CHAPTER ONE

INTRODUCTION

1.1 Preface

The prediction of the future has always fascinated mankind due to the possible benefits of this knowledge (Thierry, 1996). This is especially true in the financial world. New tools and techniques for prediction are growing away from their original environment like the mathematical and computing world and find their way into all kinds of professional applications such as finance or engineering (B.K. Wong, 1995).

Modeling the markets using advanced financial engineering techniques has lately attracted a great deal of attention. Active managers have grown in number and many of them make the arguably reasonable assumptions that all people do not react similarly to publicly available data and that everybody does not react at the same speed (Ganesh *et al.*, 1995). These assumptions open up the possibility that one can beat the consensus by performing better or more efficient analysis, using advanced computer and mathematical tools as well as time-series modeling techniques (Farmer and Sidorowich, 1998; Weigend and Gershenfeld, 1994). Human information processing has limits; the new machine-aided approaches help expand those limits.

Various domain of interest have been explored in time series prediction. For example, the study of bioactivity prediction and compound classification using large collection of biological compounds (Rahayu, 2004), the study of electricity load forecasting based on electricity load demand data (Chen *et al.*, 2004), the study of

stock price forecasting using Haier closing prices data (Bao *et al.*,2004) and the study of rainfall forecasting using meteorological variables like potential temperature, vertical component of the wind, specific humidity, air temperature, precipitable water, relative vorticity and moisture divergence flux (Valverde *et al.*, 2005).

The success of SVM in prediction technique is evidence from several researches in electricity load forecasting (Chen *et al.*, 2004), stock price forecasting (Bao *et al.*, 2004), traffic speed prediction (Vanajakshi and Rilett, 2004), travel time series prediction (Wu *et al.*, 2004) and rainfall runoff modeling (Dibike *et al.*, 2001). Although SVM has been widely implemented in time series prediction, there is yet another area of interest which has not been explored by SVM which is in the KLSE stock price prediction.

Currently artificial neural network that utilize a back propagation algorithm has proved its superiority in predicting Wall Street Journal's Dow Jones Industrial Index (Darmadi, 1994) and modeling NASDAQ-GEM stock price relationship (NG *et al.*, 2000). Thus, the effectiveness of BPNN needs to be investigated in predicting KLSE stock prices.

Therefore, this project examines the feasibility of Support Vector Machine technique with selection of two different kernel functions that are RBF and polynomial in predicting future KLSE stock price in Malaysia. In addition to that, it is aimed at comparing and contrasting the performance of SVM with ANN in predicting KLSE stock price.

1.2 Problem Background

In recent years, many attempts have been made to predict the behavior of bonds, currencies, stocks, stock markets or other economic markets (Chakraborty *et al.*, 1992). These attempts were encouraged by various evidences that economic

markets do not behave randomly, but rather perform in a chaotic manner (Malliaris *et al.*, 1994).

Is stock price really predicable? In the earlier stage, under the assumption of efficient market investors believed that the movement of stock price presents a state of random walk. That means it is impossible to predict the change of stock price by its historical data. Nevertheless, some researchers who did empirical studies applying investment portfolio found historical information is actually useful in prediction (Osborne, 1964).

As the description above about the uncertainty of price movement, therefore, it is understandable that investment risk of stock is not low. In the traditional theory of investment portfolio, risk of stock can be divided into systematic risk and unsystematic risk (Sheng *et al.*, 1999).

Table 1.1: Stock market risk types

Risk Types	**Descriptions**
Systematic risk	• Causes reward change of the whole market on a single stock. • Originated by the changes of politics, society and whole economic environment (Sheng *et al.*, 1999). **Example:** Asia financial crisis from the end of 1997 had caused the stock markets in Southeastern Asia and Eastern Asia drastically dropped off. **Impact:** • Usually not easy to avoid through investment portfolio. • Called market risk or unavoidable risk.

Table 1.1: Stock market risk types (Cont'.).

Risk Types	**Descriptions**
Unsystematic risk	• Determined by the fluctuation of stock reward ratio which is influenced by circulation volume of stock, supply and demand of stock, and management performance of the enterprise. • Related to the business risk of the enterprise itself. **Example:** The unit price of Dynamic Random Access DRAM) lower than its cost in 1998, the profits of the related businesses, such as manufacturing and packaging of integrated circuits, were shrunk. **Impact:** • Stock prices fell down. • Called risk of the particular stock or idiosyncratic risk (Sheng *et al.*, 1999).

Typically, there are six traditional statistical models (Bao *et al.*, 2002):

1. Simple exponential,
2. Holt-Winters smoothing,
3. Regression method,
4. Causal regression,
5. Time series method, and
6. Box-Jenkins (Box *et al.*, 1994)

In addition to that, some of the models mentioned above are applicable for the issue where data distribution characterizes the periodic variation or normal spread

(Diebold, 1998), e.g. the seasonal or cyclical time series. An intelligent method, e.g. neural network (Castillo *et al.*, 2001), is also introduced to improve modeling capacity in recent years. This is because modeling a neural network is nothing to do with the problem of linear or nonlinear system, the appropriate order of the function, and the fitness test of model. Unfortunately, it faces the issue about generalization capability that decides the network performance (Castillo *et al.*, 2001).

Table 1.2: Limitations of traditional statistical models.

Statistical Models	**Problems**
Holt-Winters smoothing, Regression method and Box-Jenkins	• Require a lot of observed data for fitting their models to build better approach (Diebold, 1998). • Not suitable for short-term forecast because of only a few data available.

Modeling a forecasting system is widely discussed and studied for years, especially the topic about the trend analysis on time series or index series in which both of series definitely can be represented as a single in-order sequence (Bao *et al.*, 2002). However, the most of traditional statistics model cannot result in the satisfactory predicted results in many forecasting applications. This is because the traditionally mathematical model has to consider whether the system is the linear or nonlinear model, what the appropriate order of function for prediction is, and how to test the fitness of forecasting model (Box *et al.*, 1994). Therefore, alternative is to seek a kind of intelligent method as the prediction tool, i.e. SVM in which it can avoid the crucial problem mentioned in the traditional statistics model.

The domain of financial time series prediction is a highly complicated task due to following reasons:

1. Financial time series often behave nearly like a random-walk process, rendering the prediction impossible. The predictability of most common financial time series (stock prices, levels of indices) is a

controversial issue and has been questioned in scope of the efficient market hypothesis (EMH).

2. Financial time series are subject to regime shifting, i.e. statistical properties of the time series are different at different points in time (the process is time-varying).
3. Financial time series are usually very noisy, i.e. there is a large amount of random (unpredictable) day-to-day variations.
4. In the long run, a new prediction technique becomes a part of the process to be predicted, i.e. it influences the process to be predicted (Asim *et al.*, 2005).

In the past 30 years, financial world has embraced a decidedly quantitative orientation from many parts of the decision making processes with the widely adopted theories based on linear models (John, 1965). This culture can be attributed to a common belief about financial markets that linear models are both efficient and simple. But many reasoned researchers in the area have revealed that the dynamical systems comprising the financial markets require more complex models than have been tried previously (William, 1964). However, many relationships in finance are nonlinear and that no simple transformation can be made to make them linear over a large enough range to be interesting (Darmadi *et al.*, 1994).

Table 1.3: Previous research on stock fundamental data

Area	**Researchers**	**Features**	**Findings**
Stock selection by using support vector machines	Alan Fan and Marimuthu Palaniswami (2001)	• SVM used for classification • Fundamental information used. • Not focus on prediction but selection	• Produce 208% return over 5 strict out-of-sample year. • The prediction accuracy are relatively lower than other classification problem.

Table 1.3: Previous research on stock fundamental data (Cont'.).

Area	Researchers	Features	Findings
Efficient stock market forecasting using neural network	Amir Atiya, Noha Talaat and Samir Shaheen (1997)	• Used fundamental indicators determination. • NN used to selected Buy, Sell, Hold or Objective Strategy • Performance determined by annual profit.	• Good selection ability for neural network using the chosen indicators. • Stop and Objective Strategy give overall 40.4% annual profit.
Construct Decision Support System (DSS) to deal stock by using neural network	Norio baba and Hisashi Handa (1996)	• Design DSS to deal stock. • Input network based on fundamental and stock price. • Performances determined by buy and sell (decision maker).	• DSS behave wisely in decreasing trend in Tokyo stock market. • Average total gains achieve more than 10 million yen annually.

Stock market prediction has been a research topic for many years (Peters *et al*, 1991). Due to the fact that stock markets are affected by many highly interrelated economic, political and even psychological factors, it is very difficult to forecast the movement of stock market (Jing *et al*, 1997). Kuala Lumpur Stock Exchange (KLSE) has been chosen because KLSE is one of the largest markets in the emerging economies in terms of capitalization. So, the use of SVM as a time-series analysis in the KLSE stock market prediction is still a miss compare with the conventional neural network. Therefore, a detail comparison between neural network and SVM is a need especially in stock market prediction.

Fundamental data was not used in the recent research because the researchers are interesting in determining the ability of using SVM in predicting the future stock

prices based on past prices alone. The problem of finding fundamental data that matched the price data in the correct time sequence was another reason for not considering it. Reliable stable fundamental data are also difficult to obtain as government bodies that issue the statistical economic figures frequently revised them thus making the data practically unreliable in forecasting future stock prices that may rely on the data (Clarence, 1993).

This study is also motivated by a growing popularity of support vector machines (SVM) for regression problems (Kwok, 2001). SVM generalization performance does not depend on the dimensionality of the input space, but many SVM regression application studies are performed by 'expert' users having good understanding of SVM methodology.

Each of these SVM regression problems have being solved for their specific domain. Hence, the regression problem in KLSE stock prediction has not been done by other researchers. Due to the nature of SVM is based on statistical learning theory (Theodore *et al.*, 2000), SVM can be used to predict the KLSE market as well.

Recently, a support vector method for density support estimation was introduced by (Scholkopf *et al,* 2001), and has been successfully applied to a number of problems, including stock market predictions and selections. This method permits the control of the number of outliers in the training set and the solution of the optimization problem leads to a decision function which classifies new points as inliers and outliers. From the researches that have been done by other researchers on stock market predictions, most of them did not focus on the outliers. So, the effects of the outliers in determining the accurateness of KLSE stock prediction will be considered and some comparisons will be done for these.

While applying SVM to stock data prediction, the first thing that needs to be considered is what kernel function is to be used. As the dynamics of financial time series are strongly nonlinear (Maddala, 1999), it is intuitively believed that using nonlinear kernel functions could achieve better performance than the linear kernel

(Cao and Francis, 2003). In this investigation, the Gaussian Radial Basis function and the Polynomial function are used as the kernel function of SVM, because these kernels tend to give good performance under general smoothness assumptions. Consequently, they are especially useful if no additional knowledge of the data is available (Hall, 1998).

Support Vector Regression (SVR) is a recently introduced approach to regression problem (Smola *et al.*, 1998). It is a variation of Support Vector Machine (SVM), which was developed by Vapnik and his co-workers (Vapnik, 1998). Nowadays, SVR has been successfully applied to time series prediction (Mukherjee *et al.*, 1997) and financial forecasting (Tay and Cao, 2001).

In general, SVR uses the ε-sensitive loss function to measure the empirical risk and minimizes the regression error based on the Structural Risk Minimization (SRM) principle (Vapnik, 1995). Therefore, SVR need to be tested its efficiency in predicting the KLSE stock price due to financial data embedded noise.

Support Vector Machine as an emerging type of machine learning based on statistical learning theory, has proved its success in regression and time series forecast. This is clearly shown in a research done by Lu that using air pollutant data for prediction and as a result, Support Vector Machine has performed better as compared to RBF network (Lu *et al.*, 2002).

In their research, the result showed that SVM with Radial Basis Function kernel function produces smaller MAE values either for 24 hour or for one-week prediction in advance than that of RBF network. As we know, the weaknesses of RBF network are derived from its belongings to the family of neural network, which possessed typical problems of over fitting training and local minima and high influences of parameter selection on the model.

Another issue of SVM modeling is the parameter complexity. The problem with large models, for example, means more parameters, which means either that we

need more data to estimate the parameters, or we are less certain in our estimates (and thus in the overall usefulness of the model). Therefore, SVM is needed to test the relationships between model complexity and reliability by comparing different parameters.

Initially, SVM is a novel type of learning machine, based on statistical learning theory, which contains polynomial classifiers, neural networks and radial basis function (RBF) networks as special cases (Scholkopf, 1997). Thus, parameter selections become a challenging task especially in stock market forecasting in order to produce better prediction results.

Meanwhile, Vanajakshi and Rilett (2004) have compared two machine learning techniques performance that are ANN and SVM. They used these two machine learning techniques to predict traffic speed for intelligent transportation system (ITS). Based on the result, it is clearly shown that the proposed Support Vector Machine model using Support Vector Regression (SVR) with selection of RBF kernel is a viable alternative to ANN in short term prediction. It is because ANN performance depends largely on the amount of data available for training the network.

Therefore, if there is a situation where the available data are less and training data is not a good representation of the whole data, Vanajakshi and Rilett (2004) suggested SVR as another option for prediction problem. This is essentially same situation as the stock market where short term and long term prediction is important for those investors to gain more profit.

Chen *et al.* (2004) have also conducted an investigation on load forecasting by comparing three different techniques that are SVM with RBF kernel function, local models and neural network. The result of their experiments showed that SVM outperformed other techniques by producing an overall lower mean absolute percentage error (MAPE). While local model produced the largest MAPE due to its unsuitability to nonlinear type of data and neural network are known for its

difficulties in parameters selection that resulted in inconsistent MAPE. So, the efficiency of Support Vector Machine in stock market forecasting needed to be examined comparing with other machine learning techniques.

Table 1.4 : Kernels used by previous researchers

Previous studies	**Sigmoid kernel**	**RBF kernel**	**Polynomial kernel**	**Best Performance**
Chen *et al.*(2004) in load forecasting	No	yes	No	RBF
Lu *et al.* (2002) in air pollutant parameter forecasting	No	yes	No	RBF
Lucy Long Cheu (2003) in freeway incident detection	yes	yes	yes	Polynomial
R. Begg *et al.* (2003) for recognizing Young-Old Gait patterns	No	yes	yes	Polynomial
Frontzek *et al.*(2001) in predicting the nonlinear dynamics of biological neuron	No	yes	Yes	RBF

As mentioned earlier, kernel functions are main issue in SVM. Thus, a variety of kernel functions have been tested with SVM and amongst them, radial basis function (RBF) have shown remarkable results (Lu *et al.*, 2002; Chen *et al.*, 2004; Zhu *et al.*, 2002).

In prediction, SVM with choices of RBF outperformed other techniques such as Multi Layer Perceptron (MLP) and classical radial basis function network (RBFN) (Lu *et al.*, 2002; Zhu *et al.*, 2002). However, the rational of selecting kernel functions to be used and the criterions that would affect the performance of SVM in

prediction are not highlighted and justified in above literatures. Moreover, the performance of kernel functions is varying between different problems, parameters and scaling methods.

In the research done by Lin and Lin (2003), they used different data sets (heart, diabetes and others), and the sigmoid kernel function have produced an at par performance with RBF kernel with a proper selection of parameters. While, Frontzek *et al.* (2001) concluded that RBF and polynomial were able to learn the nonlinear dynamics of biological data but sigmoid failed in learning the problem. Therefore, the accuracy of SVM with RBF and polynomial kernel functions need to be tested in order to produce promising results in KLSE stock prediction.

Meanwhile, Ali and Smith (2003) and Parrado-Hernandez *et al.* (2003) also claimed that no specific kernel functions has the best generalization performance for all kind of problem domains and a priori information on which kernel function is the most appropriate to be used is ambiguous such that combining different type of kernels are suggested to solve a given problem in SVM. Moreover, there is no literature that compared the performance of different kernels functions in predicting KLSE stock price. Thus, the comparison on the performance of RBF and polynomial kernel functions as well as neural network in KLSE stock prediction need to be justified.

Besides, Rahayu (2004) in her research also have compared three different kernel functions, which are RBF, linear and polynomial in bioactivity prediction and compound classification. Based on result obtained, RBF kernel outperform other kernels due to its ability to handle non-linear relation between class labels and attributes and has hyper parameters that influence the complexity of model (Hsu *et al.*, 2003). Although RBF kernel function is popular among researchers, another high dimensional kernel function like polynomial has the potential of producing promising results (Frontzek *et al.*, 2001).

As a conclusion, this project is aimed at investigating the potential of applying two different kernel functions namely RBF and polynomial in Support

Vector Machine, of which only one can result in better stock price prediction in Kuala Lumpur Stock Exchange domain. Besides that, ANN is used to compare and evaluate the prediction performance of SVM and a detail comparison between ANN and SVM will be discussed on their strengths and weaknesses. Last but not least, this study is also concerned in evaluating the effectiveness of various data segments in stock data prediction against SVM and ANN models.

1.3 Problem Statement

In order to cater the problems stated in section 1.2, this project is carried out in order to answer the following questions:

1. How is the stock prices prediction performance while applying Support Vector Regression in stock market forecasting?

2. Which kernel functions (Polynomial or Radial Basis Function) give better results on stock data prediction?

3. Is SVM specifically tailored for one single stock or is that model general enough to predict more than one stock or even for KLSE index?

4. How accurate is the SVM prediction if comparing with ANN?

5. Does data segmentation and data transformation help to improve prediction performance of SVM and ANN?

1.4 Project Objectives

Various prediction techniques were studied in stock market prediction field and still nowadays researchers are focusing on implementing the latest technique in order to improve the stock market prediction model. Therefore, this project is carried out in order to fulfill the following objectives:

1. To predict the future stock prices value that can help trigger warning on potential buy or sell.
2. To determine which kernel functions namely Radial Basis Function (RBF) and Polynomial give better performance in predicting future Kuala Lumpur Stock Exchange stock prices value.
3. To make comparisons between two techniques which are Support Vector Machine and Neural Network that can improve the accuracy of KLSE stock market prediction.
4. To find out whether data segmentation and data transformation can improve stock prices prediction performance.

Realization of the fact that "Time is Money" in business activities, decision making plays an important role especially in stock exchange market. Therefore, it is a must for this study to investigate the accuracy of different learning techniques (SVM and NN) in predicting stock price. Then, the managers can make their valuable decision where time and money are directly related.

Since the main objective of this study is predicting future KLSE stock prices, the performance of SVM with different kernels (radial basis function and polynomial kernel) need to be determined. Beside that, different data segments will be used in this study to find out the superior learning techniques that produce better results in stock price prediction.

1.5 Project Scopes

The objectives of this study have been stated at the previous page. In order to achieve these objectives, it is important to identify the research areas, which cover the following aspects:

1. This project is focused on KLSE stock market domain by using End Of Day (EOD) data obtained from Kuala Lumpur Stock Exchange (KLSE) from year 1992 until 2006.
2. Support Vector Machine technique was implemented in predicting the stock prices. Radial basis function and polynomial kernel functions were applied and result from both were compared to find the suitable kernel function to be used in SVM for KLSE stock data prediction.
3. Only two machine learning techniques which are Support Vector Machine and Back Propagation Neural Network will be taken into consideration for stock prices prediction.
4. The predicted output obtained using Support Vector Machine was compared with the actual output and the performance was compared using Mean Square Error (MSE)
5. Performance benchmark on prediction was compared with Back Propagation Neural Network.

KLSE stock data are collected from year 1992 until 2006 because it is important to triggered the moving of the prices before and after the economy crisis at 1997. Besides that, the terrorists attacked that happened on 11th of September will also take into considerations to examine the events that will affect the performance of Support Vector Machine in stock market prediction.

1.6. Importance of The Study

This study is carried out with the main objective of evaluating the performance of Support Vector Machine in predicting KLSE stock prices by using different kernels. Therefore, some importance of this study is stated and based on the results obtained, it is hoped that this study is able to:

1. To encourage more works in exploring the advantages of Support vector Machine in term of different kernel selections for financial data prediction (not only stock market but also included currency rates, bonds, credits and others) in different domain, different segment of data and different time periods.
2. To give exposure on another promising technique of stock market prediction (Support Vector Machine) that could offer superior or at least same performance as the existing techniques (ANN or statistical approaches).
3. To provide basis for researchers who are interested in applying Support Vector Machine algorithm in fundamental data such as accounting information and company development or historical data (closing price, open prices, high prices, low prices and volume).
4. To encourage more studies on Support Vector Machine in term of different kernel selections (sigmoid kernel, linear kernel, RBF kernel or polynomial kernel) for stock prediction or stock selection.
5. To give an introduction on Kuala Lumpur Stock Exchange stock data characteristics and encourage more researches on Malaysia stock market.

1.7 Theoretical Framework

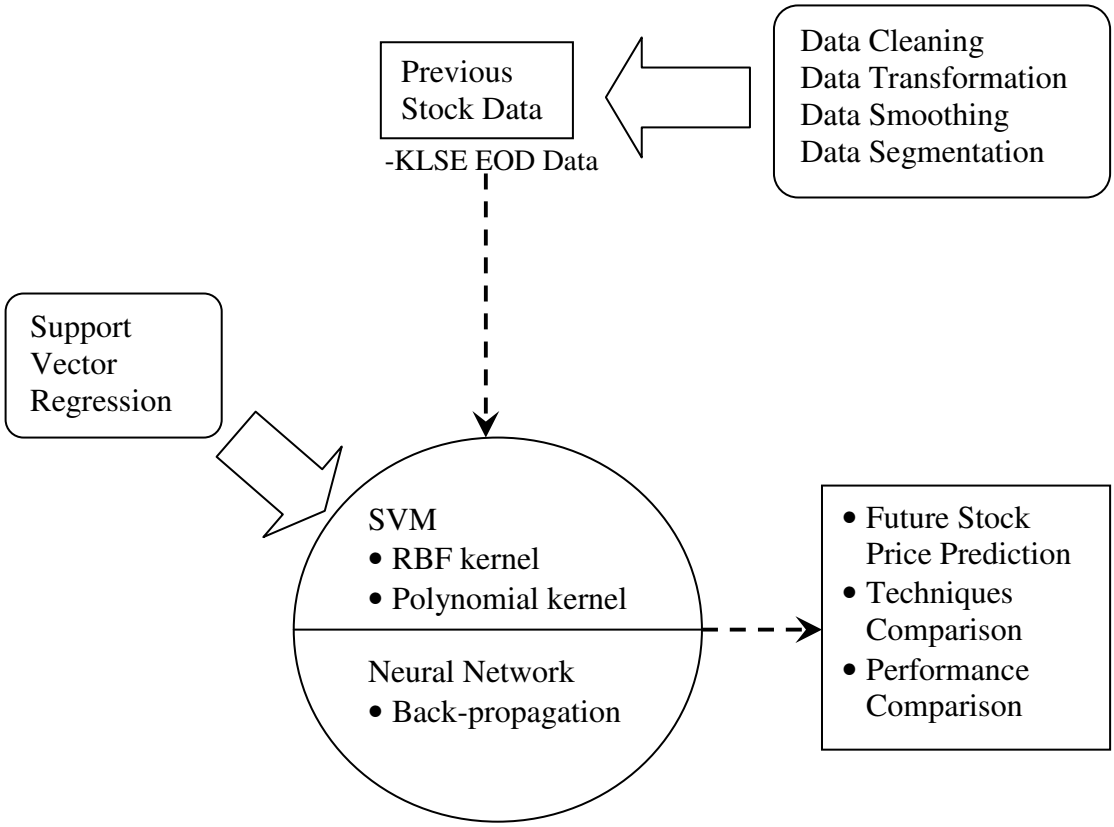

Figure 1.1 Theoretical framework

From the theoretical framework, it is clear that the input data consisted of a series of past End Of Day stock market data obtained from Kuala Lumpur Stock Exchange. These data will go through some pre-processing process such as data transformation and smoothing before evaluated by the chosen techniques. Then, SVM and Back-propagation network will be used to predict the future stock price by using different data segments and different time frame. Finally, the predicted output will compare with the actual stock prices for performance evaluation. In additions, some techniques comparison will be stated and the performance between SVM (Radial Basis Function and Polynomial kernel) and BP network will be discussed.

1.8 Report Organization

The organization of report comprised of five chapters. Chapter One explains an overall introduction on project's problem background, problem statement, objectives, scopes, importance of study, theoretical framework and definition of terms. In Chapter Two, a detailed review of past researchers studies will be discussed. It involves description on techniques in stock market prediction, KLSE Stock data, time series modeling and stock prediction, data mining operations and techniques and lastly support vector machine and artificial neural networks. After the analysis of past research, it is followed by Chapter Three, which presents the project framework and methodology. The project framework starts with the data collection, analysis on KLSE stock market domain, design of support vector machine structure that includes data prepossessing and parameter selection, experiments on the prepared data, analysis and comparison on experimental results.

Chapter Four explains the implementation parts of Support Vector Machine in the project, which consist of data preprocessing, selection of parameters, procedure in building an SVM model and experiments on prepared datasets. It is then followed by Chapter Five, which presents the analysis of the prediction results by comparing the performance of SVM and BP models on various data segments and data samples. Finally, Chapter Six concludes the overall findings, advantages, contribution and recommendations for future works based on the results obtained.