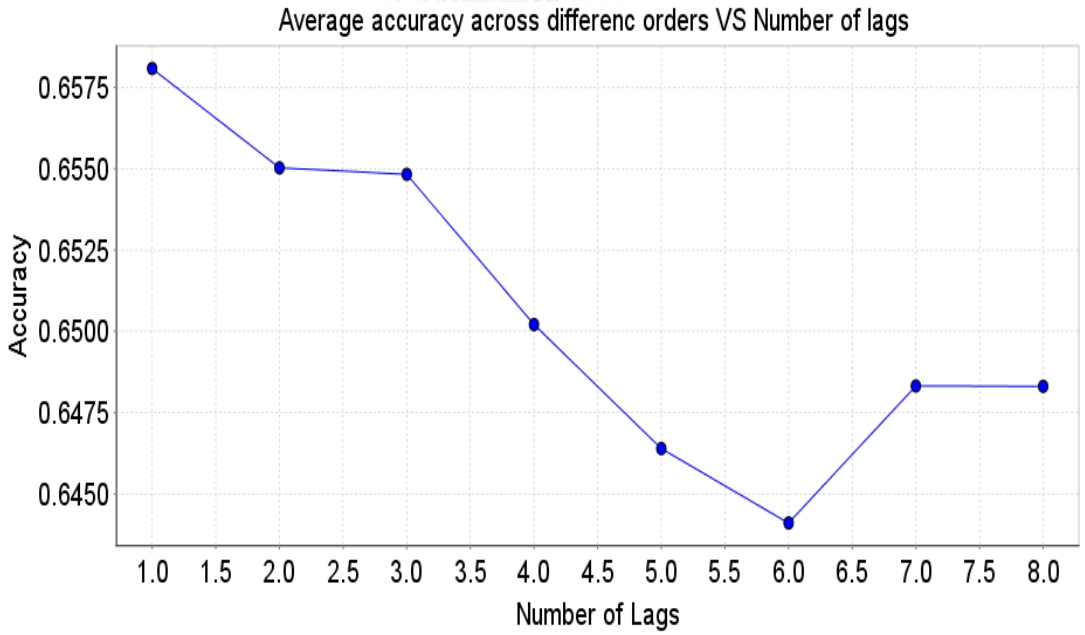


Figure 9 Average accuracy percentage along the number of difference versus the number of lags

In terms of difference orders, Table VIII shows the p-value of the test statistics on the null hypothesis which is the additional difference order provide indifference accuracies. If the p-value is below 0.05, it means that there is a significant change in accuracies when adding a difference order at 95% confidence. Table IX shows the p-value of the test statistics on the null hypothesis which is the specific number of difference orders provide indifference accuracies to the first difference order. If the p-value is below 0.05, it means that there is a significant change in accuracies when compared with the first difference order at 95% confidence. The test statistic from Table VIII and Table IX shows that the increasing of the accuracy is not statistically significant. After the difference order goes beyond 4, it shows the tendency of decreasing in the accuracy.

Table VIII The Wilcoxon signed-rank test statistic on accuracy change due to an additional difference order

Test Statistics							
	Order2 - Order1	Order3 - Order2	Order4 - Order3	Order5 - Order4	Order6 - Order5	Order7 - Order6	Order8 - Order7
P-value	.161	.686	.291	.235	.055	.271	.150

*The test statistics show the p-value of the null hypothesis which is the additional difference orders provide indifference in accuracies. If the p-value is below 0.05, it means that there is a significant change in the accuracies when adding a difference order at 95% confidence.

Table IX The Wilcoxon signed-rank test statistic on accuracies of 1st order to other difference orders

Test Statistics							
	Order2 - Order1	Order3 - Order1	Order4 - Order1	Order5 - Order1	Order6 - Order1	Order7 - Order1	Order8 - Order1
P-value	.161	.263	.261	.400	.574	.624	.944

*The test statistics show the p-value of the null hypothesis which is the specific number of difference orders provides indifference in accuracies to the first difference order. If the p-value is below 0.05, it means that there is a significant change in the accuracies when compared with the first difference order at 95% confidence.

In terms of lags, Table X shows the p-value of the test statistics on the null hypothesis which is the additional lag provide indifference accuracies. If the p-value is below 0.05, it means that there is a significant change in accuracies when adding an additional lag at 95% confidence. Table XI shows the p-value of the test statistics on

the null hypothesis which is the specific number of lag provide indifference accuracies to the first difference order. If the p-value is below 0.05, it means that there is a significant change in accuracies when compared with the first difference order at 95% confidence. The test statistic from Table XI shows that there is a significant change in the accuracy when using six lag returns. As shown in Figure 8 and Figure 9, cumulating more lags as the inputs result in the drop of the accuracy since more lags of the stock index returns do not provide essential information on predicting the stock directions.

Table X The Wilcoxon signed-rank test statistic on accuracy changes due to an additional lag order

Test Statistics							
	Lag2 - Lag1	Lag3 - Lag2	Lag4 - Lag3	Lag5 - Lag4	Lag6 - Lag5	Lag7 - Lag6	Lag8 - Lag7
P-value	.352	.833	.116	.150	.119	.058	1.000

*The test statistics show the p-value of the null hypothesis which is the additional lag provide indifference in accuracies. If the p-value is below 0.05, it means that there is a significant change in the accuracies when adding a lag at 95% confidence.

Table XI The Wilcoxon signed-rank test statistic on accuracy of 1st lag order to other lag orders

Test Statistics							
	Lag2 - Lag1	Lag3 - Lag1	Lag4 - Lag1	Lag5 - Lag1	Lag6 - Lag1	Lag7 - Lag1	Lag8 - Lag1
P-value	.352	.440	.106	.058	.017	.079	.062

*The test statistics show the p-value of the null hypothesis which is the specific number of lags provides indifference in accuracies to the first lag. If the p-value is below 0.05, it means that there is a significant change in the accuracies when compared with the first difference order at 95% confidence.

At the peak accuracy of 67.77%, the input variables consist of the four difference orders and two lags. Table XII shows the detailed of its performances which are precisions and recalls. The precision for “Up” is high at 69.40% which means when the model predicts “Up” almost 70% of the actual result is actually “Up”. Also, the recall for “Up” of this model seems to be high at the value of 73.41%, which means the model can capture 73% of “Up” SET50 index directional returns in the test period.

Therefore, the results at low difference orders from 1 to 4 show slight accuracy increase with the increase of the difference orders but it is not statistically significant. Thus, the difference order does not provide enough information for the directional prediction. However, at the high difference orders from 5 to 7, increasing too much difference order show the decrease in the accuracy.

Table XII Examples of the Precision and the Recall of the daily SET50 directional prediction using the SVM model

Variable	Accuracy	Precision Up	Precision Down	Recall Up	Recall Down
#Diff. orders 5, #Lag 1	67.77%	68.00%	64.73%	73.70%	58.19%
#Diff. orders 4, #Lags 2	66.67%	69.40%	65.54%	73.41%	60.98%

In order to provide the best accuracy using this finding, an additional difference order or the lags of inputs should be added one by one as long as the accuracy is still improving. When the accuracy is dropping, no more difference order or lag should be added as inputs.

Table XIII reports the stock predictive accuracy on each study, ranging from 53.3% to 67.28%. According to the predictive accuracy from Table XIII, the support vector machine with the higher difference orders and lags proposed in this thesis yields good directional predictive accuracy compared with other models. The cause of high predictive accuracy in this study can be occurred from the HSI return and STI return input factors. The HSI return and STI return in this thesis used the opened to closed return, see Table I. These factors are already included the morning information just before the SET50 market is open. Thus, these factors are very informative to the SET50 market. However, the directional SET50 index prediction calculated from closed to closed return, thus it is impossible to take the position at the close of the previous day when using this model. The model is only useful for the forecasting propose.

Table XIII Daily stock indices prediction accuracy comparisons

Author	**Model**	**Daily Prediction Accuracy (%)**	**Exchange Market**
Fernandez-Rodriguez et al. (2000)	NN	58	IGBM
Harvey et al. (2000)	NN	59	NYSE
Lendasse et al. (2000)	RBFN	57.2	BSE20
Francis E.H. Tay, Lijuan Cao(2001)	SVM	58.29	CME-SP
Kim K-j. (2003)	SVM	57.8313	KOSPI
Halliday (2004)	NN	55.57	NYSE
Doesken et al. (2005)	M-FIS	53.31	NYSE
Doesken et al. (2005)	TS-FIS	56	NYSE
S.-H. Hsu et al. (2009)	SVM+SOM	59.07	HSI
P. Samurwong et al. (Proposed)	SVM	67.28	SET

- NN refers to a neural network
- RBFN refers to a radial basis function neural network
- SVM refers to a support vector machine
- M-FIS refers to the Mamdani fuzzy system
- TS-FIS refers to the Takagi-Sugeno fuzzy system
- SOM refers to a self-organized map

Moreover, when compare the accuracy of the SET50 index return prediction of the SVM model with the NN model using 1 to 5 lags and 1 to 5 difference orders, the SVM accuracies statistically beat the accuracies of the NN model in the SET50 index return with 95 percent confidence and p-value = 0.000012 by the Wilcoxon signed-rank test. This result harmonizes with the result in Kim (2003), which showed that the SVM beat the NN in terms of the stock directional prediction.

From the daily SET50 return prediction performance, we can see that the momentum effect helps predict the direction of the SET50 index returns. This evidence also agrees with some studies such as Rouwenhorst (1999) which stated that there is a momentum effect in the emerging markets and Cakici, et al. (2013) which showed the evidence of the momentum effect in the emerging market. Moreover, the factors that affect the SET50 index are not only its previous lag but also those multiple lags and those multiple difference orders. According to many studies in the market efficiency (Schwert (2003)), this implies the inefficiency in the Thai market.

4.2 Experimental Result for the Two-Stage Model Prediction Compared with the NN Model Prediction

A. Benchmark Performance

In this paper, the performance of the two-stage model is determined using two benchmarks. The first one is the accuracy percentage which can be calculated by the equation below

$$Accuracy\ \ Percentage = \frac{Number\ of\ Correct\ \Pr ediction}{Total\ \ Number\ of\ \ \Pr ediction} \times 100 \qquad (4.1)$$

Another benchmark used to measure the performance of the model is the mean absolute error which can be calculated by

$$Mean\ \ Absolute\ \ Error = \frac{\sum \left| \Pr edicted\ \ Value - Actual\ \ value \right|}{Total\ \ Number\ of\ \ \Pr ediction} \times 100 \qquad (4.2)$$

F. The Two-stage Model Performance Result

According to the data preprocessing, the selected number of attributes are one attribute, two attributes and three attributes. Table XIII shows the average mean absolute error of the two-stage model and the NN model across all possible combinations of attributes at each level of percentile filtering. Also, Table XVI shows the two-stage model and the NN model's average mean absolute error performance across all possible combination of attributes at each level of percentile filtering.

B. Result Analysis

The values in Table XIV are the average mean absolute errors across all combinations of attributes for all numbers of attributes, and percentile ranks. According to the results in Table XIV, the average mean absolute errors across all of the percentile ranks of the neural network model and the two-stage model are slightly different. The average mean square error across all percentile ranks using one attribute in the neural network model is 0.000832. However, the average mean absolute error of the two- stage model using one attribute is 0.000846 which is slightly more than the error in the neural network model. However, the two-stage model using two attributes shows a better performance compared with the neural network model. The average mean square error across percentile ranks using two attributes in the neural network model is 0.000859, but the average mean absolute error of the two-stage model using two attributes is 0.000854. For three attributes, the neural network shows less error than the two-stage model. Upon compared the performance of both models, it is inconclusive that which model is better in terms of the mean absolute error. Also, from Table XVI, the Wilcoxon signed-rank test shows that the MAE of the two-stage model is not significantly better than the MAE of the neural network model. This result confirms the inconclusiveness of the MAE between these models.

Table XIV The two-stage model and the NN model average mean absolute error performance across all possible combination of attributes at each level of percentile filtering

Number of Attribute	1	1	2	2	3	3
Percentile Rank	NN Average of MAE ($\times 10^{-4}$)	Two-stage Average of MAE ($\times 10^{-4}$)	NN Average of MAE ($\times 10^{-4}$)	Two-stage Average of MAE ($\times 10^{-4}$)	NN Average of MAE ($\times 10^{-4}$)	Two-stage Average of MAE ($\times 10^{-4}$)
0%	7.95	8.4	8.18	8.11	8.18	7.89
10%	7.89	7.84	8.23	8.02	8	8.11
20%	7.66	8.31	7.84	8.04	7.97	8.06
30%	9.41	8.81	8.09	7.75	7.94	8.1
40%	8.55	7.93	7.86	8.34	7.88	8.43
50%	8.07	8.13	8.48	8.25	8.31	8.31
55%	8.76	7.63	8.58	8.44	8.44	8.08
60%	7.81	7.81	8.38	8.19	8.44	8.92
65%	8.31	8.07	8.43	8.28	8.42	8.57
70%	8.19	8.53	8.66	8.67	8.39	8.67
75%	8.38	8.56	8.83	8.97	8.52	8.91
80%	7.89	9.26	8.65	8.71	9.02	8.81
85%	8.1	9.98	8.98	9.71	9.76	9.42
90%	8.92	9.25	9.83	8.98	9.53	9.63
95%	8.95	8.37	9.91	9.58	10.4	10.6
Average All Percentile Rank	8.32	8.46	8.59	8.54	8.61	8.70

On the other hand, from Table XV, the average accuracy across percentile ranks using one attribute in the neural network model is 0.503 or 50.3%. However, the average accuracy of the two-stage model using one attribute is 0.506 or 50.6% which is slightly larger than the accuracy of the neural network model. Also, for two attributes and three attributes, the two-stage model shows better performance. Since the average accuracy across percentile ranks using two attributes and three attributes in the neural network model is 51.4% and 51.29%, respectively, but the average accuracy of the two-

stage model using two attributes and three attributes is 51.8% and 53.2%, respectively. The two-stage model shows slightly more accuracies than the neural network model. The result from the Wilcoxon signed-rank test shows that the two-stage model predictive accuracies is significantly higher than the accuracies of the neural network model at 90% confidence with p-value = 0.099. It is quite obvious that adding the direction from the support vector machine prediction to the neural network model will improve the accuracy.

Table XV The two-stage model and the NN model average mean absolute error performance across all possible combination of attributes at each level of percentile filtering.

Number of Variable	1	1	2	2	3	3
Percentile Rank	NN Average of Accuracy	Two-stage Average of Accuracy	NN Average of Accuracy	Two-stage Average of Accuracy	NN Average of Accuracy	Two-stage Average of Accuracy
0%	0.511	0.508	0.502	0.507	0.499	0.504
10%	0.497	0.496	0.503	0.503	0.503	0.506
20%	0.502	0.499	0.508	0.499	0.503	0.504
30%	0.494	0.505	0.504	0.505	0.505	0.504
40%	0.500	0.495	0.506	0.499	0.514	0.502
50%	0.499	0.493	0.509	0.510	0.510	0.503
55%	0.498	0.505	0.510	0.502	0.514	0.507
60%	0.500	0.517	0.502	0.502	0.505	0.516
65%	0.500	0.489	0.500	0.524	0.514	0.515
70%	0.495	0.520	0.509	0.512	0.531	0.519
75%	0.499	0.509	0.522	0.521	0.531	0.536
80%	0.515	0.513	0.519	0.534	0.542	0.551
85%	0.506	0.484	0.535	0.515	0.557	0.559
90%	0.508	0.498	0.532	0.556	0.578	0.594
95%	0.521	0.559	0.553	0.584	0.622	0.656
Average All Percentile Ranks	0.503	0.506	0.514	0.518	0.529	0.532

Table XVI The comparison of mean absolute error and the accuracy of the two-stage model and the neural network model using paired sample t test

Paired Samples t Test		
	TwoStageMAE - NNMAE	TwoStageAcc - NNAcc
P-value	0.212	0.099

*The test statistics show the p-value of the null hypothesis which is the two-stage model and the neural network model prediction yield indifference in terms of the accuracy or the mean absolute error. If the p-value is below 0.05, it means that there is a significant difference in the accuracy or the mean absolute error at 95% confidence.

In terms of the number of attributes, it is known that the more numbers of attributes are the more information is. The accuracies of both models are increasing. From Table XIV, for one attribute in the neural network model, the accuracy is at 50.3%. When the number of attributes increases to two and three attributes, the accuracy is improved to 51.4% and 52.9%, respectively. The number of attributes also impacts the accuracy of the two-stage model. The two-stage model's average accuracy increases from 50.6% to 53.2% when using more attributes.

In the stock prediction, the small value of the intraday returns can occur from noise trading. When there is news, the volatility is expected to increase; the size of the return will also be larger (Campbell and Hentschel (1992)). From this intuition, we filter out the small returns and predict the SET50 return only when the size of the returns has reached certain percentile ranks. Figure 10 plots the accuracy of the two-stage model and the neural network model when using one two and three attributes from 0 percentile filtering to 95 percentile filtering. From Figure 10 and Table XV, the average accuracy across all attribute combinations of one, two, and three attributes of the neural network model has increased from 51.1% to 52.1%, from 50.2% to 55.3% and from 49.9% to 62.2%, respectively, by filtering out more returns with small magnitude.

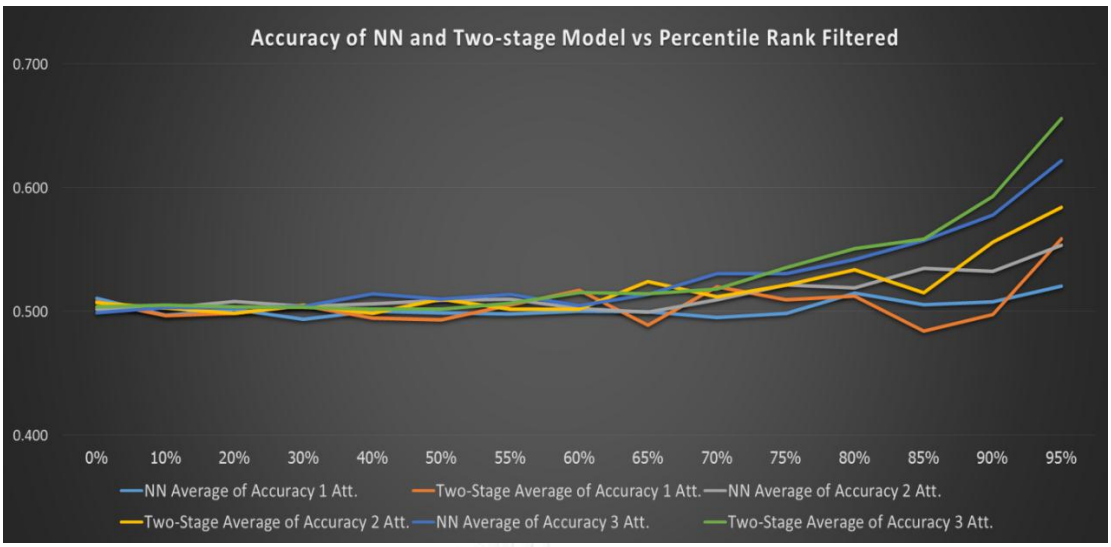


Figure 10 Accuracy of the NN and the two-stage model VS percentile rank example filtered

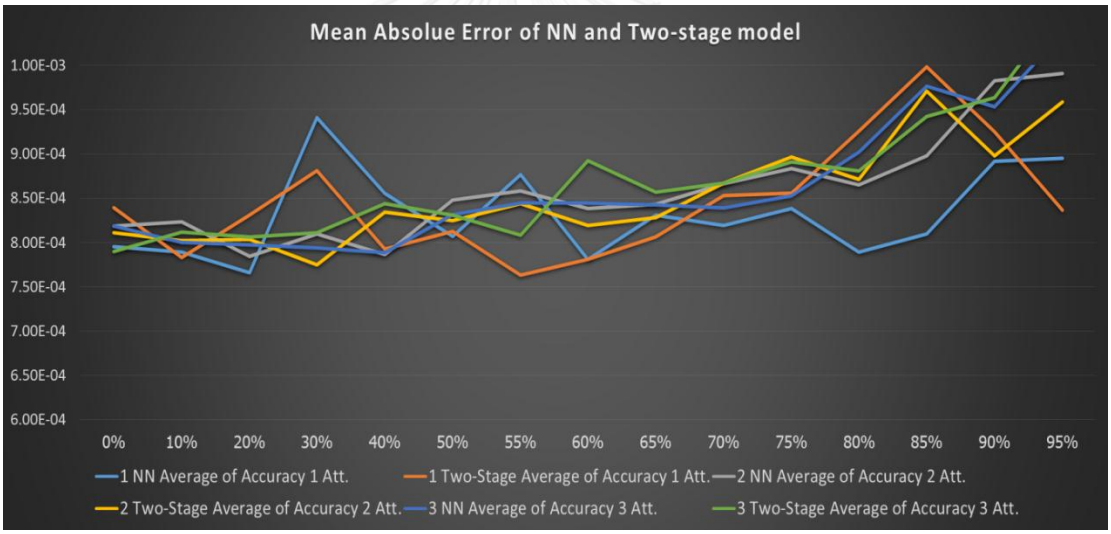


Figure 11 Mean absolute error of the NN and the two-stage model VS percentile rank example filtered

Table XIX reports the p-value that shows the significant difference of the neural network model's accuracy between 0 percentile rank and other percentile ranks from 10 to 95. Table XX reports the p-value that shows the significant difference of the two-stage model's accuracy between 0 percentile rank and other percentile ranks from 10 to 95. From the Wilcoxon signed-rank test in Table XIX and Table XX, the prediction accuracies for the filtered neural network and the filtered two-stage model are significantly increased compared with the non-filtered model's, especially in the case of filtering at the 70 percentile rank or higher. Thus, the return of the inputs at the high percentile rank has more predictive power than the return of the inputs at the low percentile rank in terms of accuracy. This result may come from the fact that returns with large magnitude tend to contain information or expectation about the future returns while the small returns are normally expected to be noises. When the size of the input returns is larger than a certain level, it is expected that it contains information that creates the five-minute momentum effect on the SET50 index return. Filtering out the low percentile rank returns could be beneficial for the stock prediction. But from Table XVII and Table XVIII, the Wilcoxon signed-rank test statistic shows that the MAE difference of the higher percentile ranks is larger than the MAE difference of the low percentile ranks. This effect may be explained by the high frequency trading returns. The high frequency returns are noisy and are normally around zero. When the low percentile rank returns are filtered out, it is hard to predict the value of the potentially large consequent returns compared with the value prediction around zero. Thus, the value predictions at the low percentile rank around zero tends to have less error.

Table XVII The Wilcoxon signed-rank test result on the mean absolute error difference of using 0 percentile data filtering and different percentile data filtering in the neural network model.

Test Statistics	Difference of MAE of NN model at each Percentile rank	P-value
	NNMAEPerc10 - NNMAEPerc0	.326
	NNMAEPerc20 - NNMAEPerc0	.009
	NNMAEPerc30 - NNMAEPerc0	.385
	NNMAEPerc40 - NNMAEPerc0	.028
	NNMAEPerc50 - NNMAEPerc0	.749
	NNMAEPerc55 - NNMAEPerc0	.155
	NNMAEPerc60 - NNMAEPerc0	.118
	NNMAEPerc65 - NNMAEPerc0	.618
	NNMAEPerc70 - NNMAEPerc0	.046
	NNMAEPerc75 - NNMAEPerc0	.036
	NNMAEPerc80 - NNMAEPerc0	.000
	NNMAEPerc85 - NNMAEPerc0	.000
	NNMAEPerc90 - NNMAEPerc0	.000
	NNMAEPerc95 - NNMAEPerc0	.000

*The test statistics show the p-value of the null hypothesis which is the neural network mean absolute error when filtering out at 10 up to 95 percentile is indifference to the neural network mean absolute error at 0 percentile. If the p-value is below 0.05, it means that there is a significant difference in the mean absolute error to 0 percentile filtering at 95% confidence.

Table XVIII The Wilcoxon signed-rank test result on the mean absolute error difference of using 0 percentile data filtering and different percentile data filtering in the two-stage model.

Test Statistics	Difference of MAE of the two-stage model at each Percentile rank	P-value
	TwoStMAEPerc10 - TwoStMAEPerc0	.803
	TwoStMAEPerc20 - TwoStMAEPerc0	.882
	TwoStMAEPerc30 - TwoStMAEPerc0	.726
	TwoStMAEPerc40 - TwoStMAEPerc0	.047
	TwoStMAEPerc50 - TwoStMAEPerc0	.459
	TwoStMAEPerc55 - TwoStMAEPerc0	.580
	TwoStMAEPerc60 - TwoStMAEPerc0	.001
	TwoStMAEPerc65 - TwoStMAEPerc0	.075
	TwoStMAEPerc70 - TwoStMAEPerc0	.000
	TwoStMAEPerc75 - TwoStMAEPerc0	.000
	TwoStMAEPerc80 - TwoStMAEPerc0	.000
	TwoStMAEPerc85 - TwoStMAEPerc0	.000
	TwoStMAEPerc90 - TwoStMAEPerc0	.000
	TwoStMAEPerc95 - TwoStMAEPerc0	.000

*The test statistics show the p-value of the null hypothesis which is the neural network mean absolute error when filtering out at 10 up to 95 percentile is indifference to the neural network mean absolute error at 0 percentile. If the p-value is below 0.05, it means that there is a significant difference in the mean absolute error to 0 percentile filtering at 95% confidence.

Table XIX The Wilcoxon signed-rank test result on the accuracy difference of using 0 percentile data filtering and different percentile data filtering in the neural network model.

Test Statistics	Difference of accuracy of NN model at each Percentile rank	P-value
	NNAccPerc10 - NNAccPerc0	.145
	NNAccPerc20 - NNAccPerc0	.020
	NNAccPerc30 - NNAccPerc0	.087
	NNAccPerc40 - NNAccPerc0	.000
	NNAccPerc50 - NNAccPerc0	.071
	NNAccPerc55 - NNAccPerc0	.018
	NNAccPerc60 - NNAccPerc0	.638
	NNAccPerc65 - NNAccPerc0	.233
	NNAccPerc70 - NNAccPerc0	.001
	NNAccPerc75 - NNAccPerc0	.000
	NNAccPerc80 - NNAccPerc0	.000
	NNAccPerc85 - NNAccPerc0	.000
	NNAccPerc90 - NNAccPerc0	.000
	NNAccPerc95 - NNAccPerc0	.000

*The test statistics show the p-value of the null hypothesis which is the neural network accuracy when filtering out at 10 up to 95 percentile is indifference to the neural network accuracy at 0 percentile. If the p-value is below 0.05, it means that there is a significant difference in the accuracy to 0 percentile filtering at 95% confidence.

Table XX The Wilcoxon signed-rank test result on the accuracy difference of using 0 percentile data filtering and different percentile data filtering in the two-stage model

Test Statistics	Difference of accuracy of the two-stage model at each Percentile rank	P-value
	TwoSAccPerc10 - TwoSAccPerc0	.885
	TwoSAccPerc20 - TwoSAccPerc0	.055
	TwoSAccPerc30 - TwoSAccPerc0	.503
	TwoSAccPerc40 - TwoSAccPerc0	.069
	TwoSAccPerc50 - TwoSAccPerc0	.879
	TwoSAccPerc55 - TwoSAccPerc0	.870
	TwoSAccPerc60 - TwoSAccPerc0	.270
	TwoSAccPerc65 - TwoSAccPerc0	.123
	TwoSAccPerc70 - TwoSAccPerc0	.058
	TwoSAccPerc75 - TwoSAccPerc0	.001
	TwoSAccPerc80 - TwoSAccPerc0	.000
	TwoSAccPerc85 - TwoSAccPerc0	.001
	TwoSAccPerc90 - TwoSAccPerc0	.000
	TwoSAccPerc95 - TwoSAccPerc0	.000

*The test statistics show the p-value of the null hypothesis which is the two-stage model mean absolute error when filtering out at 10 up to 95 percentile is indifference to the two-stage model mean absolute error at 0 percentile. If the p-value is below 0.05, it means that there is a significant difference in mean absolute error to 0 percentile filtering at 95% confidence.

The top three sets of accuracies and variables in terms of the accuracy are shown in Table XXI. The best performers are fifteen-minute returns of the HSI futures (HSIFMin15), five-minute returns of the Dow Jones futures (DJFMin5), and ten-minute returns of the Dow Jones futures (DJFMin10). These factors can best predict the five-minute SET50 index with 100% accuracy at the 95% rank return filtered. Also, it yields 100% precision and 100% recall in both "Up" and "Down". Moreover, the five-minute of the Dow Jones index future returns is the common factor in the top three best performers. Thus, five-minute of Dow Jones index future returns is able to contribute the most predictive power to the SET50 index return compared with the others in Table IV.

Table XXI Examples of variables used in the two-stage model with the three highest accuracies

Variable	Accuracy	Precision Up	Precision Down	Recall Up	Recall Down
HSIFMin15, DJFMin5, DJFMin10	100%	100%	100%	100%	100%
HSIFMin5, DJFMin5, DJFMin15	90%	100%	83%	80%	100%
HSIFMin5, DJFMin5, DJFMin30	86%	100%	67%	80%	100%

Moreover, the performance of the results of the five-minute SET50 return prediction is somewhat predictable. In some cases such as the model with three attributes filtered at 95 percentile rank, the accuracy of the prediction is as high as 65.6%. Thus, there is an impact from the return of the Hangseng index futures and the Dow Jones index futures on the SET50 index when the returns of these input index futures are high. The result also implies the inefficiency in the five-minute of the Thai market. This result agrees with Busse and Green (2002) which studied the US market efficiency in real time using Morning call and Midday call CNBC news and determined its real time effect on the US stock price.

CHAPTER V
CONCLUSION

In the first part of this thesis, it is inconclusive that adding the difference orders will help improving the accuracy for the daily stock return directional prediction using the support vector machine. Moreover, adding too many difference terms in the higher orders may result in lowering the accuracy, since generating and using useless information are more likely to create a noise for the predictive model.

Having considered the lag orders, more lagging terms tend to worsen the accuracy in the support vector machine directional prediction, since, in this case, the most recent lag term provides the most useful information and further lags seem to be unrelated to the daily prediction return. The difference orders somewhat improve the prediction accuracies, however, the accuracies improvement are not statistically significant. In conclusion, both difference orders and lags do not necessarily provide improvement to the predictive accuracy for the stock prediction.

Compared with other models, the directional support vector machine prediction using the higher order differences and lags proposed by this thesis yields the better result in terms of the accuracy. The predictive accuracy from this thesis is higher than other on other exchanges. Using the data from Thai stock exchange, this thesis confirms that the support vector machine significantly yields better accuracy than the neural network for the directional prediction.

In the second part of this thesis, the predictive accuracies are significantly higher when the higher percentile ranks of the absolute five-minute SET50 index return are filtered out. The results show that the high percentile ranks of absolute returns or the outliers of the returns significantly affect the predictive accuracy. The result implies that the outliers of the returns are informative and useful for the short-term stock prediction. It can be further applied to a momentum strategy and maybe other strategies as well.

The predictive accuracy results from the two-stage model are on average slightly higher than the predictive accuracy of the neural network model. Also, the improvement of the accuracy is significantly higher compared with the neural network model. The mean absolute error of the two-stage model compared with the neural network model is statistically indifferent. It can be concluded that the two-stage SVM-NN model can provide better predictive accuracy than the NN model, but provide indifferent performance in mean absolute error.

Also, the accuracy in both parts of the thesis are significantly more than 50% which implies that the market is somewhat predictable, thus the SET50 market is not efficient. There is a room for using momentum strategy along with the proposed factors to make a profit from the SET50 market and maybe from other markets.

This thesis covers the scope of the SET50 daily and the five-minute prediction, hence further study can be applied to other frequencies such as weekly, monthly and quarterly as well as other exchanges including the Hangseng index, the NYSE, the Nasdaq, etc.

REFERENCES

Barber, Brad M, and Terrance Odean, 1999, The courage of misguided convictions, *Financial Analysts Journal* 55, 41-55.

Barber, Brad M, and Terrance Odean, 2000, Trading is hazardous to your wealth: The common stock investment performance of individual investors, *The journal of Finance* 55, 773-806.

Bekaert, Geert, and Campbell R Harvey, 2000, Foreign speculators and emerging equity markets, *The Journal of Finance* 55, 565-613.

Benartzi, Shlomo, and Richard H Thaler, 2001, Naive diversification strategies in defined contribution saving plans, *American economic review* 79-98.

Bessembinder, Hendrik, and Kalok Chan, 1995, The profitability of technical trading rules in the asian stock markets, *Pacific-Basin Finance Journal* 3, 257-284.

Busse, Jeffrey A, and Clifton T Green, 2002, Market efficiency in real time, *Journal of Financial Economics* 65, 415-437.

Cakici, Nusret, Frank J Fabozzi, and Sinan Tan, 2013, Size, value, and momentum in emerging market stock returns, *Emerging Markets Review* 16, 46-65.

Campbell, John Y, and Ludger Hentschel, 1992, No news is good news: An asymmetric model of changing volatility in stock returns, *Journal of financial Economics* 31, 281-318.

Cao, Qing, Karyl B Leggio, and Marc J Schniederjans, 2005, A comparison between fama and french's model and artificial neural networks in predicting the chinese stock market, *Computers & Operations Research* 32, 2499-2512.

Chaigusin, Suchira, Chaiyaporn Chirathamjaree, and Judy Clayden, 2008, The use of neural networks in the prediction of the stock exchange of thailand (set) index, Computational Intelligence for Modelling Control & Automation, 2008 International Conference on (IEEE).

Doeksen, Brent, Ajith Abraham, Johnson Thomas, and Marcin Paprzycki, 2005, Real stock trading using soft computing models, Information Technology: Coding and Computing, 2005. ITCC 2005. International Conference on (IEEE).

Epps, Thomas W, 1979, Comovements in stock prices in the very short run, *Journal of the American Statistical Association* 74, 291-298.

Fernandez-Rodrıguez, Fernando, Christian Gonzalez-Martel, and Simon Sosvilla-Rivero, 2000, On the profitability of technical trading rules based on artificial neural networks:: Evidence from the madrid stock market, *Economics letters* 69, 89-94.

Hagan, Martin T, Howard B Demuth, Mark H Beale, and Orlando De Jesús, 1996. *Neural network design* (PWS publishing company Boston).

Halliday, R, 2004, Equity trend prediction with neural networks.

Hsieh, Tsung-Jung, Hsiao-Fen Hsiao, and Wei-Chang Yeh, 2011, Forecasting stock markets using wavelet transforms and recurrent neural networks: An integrated system based on artificial bee colony algorithm, *Applied soft computing* 11, 2510-2525.

Hsu, Sheng-Hsun, JJ Po-An Hsieh, Ting-Chih Chih, and Kuei-Chu Hsu, 2009, A two-stage architecture for stock price forecasting by integrating self-organizing map and support vector regression, *Expert Systems with Applications* 36, 7947-7951.

Huang, Wei, Yoshiteru Nakamori, and Shou-Yang Wang, 2005, Forecasting stock market movement direction with support vector machine, *Computers & Operations Research* 32, 2513-2522.

Inthachot, Montri, Veera Boonjing, and Sarun Intakosum, 2015, Predicting set50 index trend using artificial neural network and support vector machine, in *Current approaches in applied artificial intelligence* (Springer).

Khumyoo, C, 2000, The determinants of securities price in the stock exchange of thailand, *Unpublished master's thesis, Ramkhamhaeng University, Thailand*.

Kim, Kyoung-jae, 2003, Financial time series forecasting using support vector machines, *Neurocomputing* 55, 307-319.

Klimasauskas, Casimir C, 1993, Applying neural networks, *Neural networks in finance and investing* 47-72.

Kwon, Yung-Keun, and Byung-Ro Moon, 2007, A hybrid neurogenetic approach for stock forecasting, *Neural Networks, IEEE Transactions on* 18, 851-864.

Lendasse, Amaury, Eric de Bodt, Vincent Wertz, and Michel Verleysen, 2000, Non-linear financial time series forecasting-application to the bel 20 stock market index, *European Journal of Economic and Social Systems* 14, 81-91.

Lewellen, Jonathan, 2002, Momentum and autocorrelation in stock returns, *Review of Financial Studies* 15, 533-564.

Malkiel, Burton G, 2003, The efficient market hypothesis and its critics, *The Journal of Economic Perspectives* 17, 59-82.

Malkiel, Burton G, and Eugene F Fama, 1970, Efficient capital markets: A review of theory and empirical work, *The journal of Finance* 25, 383-417.

McCulloch, Warren S, and Walter Pitts, 1943, A logical calculus of the ideas immanent in nervous activity, *The bulletin of mathematical biophysics* 5, 115-133.

Pérez-Cruz, Fernando, Julio A Afonso-Rodriguez, and Javier Giner, 2003, Estimating garch models using support vector machines*, *Quantitative Finance* 3, 163-172.

Rouwenhorst, K Geert, 1999, Local return factors and turnover in emerging stock markets, *The Journal of Finance* 54, 1439-1464.

Schulmeister, Stephan, 2009, Profitability of technical stock trading: Has it moved from daily to intraday data?, *Review of Financial Economics* 18, 190-201.

Schwert, G William, 2003, Anomalies and market efficiency, *Handbook of the Economics of Finance* 1, 939-974.

Sirijunyapong, Warut, Adisom Leelasantitham, Supapom Kiattisin, and Waranyu Wongseree, 2014, Predict the stock exchange of thailand-set, Information and Communication Technology, Electronic and Electrical Engineering (JICTEE), 2014 4th Joint International Conference on (IEEE).

Smola, Alex J, and Bernhard Schölkopf, 2004, A tutorial on support vector regression, *Statistics and computing* 14, 199-222.

Tay, Francis EH, and Lijuan Cao, 2001, Application of support vector machines in financial time series forecasting, *Omega* 29, 309-317.

Tolvi, JUSSI, 2002, Outliers and predictability in monthly stock market index returns, *Liiketaloudellinen aikakauskirja* 369-380.

Trippi, Robert R, and Efraim Turban, 1992. *Neural networks in finance and investing: Using artificial intelligence to improve real world performance* (McGraw-Hill, Inc.).

Watts, Ross L, and Jerold L Zimmerman, 1978, Towards a positive theory of the determination of accounting standards, *Accounting review* 112-134.

Wilcoxon, Frank, 1945, Individual comparisons by ranking methods, *Biometrics bulletin* 1, 80-83.

Wu, F., 2011, Comparison between svm and bp in predicting stock index trends, (Shanghai Jiao Tong University).

Yao, Jingtao, Chew Lim Tan, and Hean-Lee Poh, 1999, Neural networks for technical analysis: A study on klci, *International journal of theoretical and applied finance* 2, 221-241.

APPENDIX

VITA

Phattradanai Samurwong was born in May 1st, 1990. In high school level, he graduated from Bunyawat Wittayalai School Lampang in 2007. Then in 2012, he earned Bachelor degree in electrical engineering from faculty of engineering, Chulalongkorn University. After that he enrolled in Master of Science in Financial Engineering program at Chulalongkorn University.