CHAPTER I
INTRODUCTION

1.1 INTRODUCTION
In a world of gizmos where everyone is running after technology and where data processing has been going through tremendous growth and changes, forecasting of stock prices is getting an ever-increasing importance. Be it the investors who are guided by the prediction of movement in stock prices or the companies who depend upon its accuracy, everyone secure their existence by giving a close look to what will happen in their future.

The stock market is comprised of a number of market participants (investors) with various risks and return characteristics with their own expectations about the stock of a particular company. They understand and react to the news of a company in different ways at different points of time. In such circumstances, it is very difficult to predict stock prices accurately. An investor can be guided to take buy and sell decision with the help of stock prediction system developed for predicting the trends or movement of dynamic stock prices accurately and consistently. Moreover, predicted trend in the stock prices will also help the regulators such as Securities Exchange Board of India (SEBI) of the stock market in taking many corrective measures. Keeping these facts in mind, this research is an attempt to forecast the future trends/values of given stock prices.

This chapter is divided into various sections where each section gives us a deeper look into the subparts of the thesis. The main theme of each section is as follows: Section1.1 gives the importance of the study; Section1.2 specifies the objectives of research; Section 1.3 briefs about developments in prediction of stock prices through traditional models, artificial neural networks and genetic algorithm. Section 1.4 communicates about the pre requisites terms and techniques used in this research. Lastly, section 1.5 concludes the chapter by outlining the present scope of our study.

1.2 OBJECTIVES OF THE STUDY
The prime objectives of the present study are as follows:

a) Forecasting of Stock data with statistical tools such as ARIMA models.

b) Forecasting of stock market data with data mining models such an ANN and GA.
c) To make a comparative analysis of statistical tools and data mining for their suitability for forecasting stock market data.

1.3 A BRIEF ABOUT TECHNICAL DEVELOPMENT IN PREDICTION OF STOCK PRICES

Research in this field possesses many theoretical and experimental challenges that give us an incentive to delve more and more into it. The most important aspect of these is the Efficient Market Hypothesis (EMH); referring Eugene Fama’s (1970) “Efficient capital markets”. This premise says that in any competent market, stock market prices are available to all citizens, whether he or she is an investor or not. Thus, probability of earning excessive profit does not exist. None of the systems is expected to outperform the market predictability, accurately and consistently. Hence, modelling any market data specifically stock prices under the assumption of EMH is only possible on the speculative, stochastic component and not on the movement of the changes in the value or other fundamental factors (He ping, 2004). Another theory related to EMH is the Random Walk theory. It states that random behaviour of the prices, i.e., previous prices along with future prices do not follow any trend or pattern.

A great deal of research has been conducted in the past about the validity of the EMH and random walk theories. However, with the introduction of computational and intelligent finance and behavioural finance, economists have tried to establish an alternative hypothesis which is called as the inefficient market hypothesis. The market is not always in a random walk, and inefficiencies always exist (Heping, 2003). The origins of disparity were studied by Mandelbrot (1960), analysing the cotton prices in New York stock exchange for developing assumptions of EMH. In his study, the results obtained proved that the data did not fit the normal distribution and resulted in producing symmetry from the viewpoint of scaling. Mandelbrot also analysed the fractals of the financial markets. Subsequently, with evolution in this field of research Pan Heping postulated the Swing Market Hypothesis (SMH) in 2003. The study analysed the efficient and inefficient nature of market and the subsequent trends also fluctuate amongst these two modes recurrently. The theory categorized the market movement into four components: physical cycles, dynamical swing, abrupt momentums and random walks.

Stock market is considered to be a chaos system. Chaos is a non-linear deterministic system which because of its irregular fluctuation is random. They are also dynamic systems, periodic, complicated and are difficult to deal with normal analytical methods. Neural
networks make fewer assumptions regarding dependencies and are proved to be very effective for learning such non-linear chaotic systems (Lawrence, 1997).

Many researchers and practitioners have proposed prediction models using various fundamental, technical and analytical techniques to give relatively an exact prediction. Fundamental analysis involves the in-depth analysis of the changes of the stock prices in terms of exogenous macroeconomic variables. It assumes the share price of a stock dependent on its inherent value and the expected return of the investors. But this expected return is subjected to change as new information pertaining to the stock is available in the market which in turn changes the share price. Furthermore, the analysis of the economic factors is also fairly subjective as the interpretation totally lays on the intellect of the analyst. On the other hand, technical analyses focus on using price, volume, and open interest statistical charts to predict future stock prices. The principle behind technical analysis is that all of the internal and external factors that affect a market at any given point of time are already factored into that market’s price. (Mendelssohn, 2000).

Further, for prediction of stock prices, some traditional time series forecasting tools are also used. In time series forecasting, the historical data is analysed to see the dependency between data. A model is developed based on the historical data and build patterns. Based on the theory, these models are then used to anticipate the future prices of the stock.

Predominantly, two approaches of time series modelling and forecasting are used, namely, linear approach and non-linear approach. Mostly, moving average, exponential smoothing, time series regression etc. are the methods used for forecasting. Autoregressive Integrated Moving Average (ARIMA) model is one of the most common and popular linear method which is also known as the [Box and Jenkins (1976)]. These model scan be represented in different types of time series, i.e., Autoregressive (AR), moving average (MA) and combined AR and MA (ARMA) series.

Many researchers and analysts have proposed the existence of the non-linearity in the financial market (Abhyankar et al, 1997). However, there exist very few evidences that substantiate the linear nature of stock market returns. It has been found that the residual variance is very high between the predicted return and the actual return of the stocks. Few parametric models such as Autoregressive Conditional Heteroscedasticity and General Autoregressive Conditional Heteroscedasticity are used for financial forecasting.

During the last decade, many researchers have developed various models and techniques for predicting stock market indices using artificial neural networks (ANN), anticipating that the market trends can be captured correctly. ANN has been used as a widespread technique
because of its ability to discover nonlinear relationship in the input data set without a prior assumption of the knowledge of relation between the input and the output (Hagen et al, 1996). Further, neural networks are analogous to non-parametric and non-linear regression model. It is proved that neural networks serve as better model in stock price prediction.

“A neural network is a massively parallel distributed processor that has a natural propensity for storing experiential knowledge and making it available for use. It resembles the brain in two respects: Knowledge is acquired by the network through a learning process and interneuron connection strengths known as synaptic weights are used to store the knowledge” (Haykin, 1994). Neural networks have the ability to interpret complex and inaccurate data. These models extract patterns and trends that are not noticeable by human or computer techniques. Neural networks are analogous to nonparametric and nonlinear regression model. Neural networks have the capability to understand the complex and chaotic nature of stock market whereas the conventional statistical models have limitations in complex relationship between the input and the output of the system. This weakness is taken care of by ANN models.

Neural Networks have a built in capability to adapt the network parameters to any change in the system under study. A neural network can be retrained and modelled to a new environment with a particular input data set to predict at the same level of environment. Further, neural networks can change its parameters, i.e., synaptic weights when the system is non- stationary and dynamic in nature.

1.4 PRELIMINARIES

In this section, the basic concepts and definitions to be used in this work are presented.

**Time Series:** “A time series is a collection of observations of well-defined data items obtained through repeated measurements over time” (Australian Bureau of Statistics-Statistical language of time series data).

**Autoregressive (AR):** Autoregressive model the series is regressed on to past values of itself. An Autoregressive (AR) Model of order p, or an AR (p) model, satisfies the equation:

\[ y_i = \phi_0 + \phi_1 y_{i-1} + \phi_2 y_{i-2} + \cdots + \phi_p y_{i-p} + \epsilon_i \]

**Moving Average (MA):** Moving average models are the time series models which can be expressed as a function of previous forecasting errors (or noise) ‘\( \epsilon_i \)’. A moving average (MA) model of order q, or an MA (q) model can be expressed with the following equation:

\[ y_i = \mu + \epsilon_i + \theta_1 \epsilon_{i-1} + \cdots + \theta_q \epsilon_{i-q} \]
**Autoregressive Integrated Moving Average (ARIMA) models:** In ARIMA Models future value of the variables is a linear combination of past values and past errors. Mathematically it can be represented as:

\[ Y_t = \Phi_0 + \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \ldots + \Phi_p Y_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \ldots - \theta_q \epsilon_{t-q} \]

where

Actual values of the data are taken as \( Y_t \), coefficients are considered as \( \Phi_i \) and \( \theta_j \). The Random error is denoted by \( \epsilon_t \) and auto regressive and moving averages are represented by integer's \( p \) and \( q \). (Ayodele et al., 2014).

**Degree of Freedom:** Degree of Freedom is known as the number of independent ways by which a dynamic system can move without violating any constraint imposed on it.

**Final Prediction Error (FPE):** It is a criterion to compare the quality of a model by simulating the situation where the model is tested on a different data set.

**Akaike information criterion (AIC):** It is a measure of the relative quality of statistical models for a given set of data. For a given collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Hence, AIC provides a means for model selection.

AICc is AIC with a correction for finite sample sizes. The formula for AICc depends upon the statistical model. Assuming that the model is univariate, linear, and has normally-distributed residuals (conditional upon regressors)

**Schwarz Bayesian information criterion (SBC)** is a criterion for model selection among a finite set of models; the model with the lowest BIC is preferred to be the best model.

**Artificial Neural Networks (ANN):** A neural network is a system composed of many simple processing elements operating in parallel whose function is determined by network structure, connection strengths, and the processing performed at computing elements or nodes. DARPA Neural Network Study (1988)

**Multilayer Perceptrons and Back Propagation Algorithm:** Multilayer Perceptrons (MLP) are usually layered Feed Forward networks. Static back propagation algorithm is always used in conjunction with MLP as the training of MLP is done with the help of back propagation algorithm. These networks are widely used in various applications requiring static pattern classification. Easy to use and approximation of input/output map is their distinct feature whereas the only disadvantages is that they are trained slowly and need a large training data set (Almeida, 1987).

**Radial Basis Function** is a type of non-linear hybrid networks which contains only one hidden layer of perceptron. The hidden layer in multilayer perceptron uses a standard
Sigmoidal functions whereas a radial basis function network uses Gaussian transfer functions. **RBF Neuron Activation Function** calculates the degree of similarity between the input vector and its prototype which is taken from the training sets. Input vectors which are more similar to the prototype give a result closer to 1.

**Hopfield Network** artificial neurons are the basic building block of Hopfield. The total number of inputs is represented by $N$ where $i$ is the input and a weight $w_i$ is associated with it. The output remains the same until the values of neuron is updated.

**Kohonen Algorithm**: It is an idea of Self Organizing Maps was given by Kohonen T (1990). This is a typical feed forward network with a single computational layer of neurons arranged in rows and columns.

**Epochs**: Epochs are full cycle of neural network training on the entire training set. This parameter defines the maximum learning epochs cycles to reach specifies minimum weights delta.

**Genetic Algorithm (GA)**: It is an algorithm which gives us the most optimal solution which is produced from fittest of a potential solution obtained from various potential solutions.

**Population**: It is a collection of potential solutions in genetic algorithms

**Selection**: It is an operator that decides which chromosome will reproduce.

**Generation**: Each cycle through the genetic algorithm is called a generation.

**Crossover**: It is an operator that perform recombination, creates two new offspring by combining the parent’s genes in new ways.

**Chromosome**: One of the candidate solutions to the problem is known as a chromosome.

**Mutation**: An operator which introduces new information to be genetic pool to protect against premature convergence is known as mutation.

**Optimization**: It is a feature in genetic algorithm which gives the best accurate solution after several iterations in the population. This feature is considered as the base-line of the genetic algorithm.

### 1.5 SUMMARY OF THE PRESENT SCOPE OF STUDY

All the chapters in this research are designed to give us better ways of predicting trends in stock market prices using artificial neural network. They are segregated on the basis of the various models used to explain different perspectives around which our study revolves. The succinct description of each chapter is mentioned underneath:
Chapter I: Introduction
This chapter draws around a brief introduction and the review of research work related to the forecasting of stock prices. It carries-forward the objectives defined for the study to identify the best forecasting methods of stock prices with the help of ARIMA models, artificial neural networks and genetic algorithms. It also presents the brief review about technical development in prediction of stock prices at different stages by different researchers. It narrates overview of the preliminary words used in this research. This chapter is summarised by giving a brief introduction of the chapter-wise research work to be carried forward to predict trends in the stock market using artificial neural network.

Chapter II: Review of Literature
This chapter presents the review of the research work carried out by various authors for using different models of forecasting, i.e., ARIMA models, artificial neural network models and genetic algorithms. A concise description of ARIMA model, ANN and Genetic Algorithm has been given to provide us a clearer view. The chapter also consists of the analysis table which presents analysis of research work done on all three models by a range of authors. A brief review of all the models used in the past has also been illustrated.

Chapter III: Descriptive Analysis of Time Series data
This chapter gives a brief introduction of the descriptive analysis followed by a range of definitions of the time series as given by earlier researchers. It also includes review of literature in relation to the application of time series analysis using ARIMA models, ANN and Genetic algorithms with specific reference to stock market data analysis. It showcases sector wise stocks of 25 BSE Sensex companies of the Indian stock market. The chapter also includes various statistics together with interpretation, mainly about measure of central tendency, dispersion, skewness and kurtosis, followed by tests of normality and test for outliers. One of the requirements of usage of times series data analysis with ARIMA models is that, time series must be stationary. The test of stationarity called ADF test is conducted to make the data stationary. The results of the ADF test are also given with comments of utility of descriptive statistics in context of ARIMA models.

Chapter IV: Analysis of Time Series data with ARIMA Model
This chapter dispenses the extensive process of building ARIMA models for short-term price prediction. The results obtained from real-life data establish the potential strength of ARIMA
models to provide stockholders a short-term prediction that could support investment decision making process.

It gives a brief impression of review of literature. It also describes the ARIMA model along with the identification of non-seasonal Box-Jenkins model. It further illustrates the methodology used and the process of experiments, its results obtained, and interpretation for non-normal time series stock, i.e., Sun Pharmaceuticals. Besides, it illustrates the same for the normal time series stock, i.e., Lupin Limited. Finally, the chapter predicts the stock prices of Sun Pharmaceuticals and Lupin Limited.

Chapter V: Analysis of Time Series Data using Artificial Neural Networks
This chapter begins with some definitions of artificial neural networks. It explains the application of neural networks and apprise about their characteristics. Subsequently, it presents the classification of artificial neural networks followed by neural network architecture. An overview of the neural network training is given which describes its methodology and data processing. The results obtained are discussed on the basis of descriptive analysis and then interpreted. Finally, the chapter concludes with the prediction of stock prices with the best ANN model.

Chapter VI: Time Series Analysis Using Genetic Algorithms
The chapter begins with the brief introduction of genetic algorithm followed by review of literature in relation to the application of time series analysis using genetic algorithms with specific reference to the stock market data analysis. It also specifies basic framework of a genetic algorithm succeeded by methodology and data processing. Further, descriptive analysis of experimental results obtained are shown and discussed. The chapter sum-up with the prediction of stock price based on the best selected method of genetic algorithm.

Chapter VII: Results and Discussions
This chapter comprises of the behaviour of stock market and the fundamental and technical analysis of stocks. It demonstrates the predicted prices of stocks forecasted on the basis of best models of ARIMA, ANN and Genetic Algorithm. In addition, it shows the descriptive analysis of predicted prices of stocks and their interpretation. Finally, the chapter concludes with inference of the study and some suggestions for future research work.