CHAPTER IV
STOCK MARKET PREDICTION
– A STUDY AND ANALYSIS

4.1 Introduction

Stock market prediction has been an important issue in the field of finance, engineering and mathematics due to its potential financial gain. As a vast amount of capital is traded through the stock market, the stock market is seen as a peak investment outlet. With the advent of faster computers and vast information over the Internet, stock markets have become more accessible to either strategic investors or the general public. As the Internet provides a primary source of event information which has a significant impact on stock markets, the techniques to extract and use information to support decision making have become a critical task [28]. To predict the stock market accurately various prediction algorithms and models have been proposed by many researchers from both in academics and in industry. This chapter discusses various prediction algorithms and models. The limitations encountered in the behavior of these techniques are identified and presented in this chapter.

4.2 Methods of Prediction

The prediction of stock market is an interesting task. In the literature there are number of methods applied to accomplish this task. These methods use various approaches, ranging from highly informal ways like studying a chart with the fluctuation of the market to more formal ways like linear or non-linear regressions.
These prediction techniques are broadly classified as depicted in figure 4.1. The criterion to this categorization is the type of tools and the type of data that each method is using in order to predict the market.

The common feature of all these techniques is that they are all used to predict the stock market aiming on enhancing the profit to be obtained from the market in future.

- Traditional Techniques
- Intelligent Techniques
  - Machine Learning Methods
    - Artificial Neural Networks
    - Genetic Algorithms
    - Support Vector Machines
  - Case Based Reasoning and Event Information
  - Hybrid Intelligent Techniques
- Combinational Techniques

Figure 4.1 Broader classifications of stock market prediction techniques

The methods and approaches adopted in all these categories of stock market prediction techniques and their limitations which have motivated this research work are discussed in the rest of this chapter.
4.3 Traditional Techniques

The Traditional Time Series Prediction analyzes the historic financial data and attempts to approximate future values of a time series as a linear combination of these historic data [101].

A number of researchers have used historical numeric time series data to predict stock markets and they achieved reasonable prediction accuracy while denying Efficient Market Hypothesis (EMH) [100]. However, there are various factors that influence stock prices such as company’s performance and robustness, trends of the market, investors’ psychology, government involvement, changes in economic activity and so forth. Thus, many researchers have agreed to the existence of significant correlation between the events which represent above factors and stock markets [28], [58], [66]. They have used the event information in addition to the numeric time series data and achieved some level of prediction accuracy.

The Econometrics is concerned with the tasks of developing and applying quantitative or statistical methods to the study and elucidation of economic principles [80]. In econometrics there are two basic types of time series forecasting: univariate (simple regression) and multivariate (multivariate regression) [75]. These types of regression models are the most common tools used in econometrics to predict time series. The way they are applied in practice is that first a set of factors that influence the series under prediction is formed or in more specific terms, a set of factors that assumed to influence the series under prediction is formed. These factors are the explanatory variables “X_j” of the prediction model.
Then a mapping between their values \( x_t \) and the values of the time series \( y_t \) is done, so that pairs \( \{ x_t, y_t \} \) are formed. The variable \( y \) is the "to be explained variable". These pairs are used to define the importance of each explanatory variable in the formulation of the "to be explained variable". In other words the linear combination of \( x_i \) that approximates in an optimum way \( y \) is defined.

- The Univariate models are based on one explanatory variable, that is \( i = 1 \).
- The multivariate models use more than one variable, that is \( i > 1 \).

These models used for predicting the stock market are limited to predicting numeric output and the lack of explanation about what has been learned is a problem.

### 4.3.1 Frequency Domain Analysis

Another approach, commonly used in scientific and engineering applications, is to analyze the financial time series in the frequency domain. Since frequencies and time periods are just inversely proportional quantities, either of them can be used in a transformed space. The spectral plot is the primary tool and basic technique used for the frequency analysis of time series. A spectral plot [45] is a graphical technique of examining the cyclic structure in the frequency domain. A cyclic structure is a smoothed Fourier transform of the auto-covariance function [5].

The spectral plot can be used to answer the following questions:

- How many cyclic components are there?
- Is there a dominant cyclic frequency?
- If there is a dominant cyclic frequency, what is it?
Trends should typically be removed from the time series before applying the spectral plot. Trends can be detected from a run sequence plot [12]. A run sequence plot is a graphical data analysis technique for preliminary scanning of the data. Trends are typically removed by differencing the series or by fitting a straight line (or some other polynomial curve) and applying the spectral analysis to the residuals [13]. Application of these techniques to predict the stock market did not result well always due to the limitations in trend removal.

4.3.2 Autoregressive (AR) Models

A common approach for modeling univariate time series of financial data is the autoregressive (AR) model illustrated in the equation 4.1:

\[ X_t = \delta + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + A_t, \quad (4.1) \]

where "X_t" is the time series, "A_t" is called as white noise, and

\[ \delta = (1 - \sum_{i=1}^{p} \phi_i) \mu, \quad (4.2) \]

where \( \mu \) is the process mean.

An autoregressive model is simply a linear regression of the current value of the series against one or more prior values of the series. The value of "p" is called as the order of the AR model.

AR models can be analyzed with one of various methods, including standard linear least squares techniques [91]. They also have a straightforward interpretation. AR
Models for a time series of financial data where the next point is dependent on the previous $n$ points: AR($n$). The evaluation of some traditional measures of sensitivity to change in price or yield of underlying security is not clear in this approach.

4.3.3 Moving Average (MA) Models

Another common approach for modeling univariate time series of financial data models is the moving average (MA) model shown in equation 4.3:

$$X_t = \mu + A_t - \theta_1 A_{t-1} - \theta_2 A_{t-2} - \ldots - \theta_q A_{t-q},$$  \hspace{1cm} (4.3)

where $X_t$ is the time series, $\mu$ is the mean of the series, $A_t$ are white noise, and $\theta_1, \ldots, \theta_q$ are the parameters of the model. The value of $q$ is called the order of the MA model.

That is, a moving average model is conceptually a linear regression [91] of the current value of the series against the white noise or random shocks of one or more prior values of the financial time series. The random shocks at each point are assumed to come from the same distribution, typically a normal distribution, with location at zero and constant scale. The distinction in this model is that these random shocks (or white noises) are propagated to future values of the time series. Fitting the MA estimates is more complicated than with AR models because the error terms are not observable. This means that iterative non-linear fitting procedures need to be used in place of linear least squares. MA models also have a less obvious interpretation than AR models.
Sometimes the Autocorrelation Plots and Partial Autocorrelation Plots suggest that a MA model would be a better model choice. Sometimes both AR and MA terms should be used in the same model.

Autocorrelation plots [6] are a commonly-used tool for checking randomness in a data set. This randomness is ascertained by computing autocorrelations for data values at varying time lags. If it is random, such autocorrelations should be near zero for any and all time-lag separations. One or more of the autocorrelations will be significantly non-zero, if it is non-random. Partial autocorrelation plots are a commonly used tool for model identification in Box-Jenkins models [6].

Although the Moving Average models can smooth the trading data, there are two main limitations of applying this method.

1. When forecasting number of items, the past observed values need to be available. Hence storage measures and safety measures are needed for the storage of huge historic data.

2. The MA methods consider each of the past time series value carrying the same weight. This disregards the possibility that the most recent data containing more information about the future forecast than do the older data.

4.3.4 Auto-Regressive Conditional Heteroskedasticity (ARCH)

Heteroskedasticity is the “serial correlation of volatility”. In econometrics, ARCH [20] is a model used for forecasting volatility which captures the conditional
“Heteroskedasticity” of financial returns. Today’s conditional variance is a weighted average of past squared unexpected returns. ARCH is an AR process for the variance. Although ARCH models usually apply to return series, financial decisions are rarely based solely on expected returns and volatilities.

4.3.5 Generalized Autoregressive Conditional Heteroskedasticity (GARCH)

GARCH [97] generalizes the ARCH model. Today’s conditional variance is a function of past squared unexpected returns and its own past values. The model is an infinite weighted average of all past squared forecast errors, with weights that are constrained to be geometrically declining.

GARCH models operate best under relatively stable market conditions [30]. GARCH is explicitly designed to model time-varying conditional variances, but it often fails to capture highly irregular phenomena. These include wild market fluctuations (for example, crashes and later rebounds) and other unanticipated events that can lead to significant structural change. GARCH models often fail to fully capture the fat tails observed in asset return series. Heteroscedasticity explains some, but not all, fat-tail behavior.

Although the ARCH / GARCH models are able to deal with the non-constant variance and non-linear relationship, still some series cannot be explained or predicted satisfactorily by these methods. This limitation has paved the path for researchers to move towards Machine Learning techniques.
4.3.6 Auto-Regressive Moving Average (ARMA) models

Box and Jenkins popularized an approach that combines the moving average and the autoregressive approaches [7]. These models are for a financial time series with no trend (the constant mean is taken as 0). They incorporate the terms in both an autoregressive (AR) model and a moving average (MA) model. In other words, the ARMA model is a combination of an autoregressive (AR) model and a moving average (MA) model. The autoregression and moving average (ARMA) models are used in time series analysis to describe stationary time series. These models represent time series that are generated by passing white noise through a recursive and through a non-recursive linear filter, consecutively.

A filter is an algorithm for processing a time series or random process. There are two major classes of problems solved by the application of filters to financial time series:

1. To estimate the current value of a time series \(X(t), t = 1, 2, \ldots\), which is not directly observable, from observed values of another time series \(Y(t), t=1,2,\ldots\), related to the time series \(X(t)\).

2. To predict the next value \(Y(t+1)\) of the observed time series \(Y\) from the current value \(Y(t)\) and previous values \(Y(t-1), Y(t-2), \ldots\).

The order of the ARMA model in discrete time \(t\) is described by two integers \((m, n)\) that are the orders of the AR and MA parts, respectively. The general expression for an ARMA-process \(y(t)\) is given in the equation 4.4:
\[ y(t) = \sum_{i=1}^{m} a(i) \cdot y(t-i) + \sum_{i=0}^{n} b(i) \cdot x(t-i) \]  

(4.4)

where

- "m" is the order of the AR-part of the ARMA model;
- \( a_1, a_2, a_3, \ldots, a_m \) are the coefficients of the AR-part of the model (of the recursive linear filter);
- "n" is the order of the MA-part of the ARMA model;
- \( b_1, b_2, b_3, \ldots, b_m \) are the coefficients of the MA-part of the model (of the non-recursive linear filter);
- \( x(t) \) are elements of the (input) white noise;

Having symmetric joint distributions, ARMA models are not ideally suited for stock market data exhibiting strong asymmetry. Also ARMA models are not suited for the stock market data exhibiting sudden bursts of very large amplitude at irregular time epochs [41].

The ARMA model is used to construct the ARIMA model of non-stationary time series discussed below.

4.3.7 Auto-Regressive Integrated Moving Average (ARIMA) models

These are the models for time series applied to financial data which resemble ARMA models except in that it is presumed the time series has a steady underlying trend.
These models, therefore work with the differences between the successive observed values, instead of the values themselves. To retrieve the original data from the differences requires a form of integration and the models are therefore called Autoregressive Integrated Moving Average models (ARIMA).

A general strategy for the analysis of time series is based on the use of ARIMA models and for seasonal financial data, SARIMA models. The first stage consists of removing trends or cycles from the data. An appropriate type of model must then be identified and its parameters are estimated. The estimated model is then compared with the original data and adjustments are made if necessary.

Although both autoregressive and moving average approaches were already known, the contribution of Box and Jenkins was in developing a systematic methodology for identifying and estimating models that could incorporate both approaches. This makes Box-Jenkins models a powerful class of models.

Both regression methods and ARIMA models for financial data fail to give an accurate forecast for some series because of their linearity inherent limitations. ARIMA model’s limitations include its requirement of a long time series. Often it is called a ‘Black Box’ model. Like the other methods discussed so far, this technique also does not guarantee perfect forecasts.

Though there are various statistical forecasting models for financial data presented above, many of them are very theoretical and hard for stock practitioners to employ for practical purposes. Professional stock practitioners always prefer to employ
models with easiness and simplicity to forecast cost-conscious time series. None of these techniques consider the wealth and health of the asset under study.

4.3.8 Fundamental Analysis

Fundamental analysis is the technique of applying the tenets of the firm foundation theory to the selection of individual stocks [10]. The analysts those who use this method of prediction use fundamental data in order to have a clear picture of the firm (industry or market) they will choose to invest on. These analyses are aiming to compute the "real" value of the asset that an investor will invest in and this value is determined by studying variables such as the growth, the dividend disbursement, the interest rates, the risk of investment, the sales level and the tax rates, and so on.

The objective of Fundamental Analysis is to calculate the intrinsic value of an asset or a stock. Intrinsic value is the actual value of a security, as opposed to its market price or book value. The intrinsic value includes other variables such as brand name, trademarks, and copyrights that are often difficult to calculate and sometimes not accurately reflected in the market price [10]. In other perspective, the market capitalization is the price (that is what investors are willing to pay for the company) and intrinsic value is the value (that is what the company is really worth).

If the intrinsic value of an asset is higher than the value it holds in the market, invest in it. If not, consider it as a bad investment and avoid it. The fundamental analysts believe that the market is defined 90 percent by logical factors and 10 percent by physiological factors. This type of analysis is not possible to fit in the objectives of our study. The reason for this is that the data it uses in order to determine the intrinsic value
of an asset does not change on daily basis. Therefore fundamental analysis is helpful for predicting the market only in a long-term basis. This technique is sometimes clubbed with Technical Analysis for better profit and returns.

4.3.9 Technical Analysis

Technical analysis [68] is the method of predicting the appropriate time to buy or sell a stock and it is used by most of the professionals with basic knowledge on stock pricing [1]. The idea behind technical analysis is that share prices move in trends dictated by the constantly changing attributes of investors in response to different forces. The technical data such as price, volume, highest and lowest prices per trading period are used for technical analysis. The technical analyst also uses charts to predict future stock movements. Price charts are used to detect trends; these trends are assumed to be based on supply and demand issues which often have cyclical or noticeable patterns.

From the study of these charts, trading rules are extracted and used in the market environment. The technical analysts are sometimes known as 'chartists'. Most chartists believe that the market is only 10 percent logical and 90 percent psychological. The chartist's belief is that a careful study of "what the other investors are doing?" will shed light on "what the crowd is likely to do in the future".

4.3.9.1 Technical Indicators

The Technical Indicators are any class of metrics whose value is derived from generic price activity in a stock or asset. Technical indicators look to predict the future price levels or simply the general price direction of a security by looking at past patterns.
Examples of common technical indicators include Relative Strength Index, Money Flow Index, Stochastic Momentum Indicator, Moving Average Convergence / Divergence (MACD) and Bollinger Bands. The list of Technical Indicators used in stock market prediction is furnished in Appendix A.

Antonia Azzini [2] demonstrated the possibilities of extracting many meaningful and reliable models of financial time series from a collection of popular technical indicators by means of evolutionary algorithms [92], [4]. The combined selection of various technical indicators to decide the stock market trend is discussed in [37], [8] which have resulted in higher percentage of profitable trade. The performance of selected combination of technical indicators may be further utilized for improving the predictability of various Intelligent Techniques.

4.4 Intelligent Techniques

Several methods for inductive learning have been developed under the common label "Machine Learning". All these methods use a set of samples to generate an approximation of the underlying function that generated the data. The following soft computing techniques applied for the prediction of stock market are generally classified as Intelligent Machine Learning Techniques [89].

- Artificial Neural Networks (ANN)
- Genetic Algorithms (GA)
- Support Vector Machines (SVM)
In the literature, it has been shown that these Intelligent Techniques offer the ability to predict market directions more accurately than other existing techniques [72]. The ability of Intelligent Techniques to discover non-linear relationships between the training Input / output pairs makes them ideal for modeling nonlinear dynamic systems such as stock markets [60].

4.4.1 Artificial Neural Networks

Artificial Neural Network (ANN) is a non-linear, data-driven approach that depends on the available data, without any a priori hypothesis about the kind of relationship, which makes it suitable for a complex data. However, neural network depends on network structure and complexity of samples, which cause over fit and low generalization [58]. This approach relies on building layers of nodes, each connected to a preceding layer through weights.

For the past several years, neural network has been going through an increment in popularity within stock market analysis. With the constant enlargement of the securities exchange market, it is unrealistic for the finance practitioners to discover patterns from thousands of hundreds of stocks in time without any professional aids. Scientists begin to design computer-based algorithms for technical pattern recognition [110]. The technical pattern recognition algorithm can mainly be classified into two categories, one is the rule-based algorithm and the other is template-based algorithm [61]. Nonetheless, either of these two categories has to design a specific rule or template for each pattern, which requires highly professional skills as well as involves considerable risks thus posing an enormous challenge for technical analysts.
Tsaih et al., [99] integrated the rule based technique and ANN to predict the direction of change of the “S&P 500” stock index futures on a daily basis. The “S&P 500” is a value weighted index of the prices of 500 large capital common stocks actively traded in the United States, published since 1957. The stocks included in the “S&P 500” are those of large publicly held companies that trade on either of the two largest American stock markets, the New York Stock Exchange and NASDAQ. Almost all of the stocks included in the index are among the 500 American stocks with the largest market capitalizations.

One of the advantages of ANN is the ability to learn relationship through the data itself rather than assuming the functional form of the relationship. As ANNs are known as “Universal Approximator”, any relation can be modeled to any degree of accuracy when sufficient data for the modeling are given. In addition, it provides a level of tolerance to noisy and incomplete data representation. Another advantage is that ANNs have non-linear, non-parametric adaptive learning properties and they have the most practical effect in modeling and forecasting. The non-linear nature of ANNs shows great potential to solve many complex problems.

The ANN has some limitations in learning the patterns when input data have high dimensionality. Dash and Liu [18] set the emphasis on the feature selection and suggested that reducing the number of input variables sometimes lead to improved model performance for a given data set. The reduction and transformation of the irrelevant or redundant features may shorten the running time and yield more generalized results [18].
The computational requirements of ANN models are heavily large for real time financial prediction. Modified Probabilistic Neural Network (MPNN) was introduced by Zaknich in 1991 [114] and further improved by Jan in 1999 [44].

Moreover, the ANN has the black box problem, which does not reveal the significance of each variable and the way they weigh independent variables [60]. As the individual role of each variable cannot be determined, it is impossible to understand how the network produces future price of stock. Although ANN offers relatively good learning ability than other techniques, they cannot always explain why they arrive at a particular solution.

Another major problem with ANN is the overtraining problem. When ANN fit the data too well, the system loses the ability to generalize. Since the generalization ability of ANN is fundamental to predict future stock prices, overtraining is a serious problem. The overtraining usually occurs by two main reasons as ANN either has too many nodes or have too long training time period (epochs). However, overtraining can be prevented by performing test and train procedures or cross validation.

As the characteristics of stock market data is of tremendous noise and non-stationary, Lawrence et al., [60] pointed out that the training of ANN is difficult for high noisy data. In this situation the networks fall into a “Naïve Solution” such as always predicting the most common output.

Furthermore, ANNs cannot always guarantee a completely certain solution. The ANN can neither guarantee arriving at the same solution repeatedly for the same training data given, nor guarantee the best solution [98].

89
4.4.2 Genetic Algorithm

The Genetic Algorithm (GA) is an evolutionary algorithm that mimics the natural selection in organisms, and is theoretically able to find optimal solution. Since GA is a stochastic algorithm, it is particularly useful in searching of solution in problem that is highly dynamic and chaotic such as stock market prediction [94].

The evolutionary algorithm GA is inspired by the genetic operations in chromosomes. It imitates the natural selection and adaptation of species in nature. Standard GA requires the definition of the parameters: selection operator, crossover operator, crossover probability, mutation operator, mutation probability, population size, number of generation, and fitness function. In addition, it requires the binary coding of input features that resembles the DNA coding [19].

The binary coding does not necessarily give significant improvement and owing to the fact that binary coding may disrupt the natural inter-feature relationships as well as introduces quantization error [94].

Over the past few years, Genetic Algorithms (GAs) as well as genetic programming (GP) have gradually become a major tool in Agent Based Computational Economics (ABCE). There are two styles of GA or GP in ABCE, namely, Single-Population GA / GP (SGA / SGP) and Multi Population GA / GP (MGA / MGP) [87]. SGA / SGP represent each agent as a chromosome or a tree, and the whole population of chromosomes and trees are treated as a society of market participants or game players.
The evolution of the stock market (society) can then be implemented by running canonical GA / GP, MGA / MGP, in contrast, represent each agent as a “society of minds”. Therefore, GA or GP is actually run inside each agent. Since, in most applications, direct conversations (imitations) among agents do not exist, this version of applications should not be mistaken as the applications of parallel and distributed GA / GP, where communications among “islands” do exist. At the current state, the SGA / SGP architecture is much more popular than the MGA / MGP architecture in ABCE [14].

In addition to its easy implementation, the reason for the dominance of SGA / SGP in ABCE is that economists would like to see their genetic operators (reproduction, crossover, and mutation) implemented within a framework of social learning, so that the population dynamics delivered by these genetic operators can be directly interpreted as market dynamics. In particular, some interesting processes, such as imitation, “following the herd”, rumors dissemination, can be more effectively encapsulated into the SGA / SGP architecture.

However, it has been recently questioned by many economists whether SGA / SGP can represent a sensible learning process at all. One of the main criticisms is given by Harrald [36], who pointed out the traditional distinction between the phenotype and genotype in biology, and doubted whether the adaptation can be directly operated on the genotype via the phenotype in social processes. If we assume (rare in practice) that agents only imitate others’ actions (phenotype) without adopting their strategies (genotype), then SGA / SGP may be immune from Harrald’s criticism.
In many situations, such as financial markets it would be hopeless to evolve any interesting agents if they are assumed to be able to learn only to "buy and hold" or "cooperate and defect" [3]. Harrald criticize by raising the issue "how can unobservable strategies be actually imitable"?

Although Harrald's criticism is well-acknowledged by researchers, it is seen no solution is perfectly proposed to tackle this issue yet [87]. Also, unfortunately, GA searches globally in the large input space, which can be time-consuming and may give suboptimal recognition performance [94].

4.4.3 Support Vector Machines

The Support Vector Machines (SVMs) were proposed by Vapnik [102] and his colleagues in the late 1970s. It has become a hot topic of intensive study due to its successful application in classification and regression tasks, especially in time series prediction and financial related applications [112].

The SVMs are a type of maximum margin classifiers. They seek to find a maximum margin hyper-plane to separate the classes, i.e., they maximize the distance of the hyper-plane from the nearest training examples. The hyper-plane thus obtained is called the Optimal Separating Hyper-plane (OSH) and the training examples that are closest to the maximum margin hyper-plane are called "Support Vectors".

If the data is linearly separable, a hyper-plane separating the binary decision classes in the two attribute case can be represented as the following equation 4.5:

\[ y = w_0 + w_1 x_1 + w_2 x_2, \]

(4.5)
where “y” is the outcome, “xi” are the attribute values, and there are three weights “wi”, which are to be used for training the learning algorithm.

The maximum margin hyper-plane can be represented as the following equation 4.6 in terms of the support vectors:

\[ y = b + \sum \alpha_i y_i x(i) \cdot x, \]  

(4.6)

where “y” is the class value of training example “x(i)”, the vector “x” represents a test example, the vectors “x(i)” are the support vectors and “\( \cdot \)” represents the dot product. In this equation, “b” and “\( \alpha_i \)” are parameters that determine the hyper-plane. Finding the support vectors and determining the parameters “b” and “\( \alpha_i \)” are equivalent to solving a linearly constrained quadratic programming problem.

If the data is not linearly separable, as in this case, SVM transforms the inputs into the high-dimensional feature space. This is done by using a kernel function as mentioned in the following equation 4.7:

\[ y = b + \sum \alpha_i y_i K(x(i), x) \]  

(4.7)

There are many different kernels for generating the inner products to construct machines with different types of nonlinear decision surfaces in the input space. Common examples of the kernel function are the polynomial kernel \( K(x; y) = (xy+1)^n \) and the Gaussian Radial Basis Function (RBF) \( K(x; y) = \exp(-(1/\delta^2)(x - y)^2) \), where \( n \) is the degree of the polynomial kernel and \( \delta^2 \) is the bandwidth of the Gaussian RBF kernel.
SVM is a very specific type of learning algorithms characterized by the capacity control of the decision function and the use of the kernel functions [101]. Established on the unique theory of the structural risk minimization principle to estimate a function by minimizing an upper bound of the generalization error, SVM is shown to be very resistant to the overtraining problem, eventually achieving a high generalization performance. A unique feature of SVM is that they are resistant to the over-fitting problem.

Recently, several applications of SVM to financial forecasting problems have been reported [56], [96], [95]. One of the well known studies using SVM in stock market prediction was performed by Kim [56]. He applied SVM to financial forecasting and compared it with the back propagation neural networks and Case Based Reasoning (CBR). The experimental results showed that SVM outperformed back propagation neural networks and CBR.

In the applications of SVM to predict the stock market, the degree of accuracy rate and the acceptability of certain prediction are measured by the predictors’ deviation from their own experiences or the ineffective data [43].

While predicting stock market, the most important matter is to improve the prediction accuracy rate [42], [15], which is yet to be achieved considerably in the case of SVM [88].

4.4.3.1 Case Based Reasoning

The Case Based Reasoning (CBR) is a reasoning technique that reuses past cases to find a solution to the new problem. Expert systems primarily capture the knowledge of
individual experts. Organizations have collective knowledge and expertise they have built up over the years. This knowledge can be captured and stored using CBR. CBR not only captures organization knowledge but also provides explanations for the derived solutions [72]. For this reason, CBR is popularly applied to many applications.

Kim [57] proposed a new hybrid model of GA and CBR for stock market prediction. From his preliminary studies, he found feature weighting or feature subset selection are very important to enhance the prediction performance of the CBR system. Thus, they used GA as a method of feature subset selection in the CBR system. He concluded that the hybrid model of GA and CBR offers a viable alternative approach to stock market prediction and the model has yet to mature.

4.4.3.2 Event information (EI)

In stock markets, there are many factors that can influence the share price. These factors can be derived from the news release about small companies or the news of superpower national economy. These incidents are called ‘events’, a qualitative factor in stock market prediction [70]. The primary reason of incorporating event knowledge in stock market prediction is based on an assumption that the future price of a stock partially depends on various political and international events as alongside the various economic indicators. Thus, many studies have used event information along with other quantitative data in predicting stock markets [58].

One of the popular studies using the prior knowledge and event-knowledge was performed by Kohara et al., [58]. They incorporated prior-knowledge in stock prediction such as newspaper information on domestic and foreign events. Event-knowledge is
extracted from the news paper headlines in accordance with certain prior-knowledge. Prior-knowledge is the information that stems from previous experience. Thus, based on the prior-knowledge, decisions can be made whether a particular event can positively influence the stock market tendencies or not.

Kohara et al., [58] selected several economic indicators such as, interest rate, price of crude oil, and New York Dow Jones average of the closing price and fed them together with event-knowledge into neural networks. Their experimental results showed incorporation of event knowledge improved the prediction ability of neural networks by reducing the error rate on the level of significance by 5%.

Meanwhile, how to incorporate the impact from news information into time series models is crucial. Maheu and McCurdy [63] specified a GARCH Jump model for return series, which can be directly measured from price data. The latent news process is postulated to have two separate components, normal and unusual news events. These news innovations are identified through their impact on return volatility.

But these methods do not provide an approach to figuring out the influential or noteworthy news of a given stocks in the face of thousands of news articles from all kinds of resources. Therefore, these methods cannot make significant improvement in practice.

In the literature, a number of researchers stated that stock prices are significantly correlated with the event information and many attempted to use both the event information and numeric time series data as input data. Fawcett and Provost [24] formulated the stock forecasting problem as an activity of monitoring the relationship between the news articles and stock prices. Fung et al., [28] proposed a system that
predicts the changes of stock trends by analyzing the influence of qualitative information (news articles). In particular, they investigated the immediate impact of news articles on the time series based on the EMH.

The evolution of the Internet with the global information infrastructure has led to an explosion in the amount of available information. Enormous event information which may have great influence on stock markets is available on the web. Whereas newspapers are updated once or twice a day, the real time news sources are frequently updated on the spot. Hong and Han [40] stated that as the popularity of the Internet increases, many newspapers expand their services by providing news information on the web in order to be more competitive and increase profit. News information includes articles on the political situation, social conditions, international events, government policies, trader’s psychology, and so forth, which we see and understand through the Internet. Such information is formulated in the form of texts, referred to as documents. Thus, text mining is required.

For example, on 13-04-2009, the news “Tech Mahindra acquires Satyam outbidding L&T”, leaving others behind, Tech Mahindra outbids other bidders and won the bid of shelling INR 2889 crore for 51% of the stake of beleaguered “Sathyam Computers” has caused its immediate increase of 3.16% per share against the previous day close price in the market [81]. It has also resulted in 12% increase of Tech Mahindra’s shares. These cannot be normally predicted by any Fundamental or Technical indicators. But qualitative analysis is not considering any price movements and intrinsic values of the company. Hence qualitative analytics alone are not capable of predicting consistently well.
Hong et al. [40] introduced an automated system that acquires event-knowledge from the Internet for the prediction of interest rates. These are the major motivations that lead to evolve the technique to extract the informative contents of web pages [52].

Most of the stock market prediction techniques or trading tools discussed so far provide just the very basic warning messages and none provide complete meaningful information about the tasks that need immediate action in critical situations [103].

4.5 Hybrid Intelligent Techniques

Recent researches tend to hybridize [85] the Artificial Intelligence Techniques [47] to forecast the stock market. Hiemstra [38] proposed fuzzy expert systems to predict stock market returns. He suggested that combining ANN and fuzzy logic (Neuro-Fuzzy Technique), capture the complexities of functional mapping and do not require the specification of the function to approximate.

Several hybrid Machine Learning Techniques have already been developed for market prediction like Neuro-Genetic approach [46]. Some are applied to predicting future price or rate of changes [104], and some are applied to recognizing certain price patterns that are characteristics of future price changes [48]. In both these models, however, only a little is considered about the training method of intelligent models. In case the numbers of training samples are uneven among categories, the intelligent models with "simple learning" try to improve only the prediction accuracy of the most dominant category which might be less important than others.
4.6 Combinational Techniques

The problem of "simple learning" or "simple training" which leads to train the network better for the most dominant category only, is addressed by performing the training process with the aid of a technical indicator or combination of well oriented multiple technical indicators.

For example: A Machine Learning model with simple learning has trained the "no-change" category of the training data set completely and is sufficient. It is because "no-change" category is a dominant category. However, as for training "selling signal" samples, the ratio of correctness is much lower than the actual number of "selling signal" in the training data set. Also, as for training "buying signal" samples, the ratio of correctness is much lower than the actual number of "buying signal" in the training data set. This means that the learning algorithms of the Machine Learning methods train mainly the samples in "no-change" category, which is most dominant of the three categories [39].

Therefore, learning methods that contribute to improve prediction accuracy of the other categories, which are more important than the dominating category, is the real need of the hour.

Chun-Teh Lee et al. [17] proposed a prediction system uses a back-propagation neural network and the KD and %R indicators for improving the efficiency of correct prediction. Hirotaka Mizuno et al., [39] proposed a technique which combines neural networks and multiple technical indicators for devising an expert system to predict the stock market. In this method, the numbers of learning samples are controlled by using information about the importance of each category. This is achieved by a normalized
learning method using many technical quantitative algorithms, one for each category of interest.

This prediction system classifies the input pattern that consists of several technical indicators, and generates a buying or selling timing signal for notifying users [59]. As shown in the Figure 4.2, the system consists of a neural network, a preprocessing unit, and a post-processing unit. The preprocessing unit normalizes the output of each technical indicator into an analog value in the range from 0 to 1. This is used to form an input pattern for training the neural network. Then the network recognizes the turning points of the market price curves from the input pattern. Finally, the post-processing unit converts the result of recognition into buying and selling timing signals.

Figure 4.2 Combinational Quantitative Technical Indicators used for Neural Network's learning process

The technical indicators presented below are the data items combined to form the input pattern used to train the neural network. This type of training is referred to as normalized training [59].
• **Moving average** - This is an average of the prices over certain past period and it is developed for allowing users to understand the trend without everyday fluctuation. There are several variations according to the period: 6, 10, 25, 75, 100, 150, and 200 days. The direction and position of moving average curves are used for predicting the price change.

• **Deviation of price from moving average** - This index is used for checking whether the price on each day is too high or too low in comparison with the expected price.

• **Psychological line** - This index is calculated by dividing the number of days of price ups by certain past period. This is used for predicting the price change from the rhythm of ups and downs.

• **Relative strength index** - This index is similar to the psychological line, and is calculated by dividing the sum of price ups by the sum of price ups and downs over certain past period.

Each of these indexes is normalized into 0 to 1 to form an input pattern to the neural network model. The neural network model with “normalized learning” demonstrated by Hirotaka Mizuno [39] has improved learning of “selling signal” and “buying signal” categories considerably.

The result of the trading rule close to –1 corresponds to the advice to sell, close to 1 corresponds to the advice to buy, and otherwise indicates no recommendation, i.e. the advice is interpreted as to do nothing.
In crucial situations most of the trading rules are expected to generate proper trading signals. But all the trading rules are not expected to generate the same trading signals. Some trading rules may be disoriented. Therefore, there is a need to build alert expert systems that combine a number of efficient trading rules which are properly oriented [76].

Piotr Lipinski et al., [76] suggested many ways of forming new trading rules by combining number of well-known trading rules. One among them suggested by these authors is forming a linear combination of known trading rules EMV and RSI as a function “f” mentioned below in the equation 4.8, which computes a result “f(K_t) ∈ {-1, 1}” on the basis of a knowledge “K_t” available at time t. The concept of knowledge is rather abstract, but in each specific case the knowledge can be defined as a set of historical data, such as daily or intra-day stock quotations.

\[
f(K_t) = 0.41 \text{EMV}(K_t) + 0.59 \text{RSI}(K_t),
\]  

(4.8)

where EMV is Ease of Movement indicator introduced by R.W. Arms and RSI is the Relative Strength Index / indicator introduced by W. Wilder [105].

Garth Garner [29] combined five technical indicators, namely, On Balance Volume (OBV), Price Momentum Oscillator (PMO), Relative Strength Index (RSI), Stochastic (%K) and Moving Average (MA) methods of time series technical analysis [31] to predict the following day’s closing price. To determine up and down days, the RSI uses the present day’s close compared to the previous day’s close [83], on the other hand, the RMI uses the close compared to the close n days ago. This technique has resulted in the generation of 50% of correct prediction. Hence it is attempted in this thesis.
to improve the performance of combinational technical indicators in terms of profitable buy/sell signal generation.

The most prominent limitations of presently available stock market prediction techniques towards the generation of profitable buy/sell signals and good returns are listed below.

- The univariate and multivariate models are limited to predicting numeric output and the lack of explanation about what has been learned is a problem.

- The frequency domain analysis does not work out for the financial time series data with no cyclic components or very few cyclic components.

- ARMA models are not ideally suited for data exhibiting strong asymmetry. Also ARMA models are not suited for the data exhibiting sudden bursts of very large amplitude at irregular time epochs.

- Many of the traditional stock market forecasting methods are very theoretical and hard for practitioners to employ for practical purposes.

- The machine learning models with simple learning attempts to improve only the prediction accuracy of the most dominant category which might be less important than other categories.

- Compounding all, the exact quantum of returns on investment was not derivable in most of the models surfaced so far is a major aspect of concern in prediction as far as the end user is concerned.
Hence a study in the perspective of designing and developing combinational technical quantitative algorithms to more accurately predict the financial time series is necessary. This area is focused in the chapter V and VI of this thesis, particularly on the returns on investment parameter. These combinational analytics when properly merged with any machine learning algorithm will improve the predictability further.