Sudan University of Science and Technology

College of Computer Science and Information Technology

Evolving Stock Market Prediction Models Using Soft Computing Techniques

تطوير نماذج للتنبؤ بسوق الاوراق الماليه عن طريق تقنيات الحوسبه المرنه

Submitted in partial fulfilment for awarding the Ph D degree in Computer Science

by

Sara Elsir Mohamed Ahmad Haimoura

M.Sc. Computer sciences , University of Khartoum, 2003

B.Sc. Computer sciences , University of Khartoum, 1995

Director:

Alaa F. Sheta, Professor

Electronic Research Institute

July, 2015

Dedication

To my Mom, Husband,Kids, Brother ,Sisters and Friends for their care, continuous support and love.

Acknowledgement

I wish to express my sincere thanks to my supervisor Dr. Prof .Alaa F. Sheta for valuable encouragement, and his friendly attitude and much valued unlimited help.

Thanks are extended to the generous staff of SUST .

I thanks my husband Imad for support, care and love.

I thanks my Mom and my brother ,sister for giving me endless encouragement and support.

My kids, have been rooting for me like nobody else, despite the intrusion that this work has been. I thank them for their tolerance, and their smiles.

I am grateful to many friends and colleagues who have listened to my questions, often provided answers and have been patient with my idiosyncrasies.

My special thanks are due to Dr. Sumai Elzain for crucial help.

Last but not least special acknowledgement and great appreciation are due to all those who helped me in one way or another.

Abstract

Evolving Stock Market Prediction Models Using Soft Computing Techniques

Sara Elsir Mohamed Ahmed

Department of computer Science

Sudan University of Science and Technology, 2015

Thesis Director:

Prof. Alaa Sheta

Stock market prediction is one of the hottest field of research lately due to its business applications owing to high stakes and the kinds of attractive benefits that it has to offer. The stock market is a dynamic, non-linear, complex, and chaotic in nature, forecasting stock market price is an important financial problem that is receiving increasing attention. During the last few years, a number of many models were presented to develop a relationship between the attributes which affect the stock index and its values. This thesis proposes a soft computing technique enhanced decision in financial management. The decision allows investors to maximize their expected return while practicing the prediction against financial risks. The importance of the research stems from the fact that it can be used to reduce the risk associated with uncertainty price movements in the stock market. The literature review shows that there are a large number of studies trying to forecast movements in the stock market, but there is a lack of literature trying to improve stock market risk management strategies with soft computing techniques. This thesis addresses this gap by applying the existing body of literature in stock index forecasting with soft computing techniques to the domain of forecasting index movements. In particular, it analyses whether there is an influence features of stocks used to predict movements of the stock index can improve forecasting the stock index movement off an investor faces use S&P 500 dataset and data related to it to create new dataset has impact in the price; The S&P 500, or the Standard & Poor's 500, is an American stock market index. The S&P 500 presented its first stock index in the year 1923. S&P 500 index found to have 27 influence features which affect the index values. A new market forecasting model based on soft computing and especially genetic programming is developed to enhance the investor decision. The model compare with traditional model Auto regression model and

Artificial Neural Network model. The system analysis stock market and futures data and makes a prediction about expected stock market conditions one day and next week. Selected new Features dataset by used GA to decrease the complexity. We were expecting that not all these features are significant in computing the index. Thus, we split our work to two phases. The first phase is to develop models based on these 27 features using Multiple Linear Regression (MLR) and ANN. Not only that, but we also explored a promising technique, multigene symbolic regression Genetic Programming to provide a mathematical nonlinear relationship between these attributes. GP found to be a powerful algorithm for providing mathematical modeling. In the second phase of this thesis, we adopted GAs as a mechanism to select the best features that contribute to the modeling process. The set of best features are once again used to build S & P 500 prediction models. Although the developed models in the second phase are with slightly less performance than in the first phase but the models are much simplex in complexity. This suggests that the stock market can be forecast using soft computing technique. Overall, this thesis concludes that the proposed model achieves a significant improvement in the prediction of stock market index.

المستخلص

التنبؤ فى سوق االاوراق الماليه من أهم مجالات البحثيه فى الآونة الأخيرة.وذلك لاهميته من الناحيه الاقتصاديه وذياده نسبه المخاطره فيه. سوق الأوراق المالية هي عملية ديناميكية، غير خطية، معقده، وذات طبيعه غير ثابته، والتنبؤ بسعرالاسهم هو من المشاكل التي تحظى باهتمام متزايد فى الاونه الاخيره للمستثمرين و التجاره. خلال السنوات القليلة الماضية، قد تم اقتراح عدد من نماذج الشبكات العصبية والنماذج الهجينة للحصول على نتائج دقيقة للتنبؤ، في محاولة ليتفوق على النمازج الخطية التقليدية.

طورت هذه الأطروحة تقنية الحوسبة المرنه لتعزيز القرار في النظم الإداريه. قرار يتيح للمستثمرين لزيادة عائداتهم المتوقعة أثناء ممارسة التنبؤ ضد المخاطر المالية. أهمية البحث تنبع من حقيقة أنه يمكن استخدامها للحد من المخاطر المرتبطة بتحركات الأسعار في سوق الأسهم. هذه الأطروحة تبحث من خلال الدراسات السابقه في طرق التنبؤ بسعر الأسهم باستخدام تقنيات الحوسبة المرنه. على وجه الخصوص، فإنها تبحث و تحلل ما إذا كانت هنالك عوامل لها تاثير فى سعر الأسهم يمكن استخدامها للتنبؤ بتحركات مؤشر البورصة لتحسين توقع حركة مؤشر الأسهم قبالة المستثمرينفى هذه الاطروحه استخدمنا مؤشرالسوق الامريكيه $S\&P500$ سعر الاغلاق مع اضافه عدد من الحقول الاخرى التي تم جمعها وتصنيفها الى سته فئات مختلفه. نتج عنهم ٢٧ حقل لهم تأثير مباشر فى سعر السهم. تم تطوير نموذج للتنبوء القائم على الحوسبة المرنة والبرمجة الجينية $(GeneticProgramming)$ باستخدام $multigenesymbolicregressionGeneticProgramming$. تم مقارنته مع نموذج الانحدار الخطى $(LinearRegression)$ ونموذج الشبكات العصبية الاصطناعية $(NeuralNetwork)$. ويستخدم نموذج لاختيار وحزف الحقول الاقل تاثيرا على التنبؤ باستخدام الخوارزمية الجينية $(GeneticAlgorithms)$ لدراسة الميزات التي يمكن أن تؤثر على التنبؤ. ونستخلص بان هذه الأطروحة اثبتت أن النماذج المطوره يمكن تطبيقها بنجاح عند التنبؤ في سوق الأسهم. جميع النتائج إيجابية وبدرجه عاليه من الثقه وهذا يشير إلى أنه يمكن التنبؤ بسوق الأسهم باستخدام تقنية الحوسبة المرنه وان النمازج المقترحه لها فائده عظمى فى التنبؤ ودرء المخاطر من ناحيه اتخاز القرارات الاداريه.

List of Publications

- Accepted: Evolving Stock Market Prediction Models Using Multigene Symbolic Regression Genetic Programming, Sara Elsir M. Ahmed , Alaa F. Sheta, Hossam Faris, AIML Journal , ISSN Print 1687-4846, ISSN Online 1687-4854, ISSN CD-ROM 1687-4862, ICGST LLC, Delaware, USA, 2015, , Impact Factor: 1.533 (2014)

  Predicting S&P 500 Stock Index: A Comparison between Multiple Linear Regression and Artificial Neural Networks Models Alaa F. Sheta, Sara E. Ahmed, Hossam Faris, International Journal of Artificial Intelligence (IJARAI) Volume 4 No 7 July 2015”.

- Under Preparation ”Improving Prediction Capability of Forecasting Model via Model Reduction”

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ANN	Artficial Neural Network
AR	Auto Regression
ARIMA	Auto Regressive Integrated Moving Average
ARMA	Auto Regression Moving Average
BPNN	Back Propagation Neural Network
CC	Correlation Coefficient
EC	Evolutionary Computational
EMH	Efficient Market Hypothesis
FF	Feed Forward Neural Network
FL	Fuzzy Logic
GA	Genetic Algorithms
GARCH	General Auto Regressive Conditional Heteroskedasticity
GP	Genetic Programming
LSE	Least Square Estimation
MAE	Mean Absolute Error
MGP	Multigene Symbolic Regression Genetic Programming
ML	Machine Learning
MLP	Multi Layer Perceptron
MLR	Multiple Linear Regression
MSE	Mean Squares Error
NN	Neural Network
PSO	Particle Swarm Optimization
RAE	Relative Absolute Error
RMSE	Root Mean Square Error
RRSE	Root Relative Squared Error
S& P500	Standard & Poor’s 500
SVM	Support Vector Machines
VAF	Variance Accounted For

Chapter One

Introduction

1.1 Overview

Today on the current state in society is that riches bring comfort and luxury. There is no doubt that the majority of the people related to stock markets is trying to achieve profit. Profit comes by investing in stocks that have a good future; this what they are trying to accomplish one way or the other is to predict the future of the market. But what determines the future? The way that people invest their money is the answer; and people invest money based on the information they hold. Investors try to forecast events that might affect stocks, such as sales expectations, and then make a decision whether the price of its stocks will increase or not. A business decision to loan or borrow money would depend on forecasts of future cash flows or expected returns. Economists in central banks are mainly interested in predict of future increase or decrease the trends, since these lead to monetary policy changes. Therefore, the development of accurate financial forecasting techniques is of extreme importance, especially in times of global economic confusion and market uncertainty. This is when financial time series are found to be most noisy and nonlinearities and structural breaks rule the common macroeconomic explanatory variables. The Stock Market prediction task divides researchers and academics into two groups those who believe that we can devise mechanisms to predict the market and those who believe that the market is well-organized and whenever new information comes up the market absorbs it by correcting itself, thus there is no space for prediction (EMH). Furthermore they believe that the Stock Market follows a Random Walk, which implies that the best prediction you can have about tomorrows value is todays value. In literature a number of different methods have been applied in order to predict stock market returns. These methods can be grouped in four major categories:

i) Technical Analysis Methods.

ii) Fundamental Analysis Methods.

iii) Traditional Time Series Forecasting.

iv) Machine Learning Methods.

Technical analysts, known as chartists, attempt to predict the market by tracing patterns that come from the study of charts which describe historic data of the market. Fundamental analysts study the intrinsic value of an stock and they invest on it if they estimate that its current value is lower that its intrinsic value. In Traditional Time Series forecasting an attempt to create linear prediction models to trace patterns in historic data takes place. These linear models are divided in two categories: the univariate and the multivariate regression models, depending on whether they use one of more variables to approximate the Stock Market time series. Finally a number of methods have been developed under the common label Machine Learning these methods use a set of samples and try to trace patterns in it (linear or nonlinear) in order to approximate the underlying function that generated the data. Such soft computing approaches have been extensively utilized in forecasting applications. Specifically, Neural Networks (NNs), Genetic Algorithms (GAs) , Fuzzy Logic (FL) and Support Vector Machines (SVMs) are very common in the financial forecasting literature.

The level of success of these methods varies from study to study and it is depended on the underlying datasets and the way that these methods are applied each time. However none of them has been proven to be the consistent prediction tool that the investor would like to have.

1.2 Motivation

There are two prices that are critical for any investor to know: the current price of the investment he owns, or plans to own, and its future selling price. Despite this, investors are constantly reviewing past pricing history and using it to influence their future investment decisions. Some investors won't buy a stock or index that has risen too sharply, because they assume that it's due for a correction, while other investors avoid a falling stocks, because they fear that it will continue to deteriorate.

The motivation of this research is to examine the factors influencing the price movements in the stock index futures markets using different soft computing techniques. It investigates whether soft computing can improve existing forecasting models. A new technique enhanced

decision will be proposed that allows investors to increase their expected return while practising qualifies against critical movements in the stock market.
Stock index forecasting is important for making informed investment decisions.

1.3 Problem Statement

In the stock market buying and selling is the most complex decision and stockholders often face some difficulties from the inability to:

- Determine and predict the stock market behaviour due to the dynamic and unpredictable environment of stock market domain. To take decision on the appropriate stock to buy or sell for better profit.
- Analyse and extract useful knowledge from a vast amount of information in order to make qualitative stock decision.

The problem is how to construct predictive model for stock market price to successfully forecast/predict index values or stock prices using the Soft Computing technique.

1.4 Research Objectives

The research have a general objective and a specific objective.

1.4.1 General objectives:

- Enhance stock market prediction model to be effective in improving the accuracy of stock market prediction.
- Enhance stock decisions of stockholders and keep the risk low.
- Explore the advantages of Soft computing techniques in solving stock market prediction problem.

- Evolve mathematical models which can be used for predicting stock market price with high confidence.

1.4.2 Specific Objectives:

- Study, survey and analyses the stock market index and the features that has an impact factor in stock index and create dataset with potential influence features that affect in prediction.
- Develop predictive model for stock market price using soft computing technique Artificial Neural Network and Multigene Genetic Programming.
- Compare this model with mathematical model based on Multiple Linear Regression model.
- Enhance these models by using feature selection (Genetic Algorithm).

1.5 Contributions

The contributions of this research to predict the stock market using soft computing are as follows:

- A new dataset based on S& P500, has been created.
- Developed and investigated Multiple Linear Regression (MLR) prediction model and used it as benchmark.
- Developed and investigated Artificial Neural Network predictive model.
- Developed and investigated Multigene Symbolic Regression Genetic Programming predictive models
- Selected new feature set by used Genetic Algorithm as feature selection.

1.6 Thesis Outline

This section describes the organization of the remaining chapters as follows:

- Chapter 2:
  The overview and survey of Stock Markets and Prediction: presents brief introduction to Stock Market and Theories of Stock Market classification and prediction.

- Chapter 3:
  Soft Computing Techniques: this chapter, presents an overview of methods that used in prediction stock market, It covers soft computing methods and survey on it.

- Chapter 4:
  Research methods: this chapter presents the methodology used in this research. A methodology is generally a guideline for solving a research problem. It contains the generic framework of the research and the steps required to carry out the research systematically.

- Chapter 5:
  Experimental Results: this chapter presents the results of the models that solve the thesis problem and success to achieved the goal and objective.

- Chapter 6:
  Conclusions : This chapter summarizes the objective and results and conclusions the research question.

Chapter Two

Literature Review and Survey

This chapter attempts to give a brief overview of some of the concepts of the stock markets and their prediction. Issues such as stocks theories, identification of available data related to the market, predictability of the market, prediction methodologies applied. All these issues are examined under the" daily and weekly basis prediction " point of view with the objective of incorporating in our study the most appropriate features.

2.1 Brief Introduction to Stock Market

Recently the Markets have become a more accessible investment tool, not only for strategic investors but for common people as well. Consequently they are not only related to macroeconomic parameters, but they influence everyday life in a more direct way. Therefore they constitute a mechanism which has important and direct social impacts. Here we are going to discuss some of the basics of stock market i.e. what is stock market, market index, stock exchange and many other concepts of the stock market. There are many different kinds of customers with different kinds of needs and preferences.

2.1.1 What is stock market?

A stock market is a public market for the trading of company stock and derivatives at an agreed price; these are securities listed on a stock exchange as well as those only traded privately. It is an organized set-up with a regulatory body and the members who trade in shares are registered with the stock market and regulatory body SEBI. The stock market is also called the secondary market as it involves trading between two investors. Stock market gets investors together to buy and sell shares in companies. Share market sets prices according to supply and demand. A stock that is highly in demand will increase in price, whereas as a stock that is being heavily sold will decrease in price. Companies that are permitted to be traded in this market place are called listed companies (Preethi and Santhi, 2012a).

2.1.2 Importance of stock market

The stock market is one of the most important sources for companies to raise money. This allows businesses to be publicly traded, or raise additional capital for expansion by selling shares of ownership of the company in a public market. History has shown that the price of shares and other assets is an important part of the dynamics of economic activity, and can influence or be an indicator of social mood. In fact, the stock market is often considered the primary indicator of a country's economic strength and development. Rising share prices tends to be associated with increased business investment and vice versa. Share prices also affect the wealth of households and their consumption. Therefore, central banks tend to keep an eye on the control and behavior of the stock market. Exchanges also act as the clearing house for each transaction, meaning that they collect and deliver the shares, and guarantee payment to the seller of a security. This eliminates the risk to an individual buyer or seller that the counterparty could default on the transaction. The smooth functioning of all these activities facilitates economic growth, lower costs; promote the production of goods and services as well as employment. In this way the financial system contributes to increased prosperity (Soni, 2011). Primary market deals with the new issues of securities. In the primary markets, securities are bought by way of the public issue directly from the company. An official prospectus is published under the Corporations Law and contains all the information that is reasonably required to allow you to make an informed investment decision about the company. Secondary market: it is where existing securities are bought and sold. Secondary market deals with outstanding securities. In the secondary market shares are traded among investors. This market is made of organized exchanges and may have a trading floor, where orders are transmitted for execution. This is where all the trading of stocks are maintained and guided by the rules set down by the exchange (Soni, 2011) .

2.1.3 Stock market basics

Stock market basics include shares and stocks. A share or stock is a document issued by a company, which entitles its holder to be one of the owners of the company.

- Share: it is directly issued by a company through IPO or can be purchased from the stock market. By owning a share one can earn a portion of the company's profit called dividend. Also, by buying and selling the shares gets capital gain. So, return is the dividend plus the capital gain. However there is a risk of making a capital loss, if selling price of the share is below than the buying price.
- Stock: it is nothing but a collection or a group of shares.The stock may be common stock or preferred stock.
- Common Stock : it represents the majority of stock. It represents ownership in a company and a claim on a portion of profits, or dividends. The dividend amount fluctuates and is not guaranteed. Shareholders are entitled to one vote per share to select board members, who oversee the major decisions made by the company's management. In the long run, common stock yields higher returns than most other investments.
- Preferred Stock: it represents a degree of ownership in a company but usually does not include voting rights. The stock holders of this type have the right to get a guaranteed fixed rate of dividend before the payment of dividend to the equity holders. They also have right to get back their capital before the equity holders in case of winding up of the company (Soni, 2011).

A stock exchange formerly a securities exchange is a corporation or mutual organization which provides ”trading” facilities for stock brokers and traders, to trade stocks and other securities, thus providing a marketplace (virtual or real). Stock exchanges also provide facilities for the issue and redemption of securities as well as other financial instruments and capital events including the payment of income and dividends.

The securities traded on a stock exchange include: shares issued by companies, unit trusts, derivatives, pooled investment products and bonds. To be able to trade a security on a certain stock exchange, it has to be listed there. Trade on an exchange is by members only. Definition done in the primary market and subsequent trading is done in the secondary market. A stock exchange is often the most important component of a stock market. There is usually no compulsion to issue stock via the stock exchange itself, nor must stock be subsequently traded on the exchange. Such trading is said to be off exchange or over-the-counter. This is the usual way that derivatives and bonds are traded. Increasingly, stock exchanges are part of a global market for securities. There are 20 major Stock Exchanges in the world (Setty et al., 2010).

Stock exchanges have multiple roles in the economy; this may includes raising capital for businesses, mobilizing savings for investment, Facilitating company growth, profit sharing, corporate governance, creating investment opportunities for small investors, government capital-raising for development projects, barometer of the economy. Listing requirements are the set of conditions imposed by a given stock exchange on companies that want to be listed on that exchange. Conditions sometimes include minimum number of shares outstanding, minimum market capitalization and minimum annual income. Companies have to meet the requirements of the exchange in order to have their stocks and shares listed and traded there, but requirements vary by stock exchange (Setty et al., 2010).

2.1.4 What is market index

An index is a statistical composite measure of the movement in the overall market or industry. Basically, indexes allow measuring the performance of a group of companies over a period of time. Companies are organized in an index according to two main methods or weighting as it is commonly termed. The movements of the prices in a market or section of a market are captured in price indexes called stock market indices, e.g., the S&P, the FTSE and the Euro next indices. Such indices are usually market capitalization weighted, with the weights reflecting the contribution of the stock to the index. The constituents of the index are reviewed frequently to include/exclude stocks in order to reflect the changing business environment. There are two major classes of indexes in use:

- Equally weighted price index: The index is calculated by taking the average of the prices of a set of companies.

  Equally weighted price Index = Sum (Prices of N companies) / divisor.

- Market capitalization weighted index: In this index, each of the N companies prices is weighted by the market capitalization of the company.

  Market capitalization weighted index = Sum (Company market capitalization * Price) over N companies/ Market capitalization for these N companies (Setty et al., 2010).

A few decades ago, worldwide, buyers and sellers were individual investors, such as wealthy businessmen with long family histories and emotional ties to particular corporations. Over time, markets have become more institutionalized buyers and sellers.

The market participants includes; investors, large institutions, issuers of securities, intermediaries.

- Stock broker : is person who is licensed to trade in shares. They also have direct access to the share market and can act as agent in share transactions. For this service, they charge a fee. Stock brokers can also offer additional services such as portfolio management or advice. The type of broker will depend on own confidence in trading shares. Often investors, who know exactly what they want to buy, will go to a discount broker to enact the trade.

  Stock broker may be full service broker or discount broker. Full-service broker will provide you with advice on which stocks to trade. They can often operate as financial planners and help with other aspects of your investment portfolio. Because they offer advice, a full service broker usually charges between 2 and 2.5 per cent fees, depending on the size of the transaction. Discount broker will execute trades, but will not provide any advice. As a result brokerage charges are low. Discount brokers generally operate via the telephone, Internet or both.

- Trader: In finance, a trader is someone who buys and sells financial instruments such as stocks, bonds and derivatives. Traders are either professionals working in a financial institution or a corporation, or individual investors. They buy and sell financial instruments traded in the stock markets, derivatives markets and commodity markets. Several categories and designations for diverse kinds of traders are found in finance, these include: stock trader, day trader, pattern day trader, swing trader, floor trader and rogue trader.

- Trading: Participants in the stock market range from small individual stock investors to large hedge fund traders, who can be based anywhere. Exchange is physical locations, where transactions are carried out on a trading floor, by a method known as open outcry. This type of action is used in stock exchanges and commodity exchanges where traders may enter verbal bids and offers simultaneously.

The other type of stock exchange (derivative exchanges) is a virtual kind, composed of a network of computers, where trades are made electronically via traders. Buying or selling at market means accepting any asked price or bid price for the stock, respectively. When the bid and ask prices match, a sale takes place on a first come first served basis (Soni, 2011).

2.1.5 What are the impact factors in stock market price?

In fact, there are many factors affecting the performance of the stock market and the investor to understand how these factors affect on the activity of the stock market, as an investor you should take advantage of understanding of these factors in a stock, which has in the gain selection, the other angle the investor must be able to determine the effects of the economic and political variables on the market in general, and that there not be an overstatement to identify these effects, as well as the case for News related to certain company must keep in mind that there is a phenomenon in securities known phenomenon overreact markets.

- Economic growth: The economic recovery is generally contributes to increased financial activity in terms of production increases depending on the growing demand for products, and returns this activity beneficial to all aspects of the economy where the rising level of employment and increasing incomes and rising revenues from corporate profits, and overall economic growth leads to an increase in the level of income community which raises the level of savings, which find their way around more investment, which drives the household sector employers to invest these funds Among the most famous investment is to invest in the stock market, thus increasing the demand for the purchase of shares will become more prices.

- Interest Rates: The stock market performance dramatically rising or falling interest rates affected, but this effect is inversely impact on the stock, as higher interest rates lead to attract a lot of savings into cash deposits which have to invest in the stock account recedes turnout rates the purchase price drops.

- Oil prices: This factor is one of the most important factors affecting stock prices, particularly in oil-producing countries, where high oil prices lead to an increase in the level of state revenue, which they can spend more money on infrastructure projects and raise the level of salaries, which contribute to raising the level of overall economic activity and provide liquidity also find their way into the stock market.

- Overreact: as the stock market is extremely sensitive to events in general, and in the short term the stock market may be affected by events that are not essential and has nothing to do with its key elements, which may be due to psychological factors among investors who have overcome their tendency optimism in certain cases the tendency to pessimism at other times, but the tendencies of optimism and pessimism and the impact of events is essential not be its impact on the long-term, but limited to the price movements in the short term time, so the investor does not deal with these price movements motivated emotional.

- Political events: the stock is a large market and political events experienced by the State affected, in the case of political instability seen the stock market is booming because the political climate stable of optimism and lure reflected to attract investors, especially foreigners, which leads to high liquidity in the market and so are therefore stock price rises applications, in contrast to the case of political tensions as we see on the Egyptian arena during the current period, which led to the deterioration of stock prices in the stock market because of investors' fears which pushed them to their money out of the market.

2.2 Data Related to the stock Market :

The information about the market comes from the study of relevant data. Here we are trying to describe and group into categories the data that are related to the stock markets. In the literature these data are divided in three major categories :

- Technical data: are all the data that are referred to stocks only. Technical data include:
  - The price at the end of the day.
  - The highest and the lowest price of a trading day.
  - The volume of shares traded per day.
- Fundamental data: are data related to the intrinsic value of a company or category of companies as well as data related to the general economy. Fundamental data include:
  - Inflation
  - Interest Rates
  - Trade Balance
  - Indexes of industries (e.g. heavy industry)
  - Prices of related commodities (e.g. oil, metals, currencies)
  - Net profit margin of a firm.
  - Prognoses of future profits of a firm
  - Etc.
- Derived data: this type of data can be produced by transforming and combining technical and/or fundamental data. Some commonly used examples are:
  Returns: One-step returns R(t) is defined as the relative increase in price since the previous point in a time series. Thus if y(t) is the value of a stock on day t,
  $R(t) = \frac{(y(t)-y(t-1))}{(y(t-1))}$
- Volatility: Describes the variability of a stock and is used as a way to measure the risk of an investment.

2.2.1 S&P 500 Stock index

The S&P 500, or the Standard & Poor 500, is an American stock market index computed according to the market capitalization of 500 large companies. These companies are having stock in the NYSE or NASDAQ. The S&P 500 index is computed by S&P Dow Jones Indices. The S&P 500 presented its first stock index in the year 1923. The S&P 500 index with its current form became active on March 4, 1957. The index can be estimated in real time. It is mainly used to measure the stock prices levels. There were a growing interest in the past few decades on measuring,analysing and predicting the behaviour of the S&P 500 stock index . John Bogle, Vanguards founder and former CEO, who started the first S&P index fund in 1975 stated that:

The rise in the S&P 500 is a virtual twin to the rise in the total U.S. stock market, so of course investors,and especially index fund investors, who received their fair share of those returns, feel wealthier.

In order to compute the price of the S&P 500 Index, we have to compute the sum of market capitalization of all the 500 stocks and divide it by a factor, which is defined as the Divisor (D). The formula to calculate the S&P 500 Index value is given as:
$IndexLevel = \frac{\sum(P_i * S_i)}{D}$

P is the price of each stock in the index and S is the number of shares publicly available for each stock.

2.3 Theories of Stock Market classification and prediction

Stock market has been studied over and over again to extract useful patterns and predict their movements. Stock market prediction has always had a certain appeal for researchers. While numerous scientific attempts have been made, no method has been discovered to accurately predict the price movement. There are various approaches in predicting the movement of stock market and a variety of prediction techniques has been used by stock market analysts. In the following section we briefly explain the two most important theories in stock market prediction. Based on these theories two conventional approaches to financial market prediction have emerged technical and fundamental analysis. The first one is efficient market hypothesis (EMH) introduced by fama in 1964 and the second one is random walk theory (Soni, 2011).

1. Efficient Market Hypothesis (EMH):

   An investment theory that states it is impossible to ”beat the market” because stock market efficiency causes existing share prices to always incorporate and reflect all relevant information. According to the EMH, stocks always trade at their fair value on stock exchanges, making it impossible for investors to either purchase undervalued stocks or sell stocks for inflated prices. As such, it should be impossible to outperform the overall market through expert stock selection or market timing, and that the only way an investor can possibly obtain higher returns is by purchasing riskier investments. The EMH states that no form of information can be used for generating extraordinary profits from the stock market, as stock prices always fully reflect all available information. Any new information which arises will be quickly and efficiently absorbed into the price of the stock. From the way that the EMH is defined, it is obvious that the result obtained in this work has a direct implication on the validity of the EMH. Fama contribution in efficient market hypothesis is significant. The EMH hypothesizes that the future stock price is completely unpredictable given the past trading history of the stock The efficient market hypothesis (EMH) states that the current market price reflects the assimilation of all the information available. This means that given the information, no prediction of future change in the price can be made. As new information enters the system the unbalanced stock is immediately discovered and quickly eliminated by the correct change in the price (Soni, 2011). The EMH exists in three forms, depending on the information which is used for making predictions weak EMH, semi-Strong EMH, strong EMH In weak EMH, any information acquired from examining the stock history is immediately reflected in the price of the stock .The weak form of EMH states that past stock prices cannot be used to predict future stock prices. Only past price and historical information is embedded in the current price. This kind of EMH rules out any form of prediction based on the price data only, since the prices follow a random walk in which successive change has zero correlation (Soni, 2011). The semi strong form goes a step further by incorporating all historical and currently public information into the price.

This includes additional trading information such as volume data and fundamental data such as profit prognosis and sales forecast (Setty et al., 2010). The strong form includes historical, public and private information such as insider information, in the share price. According to fama in his article efficient capital market states that the efficient market hypothesis surely must be false the strong form, due to the shortage in data, has been difficult to be tested. The strong form of EMH states that nothing can be used to predict future stock prices as all information is already reflected in the current price of the stock. By investigating the performance of mutual fund managers in (Quah, 2007), empirical evidence showed that the fund managers could not make use of any privileged information to achieve higher profits. The weak and semi-strong form of EMH has been fairly supported into a number of research studies (Soni, 2011).

2. Random walk theory:

   The random walk hypothesis claims that stock prices do not depend on past stock. Prices, so patterns cannot be exploited since trends to not exist. With the advent of more powerful computing infrastructure (hardware and software) trading companies now build very efficient algorithmic trading systems that can exploit the underlying pricing patterns when a huge amount of data-points are made available to them. Clearly with huge datasets available on hand, machine learning techniques can seriously challenge the EMH (Soni, 2011). It is a different perspective on prediction stock market prediction is believed to be impossible where prices are determined randomly and outperforming the market is infeasible. Random walk theory has similar theoretical to semi-strong EMH where all public information is assumed to be available to everyone. However, random walk theory declares that even with such information, future prediction is ineffective (Soni, 2011).From EMH and random walk theories, two distinct trading philosophies have been emerged. These two conventional approaches to financial market prediction are technical analysis and fundamental analysis (Setty et al., 2010).

3. Moving Average

   Moving average also called rolling average or rolling mean or running average is a type of finite impulse response filter used to analyse a set of data points by creating a series of averages of different subsets of the full data set in the stock market area. This is used to smooth out the short-term fluctuations with the help of time series analysis data and highlight longer-term stock market trends or cycles. This will use in technical analysis of financial data such as stock price, stock returns or trading volumes(Preethi and Santhi, 2012b).

2.4 Stock Market forecasting Techniques

Forecasting the stock market behaviour has always been in the foundation of scientific research by academics, investors, practitioners, market speculators and government institutions. This task has proven to be extremely challenging and hot due to the noisy and nonlinearity nature of financial time series. In order to measure the results of financial forecasts in practical market terms.

Financial time series forecasting was explored the past. They have shown many characteristics which made them hard to forecast due to the need for traditional statistical method to solve the parameter estimation problems.

In literature a number of different methods have been applied in order to predict Stock market index, we can classify these methods used to solve the stock market prediction problems to two folds:

- Traditional Techniques: These are statistical based approaches such as time series analysis Technical analysis, fundamental analysis, linear regression, Auto-regression and Auto-regression Moving Average (ARMA) (Harvey and Todd, 1983),(Wijaya et al., 2010). There are number of assumptions need to be considered while using these models such as linearity and stationary of the financial time-series data. Such non-realistic assumptions can affect the quality of predictions (Yu et al., 2009),(Walczak, 2001).

- Soft Computing Techniques: Soft computing is a term which covers artificial intelligence which mimics biological processes. These techniques includes Artificial Neural Networks (ANN) (Cubiles-de-la Vega et al., 2002), (Majhi et al., 2009) , Fuzzy logic (FL) (Hassan, 2009),Support Vector Machines (SVM) (Huang et al., 2005), particle swarm optimization (PSO) (Majhi et al., 2008) and many others.

2.5 Traditional Techniques

The literature categories the number of Traditional models in order to predict Stock market index into three major models:

- Fundamental Analysis Methods.
- Technical Analysis Methods.
- Traditional Time Series Forecasting.

2.5.1 Fundamental analysis

The fundamental analysis involves the in-depth analysis of a company's performance and the profitability to measures its fundamental value by studying the company physically in terms of its product sales, personal power quality, infrastructure, profitability on investment. It uses revenues, earnings, future growth, return on equity, profit margins, and other data to determine a company's underlying value and potential for future growth. (Chaigusin et al., 2008) To a fundamentalist, the market price of a stock tends to move towards its real value or intrinsic value . If this value of a stock is above the current market price, the investor can decide to purchase the stock because the stock price will bound to rise and move towards its intrinsic or real value . If this value of a stock is below the market price, the investor may decide to sell the stock because the stock price is bound to fall and come closer to its intrinsic value. To start finding out the intrinsic value, the fundamentalist analyser makes an examination of the current and future overall health of the economy as a whole. (Nguyen, 2003).

The fundamental analysis assumes that the investors are more logical and stock price (current and future) depends on its intrinsic value. As per fundamentalist, the market price of a stock tends to move towards its real value or intrinsic value . To find the intrinsic value of a particular stock the current and future overall health of the stock as well as the economy is required to be examined. The advantages of fundamental analysis are its systematic approach and its ability to predict changes before they show up on the charts. Fundamental analysis is a superior method for long-term stability and growth .But it is hard to time the market using fundamental analysis.(Agrawal et al., 2013) The origin of Fundamental analysis for the share price valuation can be dated back to Graham and Dodd (1934) in which the authors have argued the importance of the fundamental factors in share price valuation. Theoretically, the value of a company, hence its share price, is the sum of the present value of future cash flows discounted by the risk adjusted discount rate. This conceptual valuation frame work is the spirit of the renowned dividend discount model developed by Gordon (1962) (Chaigusin et al., 2008) . However, the dividend discount model valuation involves the forecast of future dividend payment which is difficult due to the changes in firms dividend policy. Thus, the subsequent studies along this line of literature searched for the cash flow that is unaffected by the dividend policy and can be obtained from the financial statements.

The author in (Ou and Penman, 1989) use financial statement analysis of income statement and balance sheet ratios to forecast future earnings. The primary motivation for this research is to identify mispriced securities. However, these authors demonstrate that the information in the earnings prediction signals is helpful in generating abnormal stock returns.

Fama and French in (Fama and French, 1995) show that value stocks (high book/market) significantly outperform growth stocks (low book/market). The average return of the highest book/market decile is reported to be one percent per month higher than the average return for the lowest book/market decile.

Author in (Jegadeesh and Titman, 2001)show that document that over a horizon of three to twelve months, past winners on an average continue to outperform past losers by about one percent per month.

In (Piotroski, 2000) the author examines whether a simple accounting based Fundamental Analysis strategy, when applied to a broad portfolio of high Book to Market firms, can shift the distribution of returns earned by an investor. The research shows that the mean returns earned by a high Book to Market investor can be increased by at least 7.5% annually through the selection of financially strong high Book to Market firms.

In (Nguyen, 2003) constructs a simple financial score designed to capture short term changes in firm operating efficiency, profitability and financial policy. The scores exhibit a strong correlation with market adjusted returns in the Current fiscal period and the same continues in the following period also.

2.5.2 Technical analysis

Technical analysis is a financial market technique that focuses on studying and forecasting the market action, namely the price, volume and open interest future trends, using charts as primary tools ? Charles Dow set the roots of technical analysis in late 18th century. The main principle of his Dow Theory is the trending nature of prices, as a result of all available information in the market. These trends are confirmed by volume and do persist despite the market noise, as long as there are not definitive signals to imply otherwise. Pring (2002) describes the following justification for the use of technical analysis:

The technical approach to investment is essentially a reflection of the idea that prices move in trends that are determined by the changing attitudes of investors toward a variety of economic, monetary, political, and psychological forces. The art of technical analysis, for it is an art, is to identify a trend reversal at a relatively early stage and ride on that trend until the weight of the evidence shows or proves that the trend has reversed. (Pring, 2002).

According to Yen and Hsu (Yen and Hsu, 2010), technical analysis is a popular trading strategy in the futures markets. Participants in the futures markets tend to rely on technical analysis rather than fundamental analysis. The authors state that in particular, measures like price movements and changes in trading volume play an important role when analysing futures prices.

The author in (Aldin et al., 2012) analyses the behaviour of investors in the stock index futures markets and find strong evidence of trend-chasing behaviour across the majority of the 32 examined markets. In particular, the authors state that a lot of traders seem to chase short-term to medium-term trends. The strongest indication of this trading behavior can be found during market downturns.

In (Neely and Weller, 2011) the author suggest that simple trend following systems do not work in the major currency markets any more. Instead, they discovered that trend following still works in exotic currency markets. However, even these exotic markets have become more efficient in recent years.

The author in (Menkhoff and Taylor, 2007) conclude that, while all four arguments may have some validity, argument number four is the most plausible. Technical analysis is used as a tool to inform about non-fundamental influences like market sentiment and psychological influences on price. The fact that technical analysis is frequently used by practitioners makes it an intrinsic part of the market. Therefore, the authors argue that researchers must understand

and integrate technical analysis into economic reasoning.

However, it is used by approximately 90% of the major stock traders. Despite its widespread use, technical analysis is criticized because it is highly subjective. Different persons can read charts in different manners.

Technical analysis evaluates the stocks by analysing statistics generated by market activity, past prices, and volume. It looks for peaks, bottoms, trends, patterns, and other factors affecting a stock's price movement. Future values of stock prices often depend on their past values and the past values of other correlated variables. Technical analysis looks for patterns and indicators on stock charts that will determine a stocks future performance. This analysis is largely preferred by the major stock traders and is good for shorter period also. Despite this fact technical analysis is criticized because it is highly subjective and different individuals can interpret charts in different manners. This analysis assumes that the market moves in trends dictated by the constantly changing attitudes of investors in response to different forces. Here it is assumed that the prices have tendency to go with the trend rather than against it and that the investors are 90% psychological, reacting to changes in the market environment in predictable ways. (Agrawal et al., 2013)

2.5.3 Traditional Time Series Forecasting

The Traditional Time Series Prediction analyses historic data and attempts to approximate future values of a time series as a linear combination of these historic data. In econometrics there are two basic types of time series forecasting: univariate (simple regression) and multivariate (multivariate regression). 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 firstly a set of factors that influence (or more specific is assumed that influence) the series under prediction is formed (Neelima Budhani, 2012) Models for time series data are ARMA, ARIMA, ARFIMA, and GARCH .

Statistical Regression Analysis

Statistical regression analysis associates relationships among a set of independent variables and one or more dependent variables. The independent variables could be historical measurements about certain events in the past while we want to estimate or predict an independent variable at this instant of time or even in the future. Many techniques for carrying out regression analysis were evolved in the past. Linear regression and ordinary least squares regression are parametric methods that use Least Square Estimation (LSE) to estimate mathematical model parameters.

Single Linear Regression Regression analysis measures the degree of influence of the independent variables on a dependent variable. In the case of simple bivariate regression where there is a single independent variable, the dependent variable could be predicted from the independent variable by the simple equation:

$$y = a + bx + \epsilon \tag{2.1}$$

y is the actual value x is the previous day value n is number of value.

Early work of using regression analysis in the futures market includes Schwager (1984), who later published a series of best selling books called Schwager on Futures. In his books, Schwager builds forecasting models using multiple regression and intermarket analysis.(Krollner, 2011)

In (TSE and CHAN, 2010) analyses the lead-lag relationship between the futures and spot markets of the S&P 500 using the threshold regression model. The authors find that the lead effect of the spot market is stronger in periods of direction less trading than in periods of strong trending markets.

The autoregressive integrated moving average (ARIMA) method is a forecasting technique which is often used as a benchmark for neural networks with purely technical inputs (Sermpinis et al., 2013). The ARIMA model consists of an autoregressive (AR), integrated (I) and moving average (MA) part. A model used for volatility forecasting is the General Auto Regressive Conditional Heteroskedasticity (GARCH) model where the variance rate follows a mean-reverting process. These models are usually employed in modelling financial time series that exhibit time-varying volatility clustering.

Author in (Fatima and Hussain, 2008) develop a model for the prediction of Karachi Stock Exchange index (KSE100) data. The authors use a combination of ANNs and ARIMA, as well as ARCH / GARCH models. The hybrid systems are compared to pure ARIMA and ARCH / GARCH models. The study concludes that the hybrid ANN model using ARCH/GARCH data is superior to the other analyzed models in predicting the KSE100 index.

In (Mohammadi and Su, 2010) evaluate the usefulness of several ARIMA-GARCH models in eleven international crude oil markets. While the authors report mixed forecasting results, the APARCH models outperfoms the other examined models.

In (Hossain and Nasser, 2011) the author compare a mixtures of ARMA-GARCH models to standard back propagation ANNs and support vector machines to predict changes in the Nikkei 225 and S&P 500 stock indices. The authors find that the support vector machines show the best performance in regards to the forecasting error, while the ARMA-GARCH model performs best in terms of predicting the correct direction of changes.

user conclude that support vector models are quite simple and have better interpret ability properties compared to complex GARCH type and ANN models and that support vector machines can still be used successfully in forecasting financial returns .

2.6 Soft Computing Techniques

Several methods for inductive learning have been developed under the common label ”soft computing”. All these methods use a set of samples to generate an approximation of the underling function that generated the data. The aim is to draw conclusions from these samples in such way that when unseen data are presented to a model it is possible to infer the to-be explained variable from these data. Machine learning approach is attractive for artificial intelligence since it is based on the principle of learning from training and experience. Connection’s models such as ANNs are well suited for machine learning where connection weights adjusted to improve the performance of a network. Difficulties and inaccurate results associated with these approaches (Neelima Budhani, 2012).

There are a lot of papers devoted stock market prediction with the use of Machine learning methods, neural networks, fuzzy logic, genetic algorithms, or a combination of it. Some examples of such papers are cited in order to illustrate the variety of possible approaches.

2.7 Stock feature Literature Review

The central idea to successful stock market prediction is achieving best results using minimum required input data and the least complex stock market model. The number of input variables used in each model differs. In general, the average number of input variables is between two and ten; however, there are cases where only one input variables are used. On the different, (Olson and Mossman 2003), (Zorin and Borisov 2002) use 59 and 61input variables, respectively.(Atsalakis and Valavanis, 2009a)

The most commonly used inputs are the stock index opening or closing price, as well as the daily highest and lowest values, supporting the statement that soft computing methods use quite simple input data to provide predictions.

Author	Method	Dataset	Feature
(Karazmodeh et al., 2013)	(PSO) Improved via Genetic Algorithm (IPSO) based on (SVM)	Nasdaq, Dow Jones, S&P 500	35 features
(Rezaiedolatabadi et al., 2013)	ANN and Imperialist Competitive Algorithm	Tehran Stock Exchange	share price, highest price, lowest price and trading volume
(Abhishek et al., 2012)	ANN	Microsoft Corporation from January 1, 2011 to December 31, 2011	open, high, low, volume and adj.
(Guresen et al., 2011)	NN model MLP and DAN2 and GARCH	NASDAG	Real exchange daily rate
(Aboueldahab and Fakhreldin, 2010)	Enhance PSO	Nasdaq 100 and S&P 500 indices	Daily open, maximum and closing values
(Olatunji et al., 2011)	ANN	Saudi Stock market historical data	closing price
Akintola et al. (2011)	Neural network in time series	Stock Prices of Intercontinental Bank Nigeria	weekly data using Average closing value
(Rahamneh et al., 2010)	Soft Computing Techniques	Amman stock exchange	the closing price, the highest price and the lowest price
(Ali et al., 2011)	ANN	Amman Stock Exchange	stock market prices by year

Table 2.1: Table of Literature Review Methods and Features

Author	Method	Dataset	Feature
(Allen and Karjalainen, 1999)	Genetic Algorithm	S&P 500 index	Average of closing prices
(Boyacioglu and Avci, 2010)	Adaptive Network-Based Fuzzy Inference System (ANFIS)	Istanbul Stock Exchange (ISE)	Republic gold selling price US Dollar exchange rates Interest rates on deposits Consumer price index Industrial production index Interest rates on Treasury bill Closing price of DJI Closing price of DAX Closing price of BOVESPA
(Madziuk and Jaruszewicz, 2011)	Neuro-genetic system	Tokyo Stock Exchange and New York Stock Exchange	opening, highest, lowest and closing values
(Choudhry and Garg, 2008)	hybrid GA-SVM	Indian Stock Market	35 technical indicator
(Liu and Yao, 2001)	Evolutionary Neural Network EPNet	Hang Seng stock index	Closing price
(Atsalakis and Valavanis, 2009b)	genetic programming technique (called Multi-Expression Programming)	Nasdaq-100 S&P CNX NIFTY	’opening value’, ’low value’, ’high value’ and ’closing value’.

Table 2.2: Table of Literature Review Methods and Features

Chapter Three

Soft Computing Techniques

The idea behind soft computing is to model the knowledge behavior of human mind. Soft computing is foundation of conceptual intelligence in machines; unlike hard computing, Soft computing is tolerant of imprecision, uncertainty, partial truth, and approximation. (Chaturvedi, 2008) Represents a significant paradigm shift in the aims of computing and well suited for real world problems.

Soft Computing is an approach for constructing systems which are:

- Computationally intelligent.
- Possess human like expertise in particular domain .
- Can adapt to the changing environment and can learn to do better.
- Can explain their decisions.

3.1 Artificial Neural Network

ANNs are mathematical models which were inspired from the understanding of some ideas and aspects of the biological systems, especially the human brain (Krose and van der Smagt, 2009) .

The artificial neural net attempts to follow the biological neuron, by modelling its response characteristics using an enhancing activation function f(s), similar to conduction of a signal through a resistance. The synapse of the neuron is imitated by a non-linear limiting function which performs amplitude limitation.

In Figure 4.2 x(i) is the ith input signal, and w(i) is the weights, associated with the input. The weight is the quantity that may get updated in a number of iterations in the learning

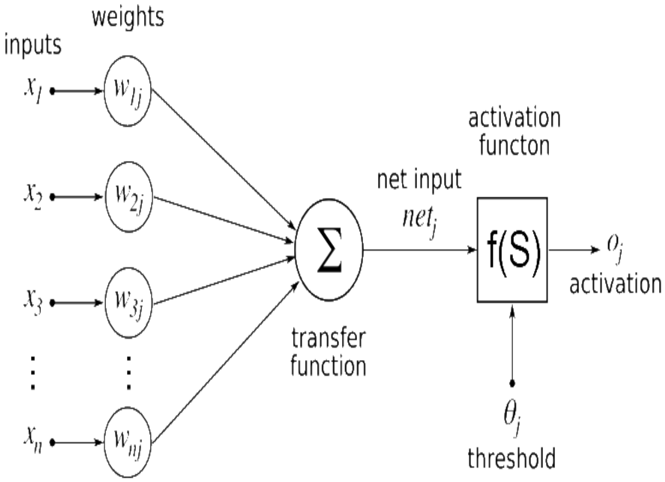

Figure 3.1: Artificial Neural Net durofy.com (durofy.com)

process, such that the neural net works more and more in the fashion of the required system model, producing more suitable outputs for a set of inputs. The net can be implemented using a mathematical Sigmoid function, where output is given as:

$$\phi(S) = \frac{1}{1 + e^{-s}} \tag{3.1}$$

Also, by other functions such as the tanh function, signum and step functions. A neural network may be considered as a data processing technique that maps, or relates, some type of input stream of information to an output stream of data. Variations of ANNs can be used to perform classification, pattern recognition and predictive tasks [(Olatunji et al., 2011), (Kara et al., 2011), (Chen et al., 2009), (Niaki and Hoseinzade, 2013)].

ANNs are considered a relatively new technology in Finance, but with high potential and an increasing number of applications. Neural network have become very important method for stock market prediction because of their ability to deal with uncertainty and insufficient data sets which change rapidly in very short period of time. In feed forward Multilayer Perceptron (MLP), which is one of the most common Neural Network systems, neurons are organized in layers. Each layer consists of a number of processing elements called neurons, each of which contains a summation function and activation function. The summation function is given by Equation 3.2 and the activation function can be a sigmoid function which is given in Equation 3.1.

$$S = \sum_{i=0}^{n} (w_i * x_i) \tag{3.2}$$

A standard neural network has at least three layers. The first layer is called the input layer (the number of its nodes corresponds to the number of explanatory variables). The last layer is called the output layer (the number of its nodes corresponds to the number of response variables). An intermediary layer of nodes, the hidden layer, separates the input from the output layer. Its number of nodes defines the amount of complexity the model is capable of fitting. In addition, the input and hidden layer contain an extra node called the bias node. This node has a fixed value of one and has the same function as the intercept in traditional regression models. Normally, each node of one layer has connections to all the other nodes of the next layer. Standard neural networks contain learning rules that modify the weights of the connections according to the input patterns [(Nigrin, 1993), (Jain et al., 1996)]. The most popular neural network learning algorithm for forecasting is back propagation neural network (Leonard and Kramer, 1990).
This layer summaries as the Flowing:

- Input layer:
  Contains source nodes that gathers information from the outside world and passes it on to the rest of the neural network .
- Hidden layer:
  Contains neurons (i.e. computational nodes) and is located between the input and output layers of the neural network .
- Output layer:
  In addition to having neurons, the output layer also provides the response of the neural network to the outside world .

3.1.1 Activation Functions

The activation function determines the output (i.e.the strength of the firing) of the neuron based on the net input and bias . The range of the output is limited to some interval by the activation function (Zimmermann et al., 2004). There exists several different kinds of activation functions, some of which are the threshold function and sigmoid functions etc. The sigmoid functions are the most common type of activation functions and they are monotonically increasing,continuous and differentiable (e.g. the logistic function and the hyperbolic tangent function) (Zimmermann et al., 2004).

Threshold Function

The threshold function is a binary valued function with the range [0,1] and neurons using this activation function are often referred to as McCulloch-Pitts model. There also exists threshold functions with other ranges (e.g. [-1;1]).

$$\varphi(u) = \begin{cases} 1 & : u \geq 0 \\ 0 & : u < 0 \end{cases}$$

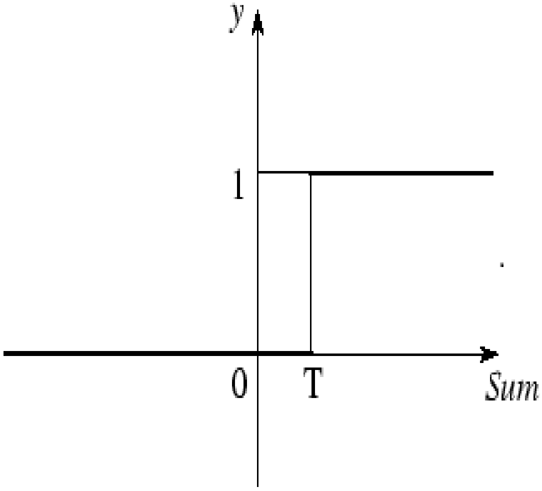

Figure 3.2: Threshold Function

Sigmoid Function

A sigmoid function with the range (0,1) in Equation 3.3. The parameter a affects the slope of the function and when a goes towards infinity,the sigmoid function becomes a threshold function (Zimmermann et al., 2004).

$$\varphi(u) = \frac{1}{1 + e^{-au}} \tag{3.3}$$

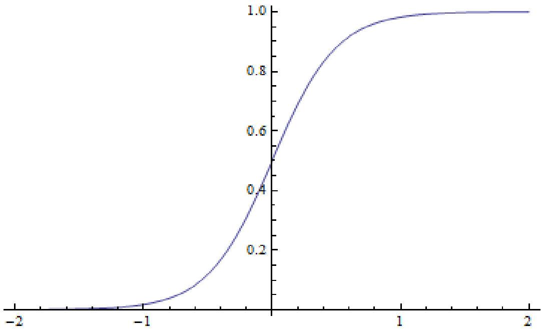

Figure 3.3: Sigmoid Function

Linear Function

The linear activation function multiplies the net input with a constant value k to produce the neuron output, which can be seen in Equation 3.4. Also note that the range of the linear activation function is $(-\infty; \infty)$.

$$\varphi(u) = k * u \tag{3.4}$$

Hyperbolic Tangent Function

The hyperbolic tangent function Equation 3.5is also a sigmoid function, with the range(-1;1)

$$\varphi(u) = \tanh(\frac{u}{2}) \tag{3.5}$$

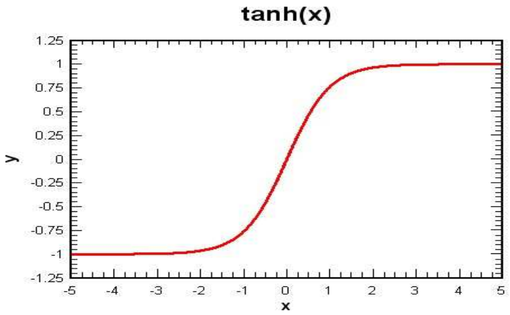

Figure 3.4: Hyperbolic Tangent Function

The hyperbolic tangent function is almost linear close to zero, which means that input values close to zero passes almost unchanged. On the other hand, large input values are squeezed to the limits of the hyperbolic tangent function (i.e. towards 1 or -1). This means that the hyperbolic tangent function has the ability to reduce the effect of outliers in the data.

3.1.2 Feed-Forward Neural Networks

Feed-forward neural networks (FF networks) are the most popular and most widely used models in many practical applications.Figure 3.5 illustrates a one-hidden-layer, Feed Forward network with inputs ,hidden and output . Each arrow in the figure symbolizes a parameter in the network. The network is divided into layers. The input layer consists of just the inputs to the network. Then follows a hidden layer, which consists of any number of neurons, or hidden units placed in parallel. Each neuron performs a weighted summation of the inputs, which then passes a nonlinear activation function, also called the neuron function. Mathematically the functionality of a hidden neuron is described by:

$$\sigma[\sum_{j=1}^{n}(w_j * x_j) + b_j] \tag{3.6}$$

where the weights (wj, bj) are symbolized with the arrows feeding into the neuron. The network output is formed by another weighted summation of the outputs of the neurons in the hidden layer. This summation on the output is called the output layer. In Figure 3.5 there is only one output in the output layer since it is a single-output problem.

Generally, the number of output neurons equals the number of outputs of the approximation

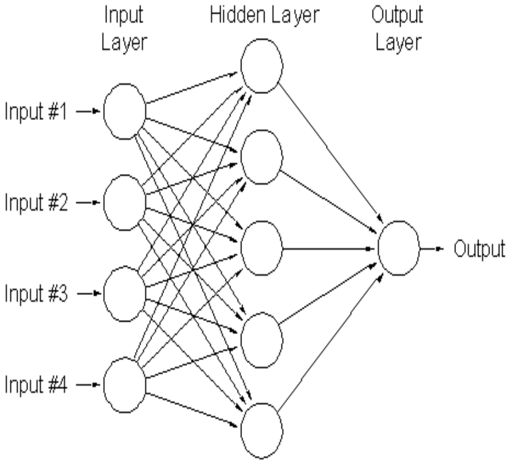

Figure 3.5: Feed-Forward Neural Networks

problem. The neurons in the hidden layer of the network in Figure 3.6 are similar in structure to those of the perceptron, with the exception that their activation functions can be any differential function. The output of this network is given by Equation 3.7:

$$y(\hat{\theta}) = g(\theta, x) = \sum_{i=1}^{m}(w_i^2 \sigma[w_{i,j}^1 * x_j + b_{i,j}^1] + b^2) \tag{3.7}$$

3.1.3 Multi-Layer Perceptron (MLP)

A multilayer perceptron is a feedforward neural network with one or more hidden layers. Typically, the network consists of an input layer of source neurons, at least one middle or hidden layer of computational neurons, and an output layer of computational neurons. The input signals are propagated in a forward direction on a layer-by-layer basis (Negnevitsky, 2005). A multilayer perceptron with two hidden layers is shown in Figure 3.6.

Each layer in a multilayer neural network has its own specific function. The input layer accepts input signals from the outside and redistributes these signals to all neurons in the hidden layer. Actually, the input layer rarely includes computing neurons, and thus does not process input patterns. The output layer accepts output signals, or in other words a stimulus pattern, from the hidden layer and establishes the output pattern of the entire network. Neu-

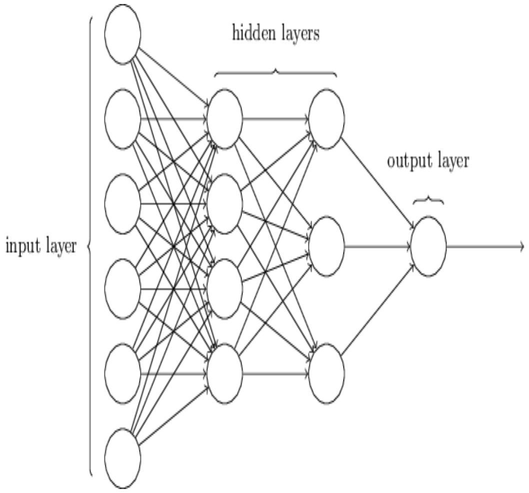

Figure 3.6: Multi-Layer Perceptron MLP

rons in the hidden layer detect the features; the weights of the neurons represent the features hidden in the input patterns. These features are then used by the output layer in determining the output pattern. With one hidden layer, we can represent any continuous function of the input signals, and with two hidden layers even discontinuous functions can be represented.

A hidden layer hides its special output. Neurons in the hidden layer cannot be observed through the input/output behaviour of the network. There is no clear way to know what the special output of the hidden layer should be. In other words, the special output of the hidden layer is determined by the layer itself. Learning in a multilayer network proceeds the same way as for a perceptron. A training set of input patterns is presented to the network. The network computes its output pattern, and if there is an error or in other words a difference between actual and desired output patterns the weights are adjusted to reduce this error.

In a perceptron, there is only one weight for each input and only one output. But in the multilayer network, there are many weights, each of which contributes to more than one output.There are more different learning algorithms are available, but the most popular method is back-propagation.

Particularly for Neural Networks there is a rich literature related to the forecast of the market on daily basis (Neelima Budhani, 2012).

The authors in (Abraham et al., 2001) proposed a hybrid intelligent system based on an artificial neural network trained using scaled conjugate algorithm and a neuro-fuzzy system for stock market analysis,for automated stock market forecasting and trend analysis.The proposed hybrid system can be easily implemented and the em- pirical results are very promising.

Another method used in (Chen et al., 2006) is Flexible Neural Tree (FNT) model for forecasting three major international currency exchange rates. Based on the pre-dened instruction/operator sets, a flexible neural tree model can be created and evolved. the FNT forecasting model may provide better forecasts than the traditional MLFN forecasting model and the ASNN forecasting model.

In (Atsalakis and Valavanis, 2009a) authors explored to use A neuro-fuzzy adaptive control system to forecast next days stock price trends the result show that the accurate predictions of stock price trends are achievable. And the proposed system clearly demonstrates the potential of neuro fuzzy based modeling for financial market prediction.

Bacterial chemotaxis optimization (BCO) and Back propagation neural network (BPNN) in Stock index prediction explored by Zhang and Wu in (Zhang and Wu, 2009) Experiments show its better performance than other methods in learning ability, the model offers less computational complexity, better prediction accuracy, and less training time.

Adaptive neural fuzzy inference system (ANFIS) is adopted to predict stock market return on the ISE National 100 Index used in (Melek Acar Boyacioglu, 2010) The study shows that the performance of stock price prediction can be significantly enhanced by using ANFIS. The prediction performance of this method shows the advantages of ANFIS. It is rapid, easy to operate, and not expensive. The result show the learning and predicting potential of the ANFIS model in financial applications.These results indicate that ANFIS can be a useful tool for stock price prediction in emerging markets.

Number of Soft computing techniques axemen in (Zainab Al Rahamneh, 2010) fuzzy logic (FL), Neural networks (NN) and Neuro-Fuzzy (NF) to deal with the stock exchange forecasting problem. All model provided good forecasting capabilities while the fuzzy model was the best.

Backpropagation neural networks (BPNN)used in (Lahmiri, 2011) the result show that backpropagation neural networks (BPNN) trained with conjugate (BFGS) and the Levenberg Marquardt (LM) provides the best accuracy according to RMSE, MAE, and MAD. Therefore, numerical techniques are suitable to train BPNN than heuristic methods when the problem of S&P 500 stock market forecasting is considered.

In (Abhishek et al., 2012) authors uses simple and efficient approach to stock prediction using Back-Propagation with Feed Forward Network.The network was trained using one year data. It shows a good performance for market prediction, however The network selected was not able to predict exact value but it succeeded in prediction the trends of stock market.

In (Erkam Guresen, 2011) the author used these models multi-layer perceptron (MLP), dynamic artificial neural network (DAN2) and the hybrid neural networks which use generalized autoregressive conditional heteroscedasticity (GARCH) to extract new input variables. The comparison for each model is done in two view points: Mean Square Error (MSE) and Mean Absolute Deviate (MAD) using real exchange daily rate values of NASDAQ Stock Exchange index. The results show that classical ANN model MLP outperforms DAN2 and GARCH-MLP with a little difference. GARCH inputs had a noise effect on DAN2 because of the inconsistencies explained in the previous section and GARCH-DAN2 clearly had the worst results. The problems mention about DAN2 structure clearly shows DAN2 architecture behaves like a statistical method rather than an artificial neural network. The overall results show that classical ANN model MLP gives the most reliable best results in forecasting time series but Hybrid methods failed to improve the forecast results.

The authors in (Fok et al., 2008) use neural networks to predict four stock indices in the United States,Europe, China and Hong Kong. The predictions from the neural network are compared to predications made by a linear regression analysis. The authors find that the neural network predictions are more accurate in terms of the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) performance metrics.

In (Kumar and Haynes, 2003) the author develop an ANN to forecast credit ratings in India and find that the ANN model provides an increased speed and efficiency of the rating process. The authors state that the ANN exhibits an superior performance compared to the benchmark models.

In (Yen et al., 2007) analyse the trend of the price spread between the Taiwan Stock Exchange Electronic Index Futures (TE) and Taiwan Stock Exchange Finance Sector Index Futures (TF). The authors explore different trading models and find that the backpropagation neural network model is superior to the genetic programming model during the analysed period. The authors conclude that it is possible to generate profits trading the spread between the two described index futures and suggest a back-propagation neural network with a momentum trading strategy as the best choice amongst the examined models.

The authors in (Hanias et al., 2007) use back-propagation neutral networks to predict the

Athens stock index. The authors use an architecture with one hidden layer consisting of seven hidden neurons and report a good prediction performance in terms of the mean squared error (MSE) up to nine days ahead.

In (Kara et al., 2011) try to predict the direction of change in the Istanbul Stock Exchange (ISE) National 100 Index. In the study, two forecasting models are developed. The first model is based on artificial neural networks and the second model employs support vector machines. The simulation results show that the neural network performs with 75.74% correct directional forecasts performs better than the support vector machine with 71.52%. The authors conclude that both models show a significant forecasting performance and that both are useful tools in the area of stock index prediction.

3.2 Genetic Programming

GP is one of the evolutionary computational (EC) methods which evolve populations of symbolic tree expressions to solve complex optimization problem was developed by John Koza in 1991(Koza, 1991),GP is known to be domain-independent method which genetically evolves a population of computer programs to solve a problem. GP iteratively evolves a population of computer programs to produce a new generation of programs by mutating and crossing over the best performing trees to create a new population. GP uses the principles of Darwinian natural selection and biologically inspired operations (Koza, 1992). Genetic programming is modification of genetic algorithms with one main variation, the population consists of individuals represented by specific data structure trees but GAs uses a chromosome of genes. In the beginning tree was representing a computer program in functional language LISP; LISP gives a freedom for the evolutionary process of GP to handle data which can be easily manipulation and evaluation. Inner nodes of the trees can represent functions (e.g. arithmetic operators, conditional operators or problem specific functions) and leaves would be terminals, external inputs, constants, zero argument functions.

Terminal s= a,b,c,d,.

non-terminals (functions) =$f_1, f_2, f_3, ...$

Suppose that we have function F that can be written in the form of LISP program as follows:

$$F : (f_1ab(f_2cd)(f_3xy)) \tag{3.8}$$

Function F represented as tree Picture as in Figure 3.7

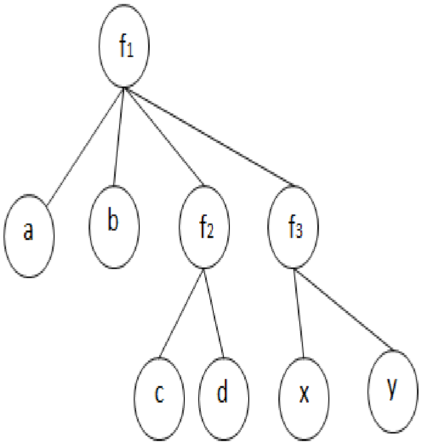

Figure 3.7: LSIP tree representation

GP is used to create experimental mathematical models based on data set collected from a process or system; it is commonly referred to as symbolic regression.
In contrast to classical regression analysis,in which the user must specify the structure of the model, GP automatically evolves both the structure and the parameters of the mathematical model. GP evolutionary process is shown in Figure 3.8. GP procedure could be summarized as follows: (Palit and Popovic, 2005)

- Initialization: generate an initial population of random structures (i.e. tress) using the predefined function set and the given terminal set of the problem at hand; following the rules that specify the legal syntax of the language, and randomly stopping at some point before the maximum initial size of an individual is reached. For a tree-based code representation, a function for the root node is randomly selected, and then the contents of each leaf is set randomly. Then recursively the leaves of any new nodes are filled in, until the randomly selected depth is reached. This process is repeated for each new member of the population until a full population is created.

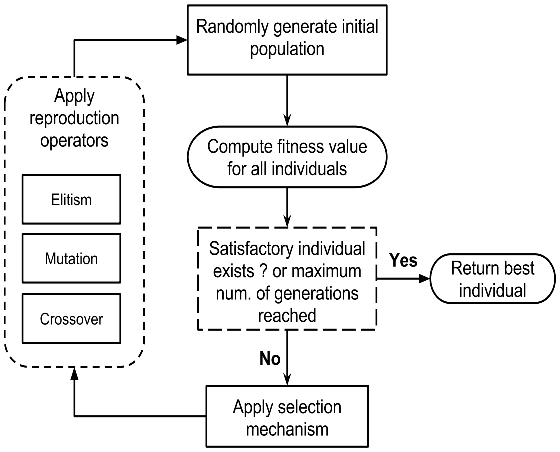
Randomly generate initial population
Compute fitness value for all individuals
Satisfactory individual exists ? or maximum num. of generations reached
Yes
Return best individual
No
Apply selection mechanism
Apply reproduction operators
Elitism
Mutation
Crossover

Figure 3.8: Main loop of the GP algorithm

- Fitness: calculate the fitness values for each tree in the population; trees are the models evolved. The calculated fitness is then stored with each program for use in the next step.
- Reproduction: produce the new population using some selection mechanism and reproduction operators; Various methods exist for selecting who reproduces. The simplest method is to repeatedly take the very best members of the population, until enough parents are chosen to produce the next generation.
- Apply genetic operators until a new population is generated; There are many ways to produce children; the three most common are crossover, mutation, and duplication. These algorithms are called genetic operators because of their conceptual similarity to genetic processes.

In other hand selection and reproduction operations stay the same like in genetic algorithm.

Mutation It is realized by choosing a random node in the tree. Subtree in the chosen node is deleted and new one is randomly generated .as shown in figure 3.9

Crossover Two candidate trees must be selected first for crossover to be applied; one random node is chosen in each parent. Then the sub trees in the chosen nodes are recombined,

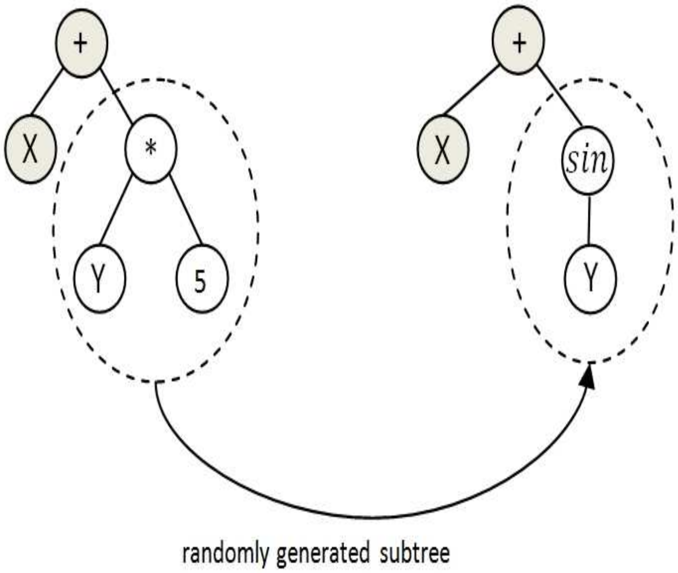

Figure 3.9: Mutation on a tree

i.e. swapped between two parents to the corresponding place.as shown in figure 3.10.

3.2.1 GP Representation

The normally adopted representation in GP is the program tree or parse tree. For example, the simple expression:

$$\sin(y) + \sqrt{y} - \cos(x * x) \tag{3.9}$$

is represented as shown in Figure 3.11. The tree includes nodes (which are also called point) and links. The nodes specify the instructions to execute. The links specify the arguments for each instruction. The internal nodes in a tree are called functions, while the trees leaves will be called terminals. The initial population (i.e. trees) of the GP model is generated randomly. The number of trees represents the size of the population.
In order to apply GP to a problem, the following setup was required by the user.

- Define the set of functions, $F = f_1, f_2, .., f_n$ with arty ¿ 0. Each function in F takes a specified number of arguments, defined as $a_1, a_2, ,$ an.

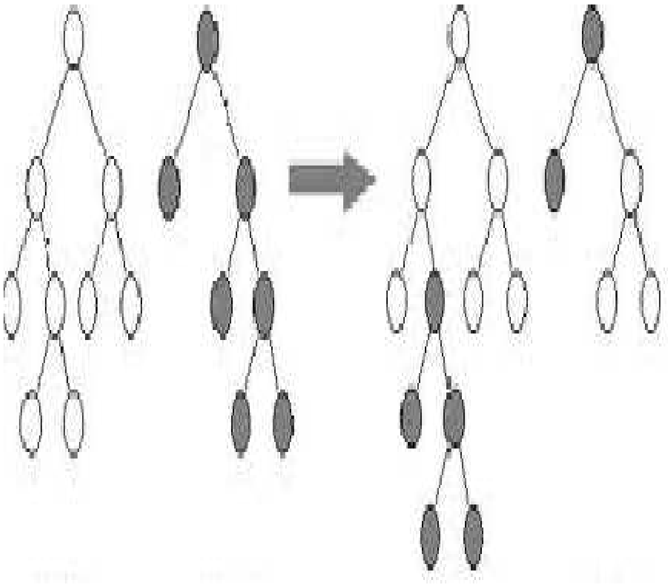

Figure 3.10: Crossover on a tree

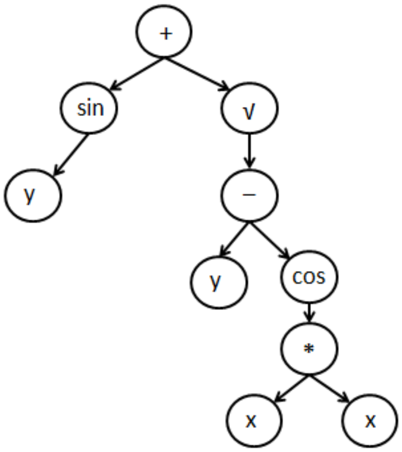

Figure 3.11: Basic tree-like program representation used in GP

- Define the set of terminals,$T = t_1, t_2, .., t_n$ of 0-arty functions (e.g.variable) or constants (e.g. real number, integer).

- Define the fitness function used to evaluate how well each program performs the designated task.

- Define some parameters for controlling the run, such as population size M,the maximum initial depth (depth of program tree), the maximum program depth allowed during the evolving process, and some pre-specified probabilities of genetic operations such reproduction,crossover, mutation, etc.

- Define the method for designating the result and the criterion for terminating a run.

3.2.2 Multigene Symbolic Regression GP

Symbolic regression method was first introduced by J.Koza in (Koza, 1991). The main goal of this method is to search the space of a symbolic description of a model(i.e mathematical expressions) while minimizing some error criteria. Traditional system identification techniques usually adopt two stages of operation: structure determination and parameter estimation. In each stage, some strategy needs to be adopted to select the suitable class of models and to estimation the model parameters using a technique such as linear regression, polynomial approaches or ANN. In contrast, symbolic regression is not similar to traditional linear and nonlinear regression methods, it searches both the space of models along with the space of all possible parameters simultaneously, given that, the model which is searched for is not of any pre-specified structure.
GP is applied to symbolic regression to help evolving a population of trees Koza (1992). For a system with u input of dimension $n * m$ to produce a model output y with dimension $n1$, we could produce a tree structure with a mathematical relationship $y = f(u).n$ is the number of observations taken and m is the number of input variables. In Multigene symbolic regression, each prediction of the output variable $\hat{y}$ is formed by a weighted output of each of the trees/genes in the Multigene individual plus a bias term. Each tree is a function of zero or more of the N input variables $u_1, ..., u_N$. Mathematically, a Multigene regression model can be written as:

$$\hat{y} = \beta_0 + \beta_1 * Tree_1 + ... + \beta_M Tree_M \qquad (3.10)$$

where β_0 represents the bias or offset term while $\beta_1, ..., \beta_M$ are the gene weights and M is the number of genes (i.e. trees) which constitute the available individual. The weights (i.e. regression coefficients) are automatically determined by a least squares procedure for each Multigene individual. In Multigene symbolic regression each symbolic model is represented by number of GP trees weighted by linear combination.

Each tree is considered as a ”gene” by itself. An example of Multigene model is shown in Figure 3.12. The presented model can be introduced mathematically as given in Equation 3.11.

$$\beta_0 + (\beta_1(\sin(y) + \sqrt{x})) + \beta_2 \tan(x * x) \qquad (3.11)$$

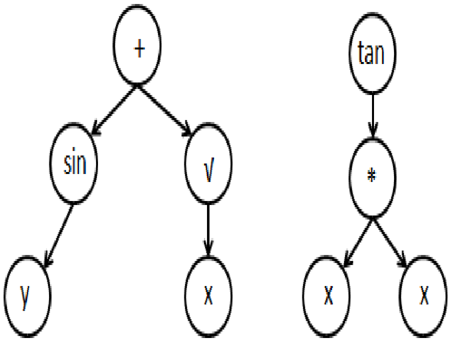

Figure 3.12: Example of a Multigene symbolic GP model

Many authors explore the used of GP (Iba and Sasaki, 1999), (Kaboudan, 2000), (Sheta et al., 2013) in prediction stock market return.

in (El-Telbany, 2005) the author explored genetic programming to forecast the Egyptian Sock Market return. Experiments results show the capability of genetic programming to predict accurate results, comparable to traditional machine learning algorithms i.e., neural networks.

Author in (Hui, 2003) explored the use of GP to perform stock time-series analysis. Given the unpredictable nature of real-life stock data, genetic programming seemed to do a reasonably good job in producing individuals which could give reasonably close predictions and generalize reasonably well during cross-validation. Learning on changes of stock prices gave much better results than learning on the raw stock prices, and having a longer window period and a larger population size also gave rise to best-of-run individuals with better fitness.

In (Refat et al., 2003) author using GP as classification method, Experiments presenting a preliminary result to show the capability of GP to mine accurate classification rules suitable for prediction, comparable to traditional machine learning algorithms.

The authors in (Mori et al., 2005) proposed a new stock price prediction model by applying the GNP to the searching for an optimal combination of two or more proper indices. The effectiveness of the proposed GNP prediction model is confirmed by using simulation studies on real stock dealing data.

In (Rajabioun and Rahimi-Kian, 2008) the author use GP as a decision model in simulation studies showed that the players make decisions in his buy/sell trends (compared to the 10-day-before moving average prices) and traded the maximum number of stocks in each trading day was the most successful one in our virtual stock market.

In (Grosan and Abraham, 2006) introduces two Genetic Programming (GP) techniques: Multi-Expression Programming (MEP) and Linear Genetic Programming (LGP) for the prediction of two stock indices. The performance is then compared with an artificial neural network trained using Levenberg-Marquardt algorithm and Takagi-Sugeno neuro-fuzzy model. Empirical results reveal that Genetic Programming techniques are promising methods for stock prediction.

3.3 Genetic Algorithm (GA)

Genetic algorithm (GA) is a general adaptive optimization search method based on a direct analogy to Darwinian natural selection and genetics in biological systems (Davis et al., 1991). It has been proved to be a promising alternative to conventional heuristic methods. Based on the Darwinian principle of survival of the fittest, GA works with a set of candidate solutions called a population and obtains the optimal solution after a series of iterative computations (Mitchell, 1998). GA evaluates each individuals fitness, quality of the solution, through a fitness function. The fitter chromosomes have higher probability to be kept in the next generation or be selected into the recombination pool using the contest selection methods. If the fittest individual or chromosome in a population cannot meet the requirement, successive populations will be reproduced to provide more alternate solutions. The crossover and mutation functions are the main operators that randomly transform the chromosomes and finally impact their fitness value. The evolution will not stop until acceptable results are obtained. Associated with the characteristics of exploitation and exploration search, GA can deal with large search spaces efficiently, and hence has less chance to get local optimal solution than other algorithms. (Zhuo et al., 2008)

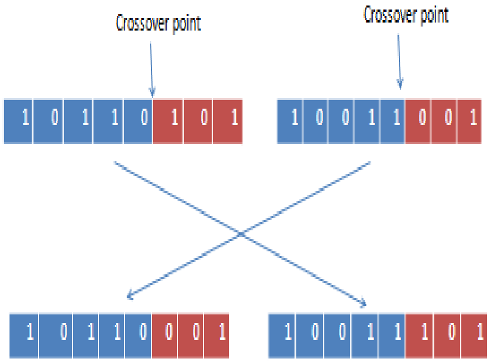

Figure 3.13: GA Crossover

As illustrates in Figures 3.13, 3.14 the genetic operators of crossover and mutation. Crossover is the critical genetic operator that allows new solution regions in the search space to be explored, is a random mechanism for exchanging genes between two chromosomes using the one point crossover, two point crossover, or homologue crossover. In mutation the genes may occasionally be altered, for example, changing the gene value from 0 to 1 or vice versa in a binary code chromosome (Zhuo et al., 2008).

The GA procedure steps is as show in Figure 3.15:

- Representation:

  Represent a chromosome by a binary vector, where each bit of the chromosome tells

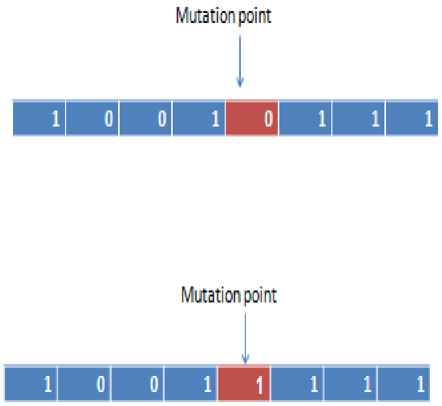

Figure 3.14: GA Mutation

whether the corresponding input feature is selected or not. Initial population

- Fitness Evaluation:

  Evaluates the value of a subset of attributes by taking into consideration the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low inters correlation are preferred.

- Selection:

  Selection, Crossover and Mutation.

- Reproduction item Repeat 3,4,5 and 6 fined good result

3.3.1 GA as feature selection

Since each feature used as part of a prediction procedure can increase the cost and running time of a prediction system, there is strong motivation to design and implement systems with small feature sets.

Genetic algorithm (GA) is best known for their ability to efficiently search large spaces about which little is known a priori. Since genetic algorithms are relatively insensitive to noise, they seem to be an excellent choice for the basis of a stronger feature selection strategy for

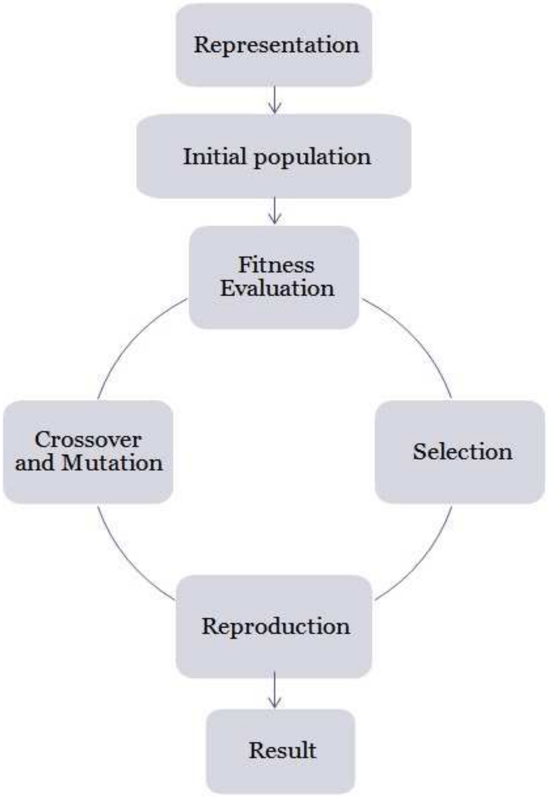

Figure 3.15: GA procedure flow

improving the performance of our quality method of prediction or classification. (Vafaie and De Jong, 1992)

The approach described here involves the use of genetic algorithm as feature selection model in order to identify and select the best subset of features. The main issues in applying GAs to any problem are selecting an appropriate representation and a sufficient evaluation function. In the feature selection problem the main interest is in representing the space of all possible subsets of the given feature set. Then, the simplest form of representation is binary representation where, each feature in the candidate feature set is considered as a binary gene and each individual consists of fixed length binary string representing some subset of the given feature set. An individual of length L corresponds to a L-dimensional binary feature vector X, where each bit represents the removal or addition of the associated feature. Then, xi = 0 represents removal and xi = 1 indicates addition of the ith feature. Genetic algorithms used in (Kim et al., 2006) to combine multiple classifiers to predict the Korea stock price index. The classifiers used consist of a number of different technical indicators. The authors describe a number of combination techniques with a majority vote being the most simple combination method. The authors report that the proposed genetic algorithm based method forecasts the index more precisely that any other of the analysed combination methods. Therefore, it is argued that genetic algorithms pose an effective decision tool for unstructured business problems like stock index forecasting.

In (Majhi et al., 2008) the author propose a clonal particle swarm optimization technique to predict the S&P 500 and DJIA indices. The performance of the model is compared to a standard particle swarm optimization technique and to genetic algorithms. The performance is measured in terms of computational complexity, learning rate, minimum MSE, training time and prediction accuracy. The simulation results show that the clonal particle swarm optimization model performs best out of the three analysed techniques. The authors conclude that the proposed model is a suitable candidate for short- and long-term stock index prediction.

The author in (Majhi et al., 2009) refine the forecasting model using adaptive bacterial foraging optimization. Compared to the particle swarm optimization and genetic algorithm models, the adaptive bacterial foraging optimization is more accurate and shows faster convergence.

In (Zhang et al., 2007) the author use an ensemble of particle swarm optimisation (PSO) and artificial neural networks to forecast the Nasdaq 100 and the S&P CNX Nifty stock index. The proposed algorithm creates several neural networks through bagging and trains the networks using a PSO algorithm. The best networks are then selected and combined. The simulation results indicate that the proposed algorithm is effective and performs better than the benchmark model represented by an genetic algorithm based selective ensemble algorithm.

In (Huang and Wu, 2008) develop a hybrid forecasting model consisting of a genetic algorithm based optimal time-scale feature extraction algorithm and a support vector machine. The input time series is decomposed employing wavelet analysis and genetic algorithms are used to extract the optimal time-scales from the decomposed features. The resulting subsets are then used as input for a support vector machine. The simulation study uses neural networks, simple support vector machines and GARCH models as benchmarks. The simulation results indicate that the proposed model can significantly reduce the forecasting error relative to the analysed benchmark models.

in (Jia et al., 2008) author combine multi expression programming (MEP), particle swarm optimization and artificial neural networks to forecast the S&P CNX Nifty stock index. The authors state that MEP is a variation of genetic programming. A traditional GP encodes a single expression (computer program). By contrast, a MEP chromosome encodes several expressions. The best of the encoded solution is chosen to represent the chromosome. The authors conclude that the proposed model is feasible and effective for stock index forecasting problems.

The author in (Wu et al., 2008) propose an ensemble of artificial neural networks and support vector machines to predict movements in stock market indices. The two machine learning techniques are combined using a linear approach and particle swarm optimisation. The authors state that a combination of these techniques performs significantly better than using each technique individually.

Chapter Four

Research Overview and Methodology

This chapter presents the different methods of this research work and discusses the methodology during the development of the proposed models to achieve the objectives of this research. In this research We develop forecasting model to predict Stock Market Exchange; We have designed three models and compared these models. we used historical data from S&P500 stock market exchanges.
The research consists of the following number of phases :

- Create and preprocess datasets.
- Create a Multiple Linear Regression model to predict the Stock Market Exchange .
- Create a Multilayer Perceptron Neural Network model to predict the Stock Market Exchange .
- Create a Genetic Programming model to predict the Stock Market Exchange for next day.
- Create Feature selection model using GA technique.
- Examine the previous models with new dataset.
- Finally evaluate all these models to get best solution.

Figure 4.1 show the phases of proposed models.

4.1 S&P 500 Index Dataset

The characteristic that all Stock Markets have in common is the uncertainty, which is related with their short and long-term future state. This feature is undesirable for the investor but it

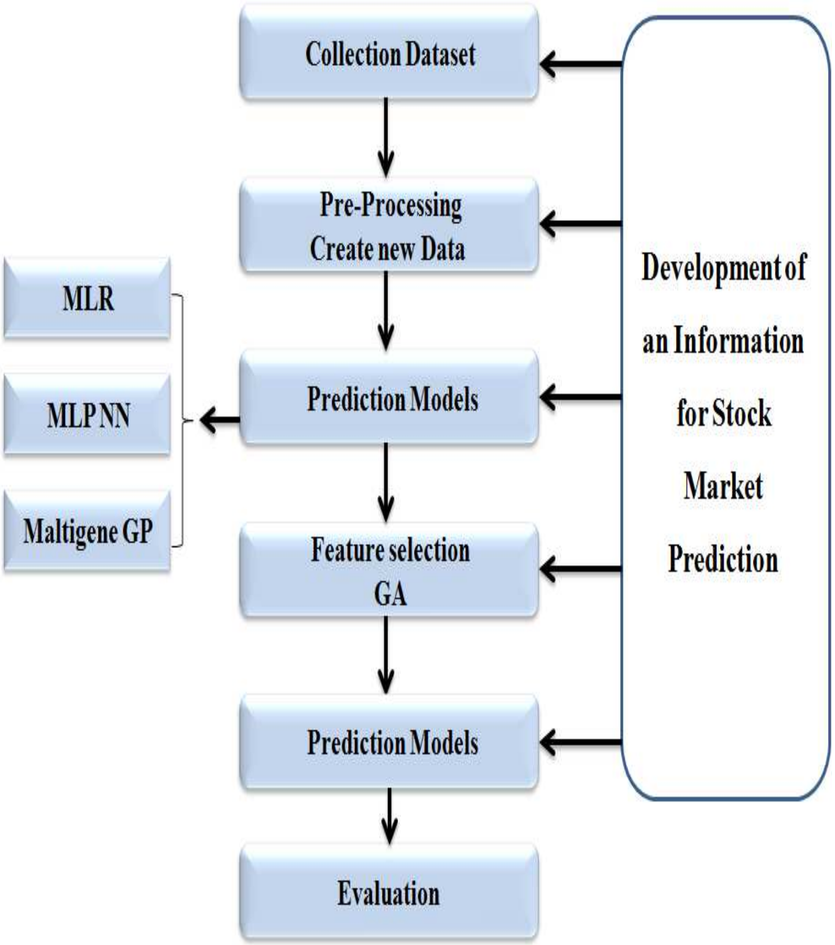

Figure 4.1: Phases of Proposed models

is also unavoidable whenever the Stock Market is selected as the investment tool. The best that one can do is to try to reduce this uncertainty.

Different factors interact in stock market such as Governments and monetary policy, interest rates, International Transactions, general economic conditions, Speculation and Expectation, Supply and Demand, etc (Khadka, 2012). For these reasons we use 27 potential financial and economic variables these factors that impact on stock movement. The main consideration for selecting the potential features is whether they have significant influence on the direction of (S&P 500) index in the next day. While some of these features were used in previous studies (Niaki and Hoseinzade, 2013).

The S& P500, or the Standard & Poors 500, is an American stock market index based on the market capitalizations of 500 large companies having common stock listed on the NYSE or NASDAQ. S& P 500 stock market data set used in our case consists of 27 feature and 1192 days of data, which cover five-years period starting 7 December 2009 to 2 September 2014.The (S& P 500) data were presented and sampled such that data for 596 days is used as training set and data for 596 days is used for testing the developed model. we can categories these feature into 6 group as :

1. G1 : S&P 500 index return in three previous days (SPYt-1, SPYt-2, SPYt-3).
2. G2 : Financial and economical indicators (Oil, Gold, CTB3M, AAA).
3. G3 : The return of the five biggest companies in S&P 500 (XOM, GE, MSFT, PG, JNJ).
4. G4 : Exchange rate between USD and three other currencies (USD-Y, USD-GBP, USD-CAD).
5. G5 : The return of the four world major indices (HIS, FCHI, FTSE, GDAXI).
6. G6 : S&P 500 trading volume (V).

We divided our work into two Solution:

- Daily data: The daily data, in split validation divided 50% to 50% percentage. sampled such that data for 596 days is used as training set and data for 596 days is used for testing in the developed models.

- Weekly data: The weekly data, in split validation divided 70% to 30% percentage. sampled such that data for 100 week is used as training set and data for 43 week is used for testing in the developed models.

The list, description, and the sources of the potential features are given in Table 4.1 show the 27 features of dataset .

No	**Feature**	**Description**
1	$SPY_(t-1)$	The return of the S&P 500 index in day $t-1$ Source data: finance.yahoo.com
2	$SPY_(t-2)$	The return of the S&P 500 index in day $t-2$ Source data: finance.yahoo.com
3	$SPY_(t-3)$	The return of the S&P 500 index in day $t-3$ Source data: finance.yahoo.com
4	Oil	Relative change in the price of the crude oil Source data: finance.yahoo.com
5	Gold	Relative change in the gold price Source data: www.usagold.com
6	CTB3M	Change in the market yield on US Treasury securities at 3-month constant maturity, quoted on investment basis Source data: H.15 Release - Federal Reserve Board of Governors
7	CTB6M	Change in the market yield on US Treasury securities at 6-month constant maturity, quoted on investment basis Source data: H.15 Release - Federal Reserve Board of Governors
8	CTB1Y	Change in the market yield on US Treasury securities at 1-year constant maturity, quoted on investment basis Source data: H.15 Release - Federal Reserve Board of Governors
9	CTB5Y	Change in the market yield on US Treasury securities at 5-year constant maturity, quoted on investment basis Source data: H.15 Release - Federal Reserve Board of Governors
10	CTB10Y	Change in the market yield on US Treasury securities at 10-year constant maturity, quoted on investment basis Source data: H.15 Release - Federal Reserve Board of Governors
11	AAA	Change in the Moody's yield on seasoned corporate bonds - all industries, Aaa Source data: H.15 Release - Federal Reserve Board of Governors

12	BBB	Change in the Moody's yield on seasoned corporate bonds - all industries, Baa Source data: H.15 Release - Federal Reserve Board of Governors
13	USD-Y	Relative change in the exchange rate between US dollar and Japanese yen Source data: OANDA.com
14	USD-GBP	Relative change in the exchange rate between US dollar and British pound Source data: OANDA.com
15	USD-CAD	Relative change in the exchange rate between US dollar and Canadian dollar Source data: OANDA.com
16	HIS	Hang Seng index return in day t-1 Source data: finance.yahoo.com
17	FCHI	CAC 40 index return in day t-1 Source data: finance.yahoo.com
18	FTSE	FTSE 100 index return in day t-1 Source data: finance.yahoo.com
19	GDAXI	DAX index return in day t-1 Source data: finance.yahoo.com
20	V	Relative change in the trading volume of S&P 500 index Source data: finance.yahoo.com
21	XOM	Exxon Mobil stock return in day t-1 Source data: finance.yahoo.com
22	GE	General Electric stock return in day t-1 Source data: finance.yahoo.com
23	MSFT	Micro Soft stock return in day t-1 Source data: finance.yahoo.com
24	PG	Procter and Gamble stock return in day t-1 Source data: finance.yahoo.com
25	JNJ	Johnson and Johnson stock return in day t-1 Source data: finance.yahoo.com
26	DJI	Dow Jones Industrial Average index return in day t-1 Source data: finance.yahoo.com
27	IXIC	NASDAQ composite index return in day t-1 Source data: finance.yahoo.com

Table 4.1: The 27 potential influential features of the S&P 500 Index

4.2 Multiple Linear Regression

Regression analysis measures the degree of influence of the independent variables on a dependent variable. In the case of simple bivariate regression where there is a single independent variable, the dependent variable could be predicted from the independent variable by the simple equation:

$$y = a + bx \tag{4.1}$$

a is constant and b is the slope. This model is linear in the parameters a. y is called the independent variable and x , is called the independent variables. The goal is to find the relationship between the dependent and independent variables. To compute the regression coefficient for the single independent variable given in Equation 4.1 , we use the formula:

$$b = \frac{\sum (x_i - \hat{x})(y_i - \hat{y})}{\sum (x_i - \hat{x})^2} \tag{4.2}$$

Where $\hat{x}$ is the mean (average) of the x values and $\hat{y}$ is the mean of the y values. The parameter a is computed by the formula:

$$a = y - bx \tag{4.3}$$

Equation 4.2 can be expanded to be:

$$b = \frac{(\sum y_i \sum x_i^2) - (\sum x_i \sum x_i y_i)}{n(\sum x_i^2) - (\sum x_i)^2} \tag{4.4}$$

In this research used Multiple Linear Regression ; Equation 4.1 can be expanded to a multivariate concept as follows:

$$y = a_1 x_{i1} + a_2 x_{i2} + ... + a_n x_{in} \tag{4.5}$$

Where x_{ij} is the ith observation on the jth independent variable. To show how the parameter estimation process works, we have a system with 27 input variables $x_1, x_2, x_3, ..., x_{27}$

and single output y. Thus, the model mathematical equation can be represented as:

$$y = a_0 + a_1 x_1 + a_2 x_2 + ... + a_{27} x_{27} \tag{4.6}$$

To find the values of the model parameters as we need to build what is called the regression matrix ϕ. This matrix is developed based on the experiment collected measurements. Thus,ϕ can be presented as follows given there is a set of measurements m : The *characteristic polynomial* ϕ of the $27 \times m$

$$X = \begin{pmatrix} 1 & x_1^1 & x_2^1 & x_3^1 & ... & x_{27}^1 \\ 1 & x_1^2 & x_2^2 & x_3^2 & ... & x_{27}^2 \\ 1 & x_1^3 & x_2^3 & x_3^3 & ... & x_{27}^3 \\ . & . & . & . & ... & . \\ . & . & . & . & ... & . \\ . & . & . & . & ... & . \\ 1 & x_1^m & x_2^m & x_3^m & ... & x_{27}^m \end{pmatrix} \tag{4.7}$$

The parameter vector θ and the output vector y can be presented as follows:

$$\theta = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ . \\ . \\ . \\ a_{27} \end{pmatrix} \tag{4.8}$$

$$y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ . \\ a_m \end{pmatrix} \tag{4.9}$$

The least squares solution of yields the normal equation:

$$\phi^T \theta = y \tag{4.10}$$

which has a solution:

$$\theta = \phi^{-1} y \tag{4.11}$$

But since, the regression matrix ϕ is not a symmetric matrix, we have to reformulate the equation such that the solution for the parameter vector θ is as follows:

$$\theta = (\phi^T \phi)^{-1} \phi^T y \tag{4.12}$$

4.3 Multilayer Perceptron Neural Network model

ANNs are mathematical models which were inspired from the understanding of some ideas and aspects of the biological systems, especially the human brain (Krose and van der Smagt, 2009) .

The artificial neural net attempts to follow the biological neuron, by modelling its response characteristics using an enhancing activation function (f(s)), similar to conduction of a signal through a resistance. The synapse of the neuron is imitated by a non-linear limiting function which performs amplitude limitation.

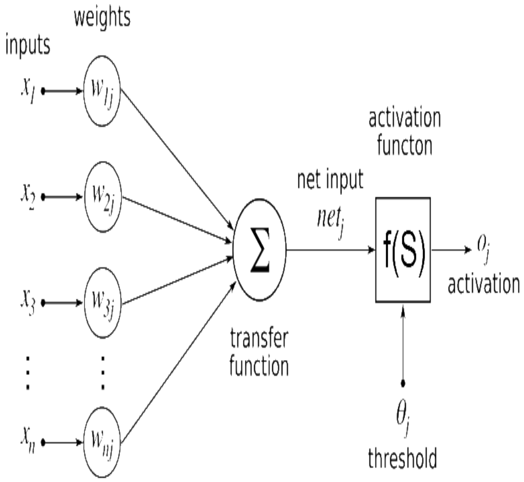

Figure 4.2: Artificial Neural Net durofy.com (durofy.com)

In Figure 4.2 x(i) is the ith input signal, and w(i) is the weight, associated with the input. The weight is the quantity that may get updated in a number of iterations in the learning process, such that the neural net works more and more in the fashion of the required system model, producing more suitable outputs for a set of inputs. The net can be implemented using a mathematical Sigmoid function, where output is given as:

$$\phi(S) = \frac{1}{1+e^{-s}} \tag{4.13}$$

Also, by other functions such as the tanh function, signum and step functions. In feed forward Multilayer Perceptron (MLP), which is one of the most common Neural Network systems, neurons are organized in layers. Each layer consists of a number of processing elements called neurons, each of which contains a summation function and activation function. The summation function is given by Equation 4.14 and the activation function can be a sigmoid function which is given in Equation 4.13.

$$S = \sum_{i=0}^{n}(w_i * x_i) \tag{4.14}$$

A multilayer perceptron is a feedforward neural network with one or more hidden layers.

Typically, the network consists of an input layer of source neurons, at least one middle or hidden layer of computational neurons, and an output layer of computational neurons. The input signals are propagated in a forward direction on a layer-by-layer basis(Negnevitsky, 2005). Mathematically the functionality of a hidden neuron is described by:

$$\sigma[\sum_{j=1}^{n}(w_j * x_j) + b_j] \tag{4.15}$$

where the weights (wj, bj) are symbolized with the arrows feeding into the neuron. The output of the network is given by Equation 4.16:

$$y(\hat{\theta}) = g(\theta, x) = \sum_{i=1}^{m}(w_i^2 \sigma[w_{i,j}^1 * x_j + b_{i,j}^1] + b^2) \tag{4.16}$$

In order to design an appropriate ANN for a particular problem, one should decide on the network topology, number of network layers, number of nodes in each layer, activation function of the nodes, and finally, the learning algorithm. The basic architecture of the MLP Network used to model the stock price prediction problem consists of three layers with single hidden layer. Thus input layer of our neural network model has 27 input nodes while the output layer consists of only one node that gives the predicted next week value. Empirically, we found that 20 neurons in the hidden layer achieved the best performance. The Backpropagation algorithm is used to train the MLP and update its weight. Table 4.2 shows the settings used for MLP.

Maximum number of epochs	500
Number of Hidden layer	1
Number of neurons in hidden layer	20
Learning rate	0.5
Momentum	0.2

Table 4.2: The Setting of MLP

4.4 Genetic Programming model

GP procedure could be summarized as follows:

- generate an initial population of random tree structures using the predefined function set and the given terminal set of the problem at hand;
- calculate the fitness values for each tree in the population; trees are the models evolved.
- produce the new population using some selection mechanism and reproduction operators;
- repeat the procedure until an appropriate solution is found or the end of iterations.

Fitness function is essential for any evolutionary computation process. In the case of GP, we adopted the Mean Squares Error (MSE) as a fitness function as given in Equation 4.17.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y - \hat{y})^2 \tag{4.17}$$

4.4.1 Multigene Symbolic Regression GP

We adopted a GPTIPS toolbox (Searson et al., 2010) to develop our results. In GPTIPS, the initial population is constructed by creating individuals that contain randomly generated GP trees with between 1 and G_{max} genes. During the run, genes are acquired and deleted using a tree crossover operator called two point high level crossover. This allows the exchange of genes between individuals and it is used in addition to the ”standard” GP recombination operators.

Some parameters have to be defined by the user at the beginning of the evolutionary process. They include: population size, probability of crossover, mutation probability and the type of the selection mechanism. User has also to setup the maximum number of genes G_{max} where a model is allowed to have. The maximum tree depth D_{max} allows us to change the complexity of the evolved models. Restricting the tree depth helps evolving simple model but it may also reduce the performance of the evolved model.

A prior knowledge on the problem domain helps in designing a function set which could speed up the evolutionary process for model development. The adopted function set to develop the GP model is given as:

$$F = \{+, -, \times\}$$

Table 5.28 show the tuning parameter of the model.

Population size	100
Number of generation	100
Selection mechanism	Tournament
Max. tree depth	7
Probability of Crossover	0.85
Probability of Mutation	0.1
Max. No. of genes allowed	6

Table 4.3: GP Tuning Parameters

4.5 Feature selection model using Genetic Algorithm (GA) technique

We obtain a set of 27 features in S&P 500 dataset. A Genetic Algorithm is now used to select a set of salient features from among them.The selected features are used as inputs to the developed models. The purpose here is to obtain an optimal subset of features which produce the best possible results. The various steps in the GA are described below:

- Representation: We represent a chromosome by a binary vector of size 27, where each bit of the chromosome tells whether the corresponding input feature is selected or not.
- Fitness Evaluation: CfsSubsetEval : Evaluates the value of a subset of attributes by taking into consideration the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low inters correlation are preferred.

- Selection: Roulette Wheel selection is used for parent selection. Thus, chromosomes with high fitness scores get selected more often.

- Crossover and Mutation are then carried out to produce a new generation.

- Stopping Condition: The GA stops when it does not find a better solution for a fixed number of generations.

A genetic algorithm follows the following pseudo-code :

Algorithm 1: Basic steps describing the GA algorithm

```
begin GA
    input : n, X, m
    Initialise generation 0;
    $J \leftarrow 0$;
    $P_J \leftarrow$ a population of $n$ randomly generated individuals;
    Evaluate $P_J$;
    Compute Fitness $(i)$ for each $i \epsilon PJ$;
    repeat
    Create generate $J + 1$;
    Copy; Select $(1 - X) \times n$ members of $P_J$ and insert into $P_J + 1$;
    Crossover; Select $X \times n$ members of $P_J$; pair them up; produce offspring; insert the
offspring into $P_J + 1$ ;
    Mutate; Select $m \times n$ members of $P_J + 1$ ; invert a randomly selected bit in each;
    Evaluate $P_J + 1$; Compute Fitness $(i)$ for each $i \epsilon PJ$;
    Increment; $J = J + 1$;
    until fitness of fittest individual in $P_J$ is high enough;
    return the fittest individual from $P_J$
    output: a optimal population
```

Population size	50
Number of generations	50
Probability of crossover	0.6
Probability of mutation	0.033
Report frequency	20
Random number seed	1

Table 4.4: GA Tuning Parameters

4.6 Examine the previous models with new data set

In this phase we examine the three models with new dataset that created by feature selection model to investigate if these new features is improve the results and enhance the prediction by decrease the attributes.

4.7 Performance Measures

The performance measures include Variance-Accounted-For (VAF) , Mean Square Error (MSE), Mean Absolute Error (MAE), Root Mean Square Error (RMSE) . They are given in Equations 4.18, 4.19, 4.20, 4.21 respectively.

- Variance Accounted- For (VAF) :

$$VAF = [1 - \frac{var(y_1 - y_2)}{var(y_1)}] * 100\% \tag{4.18}$$

- Mean Squared Error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4.19}$$

- Mean Absolute Error (MAE):

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|(y_i - \hat{y}_i)| \tag{4.20}$$

- Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y-\hat{y})^2} \tag{4.21}$$

where y is real actual value, $\hat{y}$ it the estimated target value. n is the total number of measurements and $\overline{y}$ is the mean of the signal y using n measurements.

4.7.1 Benchmarks

When evaluating a forecast model, performance measures are usually not enough. This since a forecast model with low performance measures can still be the best available alternative. Thus by comparing the performance measures of the prediction model with benchmarks, the evaluation becomes more meaningful.

An example of an easy to use benchmark is the linear regression, where a forecast is based on the previous value (of the time series). More advanced prediction models can also be used as benchmarks, e.g. multi layered perceptron networks . Another possibility is to use a suitable index as a benchmark. Obviously, when comparing a forecast model with a benchmark, it is very important that the performance measures are based on the same time periods (i.e. the same generalization set) and have the same scale. When comparing the models used Variance Accounted- For (VAF) of a forecast model with a benchmark,it can be helpful to calculate the quotient between them, as done in Equation 5.10.

$$VAF_q = \frac{VAF_f}{VAF_b} \tag{4.22}$$

If the quotient is nearest to one or one the model is good or there’s no different.

4.8 Validation criteria

X-Validation The X-Validation is a nested operator. It has two sub processes: a training sub process and a testing sub process. The training sub process is used for training a model. The trained model is then applied in the testing sub process. The performance of the model is also measured during the testing phase.

The data is partitioned into k subsets of equal size. Of the k subsets, a single subset is retained as the testing data set (i.e. input of the testing sub process), and the remaining k 1 subsets are used as training data set (i.e. input of the training sub process). The cross-validation process is then repeated k times, with each of the k subsets used exactly once as the testing data. The k results from the k iterations then can be averaged (or otherwise combined) to produce a single estimation. The value k can be adjusted using the number of validations parameter.

Split Validation The Split Validation is a nested operator. It has two sub processes: a training sub process and a testing sub process. The training sub process is used for learning or building a model. The trained model is then applied in the testing sub process. The performance of the model is also measured during the testing phase. The data set is partitioned into two subsets. One subset is used as the training set and the other one is used as the test set. The size of two subsets can be adjusted through different parameters. The model is learned on the training set and is then applied on the test set. This is done in a single iteration, as compared to the X-Validation operator that iterates a number of times using different subsets for testing and training purposes.

Chapter Five

Experimental Results

5.1 Solution 1: Daily Data prediction

5.1.1 Multiple Linear Regression Model

The regression model have the following equation system.

$$y = a_0 + \sum_{i=1}^{27} a_i * x_i \tag{5.1}$$

y is the actual value on day x_i is the previous day value.
The values of the parameters a_i be estimated using LSE to produce the optimal values of the parameters θ.
The produced linear regression model can be presented as given in Equation 5.3 .All evaluation results are shown in Table 5.1 and 5.2shown the split and cross validation the performance is good and the result in split validation look better tan another one. The actual and Estimated S&P 500 index values based the MLR in both training and testing cases are shown in Figure 5.1 and Figure 5.2, respectively. The scattered plot for the developed MLR model is shown in Figure 5.3. in this research we use the MLR model as benchmark model because its a classical model of prediction.

	MLR MODEL	
	Training	Testing
VAF	99.8266	99.9400
MSE	0.1628	0.2542
MAE	0.3096	0.4017
RMSE	0.4828	0.5042

Table 5.1: Evaluation Criteria for the developed MLR model in split validation

	MLR MODEL
	Cross-validation
MSE	0.5484
MAE	0.5815
RMSE	0.7406

Table 5.2: Evaluation Criteria for the developed MLR model in cross validation

$$\begin{aligned}
SPY &= 0.0925 * SPY(t-1) + 0.0333 * SPY(t-2) \\
&+ (-0.0184) * SPY(t-3) \\
&+ 0.0362 * OIL + (-0.0095) * GOLD + (-8.3406) * CTB3M \\
&+ 5.6003 * CTB6M * (-1.5881) * CTB5Y + 1.2858 * CTB10Y \\
&+ 0.0426 * USD/JPY + 16.5726 * USD/GBP + 3.9301 * USD/CAD \\
&+ (-0.0002) * HIS + 0.1504 * FCHI + (-0.0004) * FTSE100 \\
&+ (-0.0004) * GDAXI + 0 * V + 0.2069 * XOM \\
&+ 0.5831 * GE + (-0.0884) * MSFT + 0.0331 * PG \\
&+ 0.0702 * JNJ + (-0.0002) * DJI + 0.0252 * IXIC \\
&+ (-1.9673) \qquad (5.2)
\end{aligned}$$

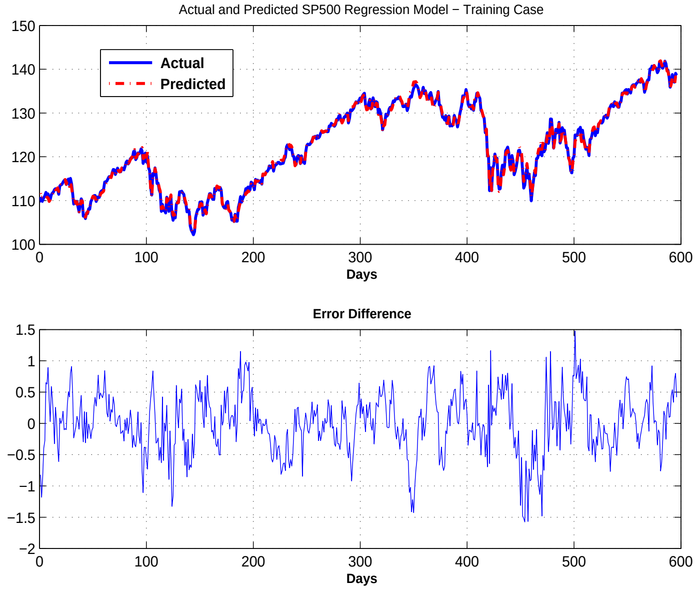

Figure 5.1: MLR Actual and Estimated S&P 500 Index values in Training Case

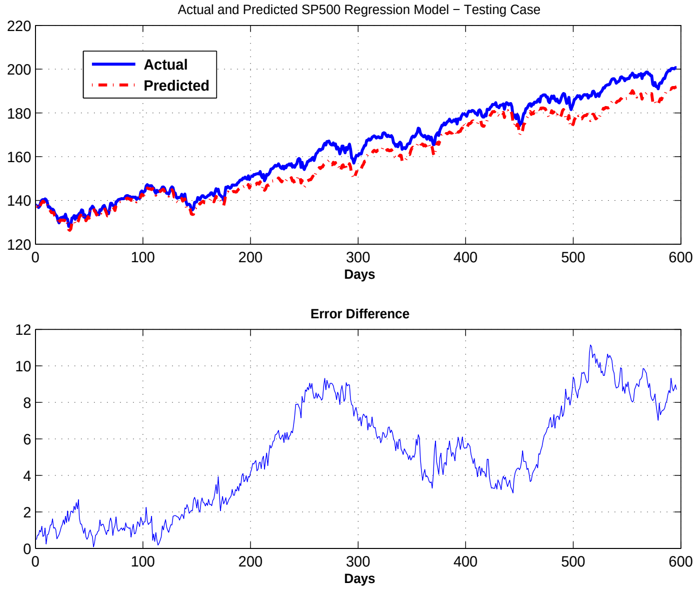

Figure 5.2: MLR Actual and Estimated S&P 500 Index values in Testing Case

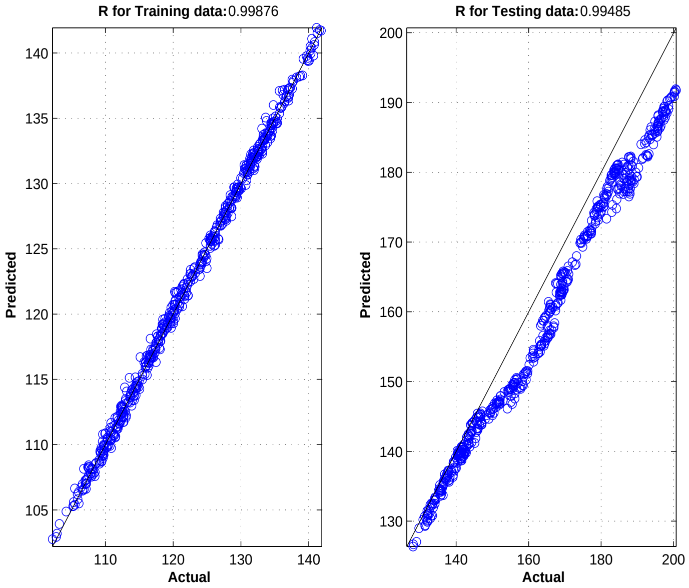

Figure 5.3: MLR Scattered

5.1.2 Multilayer Perceptron Neural Network model

The basic architecture of the MLP Network used to model the stock price prediction problem consists of three layers with single hidden layer. Thus input layer of our neural network model has 27 input nodes while the output layer consists of only one node that gives the predicted next day value. Empirically, we found that 30 neurons in the hidden layer achieved the best performance 5.5. The Backpropagation algorithm is used to train the MLP and update its weight. as we describe in Research Overview chapter.

The training phase of the MLP run converged to the minimum MSE value after 30 epochs as shown in Figure 5.6. All evaluation results are shown in Table 5.3and 5.4 the cross validation seems better . Figures 5.7 and Figure 5.8 depict actual and predicted stock prices for training and testing cases of the developed MLP. The scattered plot for the developed MLP model is shown in Figure 5.9.

5.1.3 Multigene Genetic Programming Model

The Multigene GP was applied using GPTIPS Tool. The parameters of the algorithm were tuned as listed in Table 5.6.first we axemen various population size and choose the best one figure 5.10 explained the convergences. The default recombination operator were used listed in Table5.5. In Figure 5.11, we show the convergence of GP over 300 generations

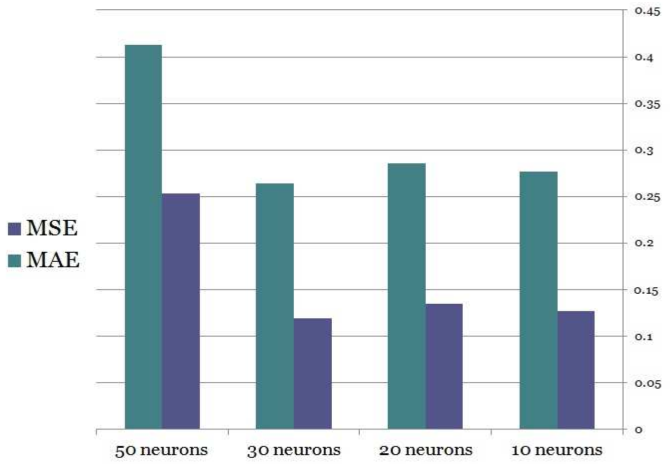

Figure 5.4: Different number of neurons in hidden layer

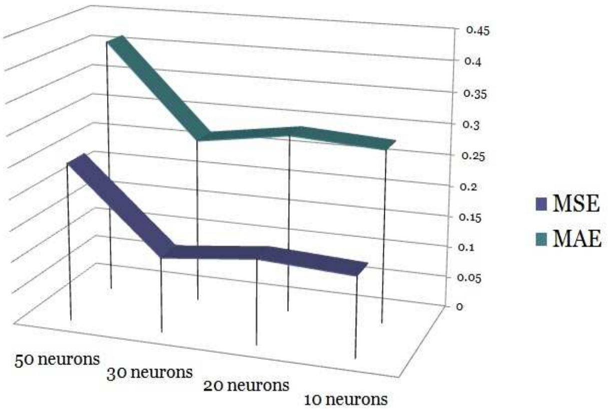

Figure 5.5: Different number of neurons in hidden layer

	MLP MODEL	
	Training	Testing
VAF	99.6895	98.7892
MSE	0.2223	1.0320
MAE	0.2774	0.390
RMSE	0.3564	0.4851

Table 5.3: Evaluation Criteria for the developed MLP model in split validation

	MLP MODEL
	Cross-validation
MSE	0.47707
MAE	0.5388
RMSE	0.6907

Table 5.4: Evaluation Criteria for the developed MLP model in cross validation

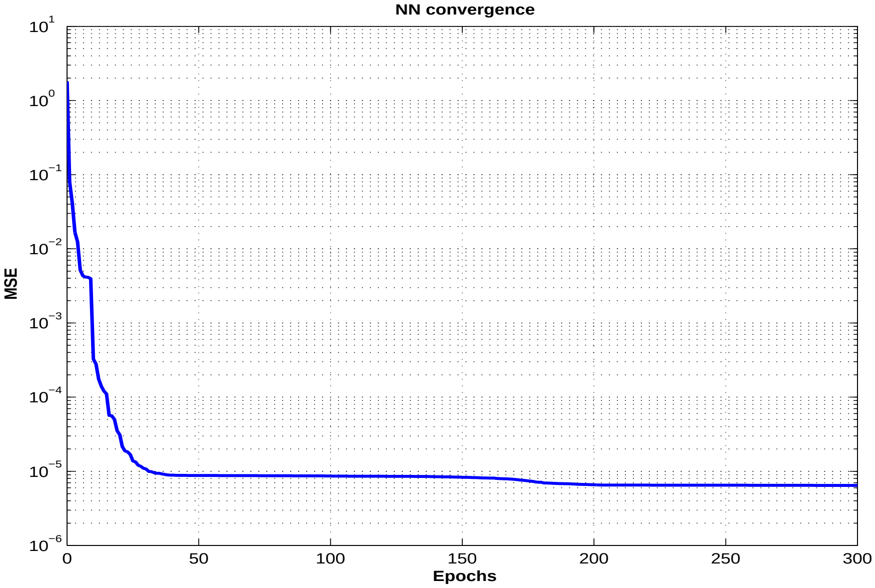

Figure 5.6: MLP Convergence

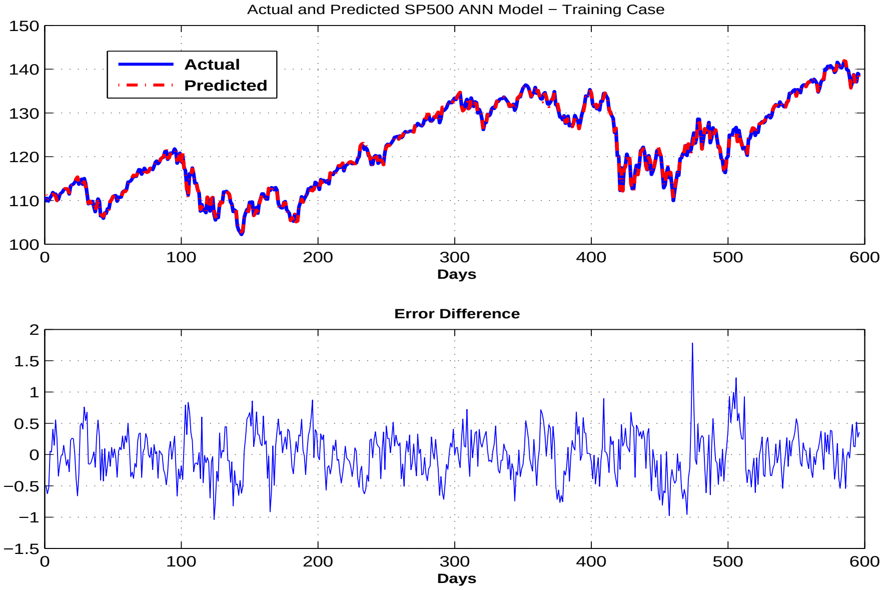

Figure 5.7: MLP Actual and Estimated S&P 500 Index values in Training Case

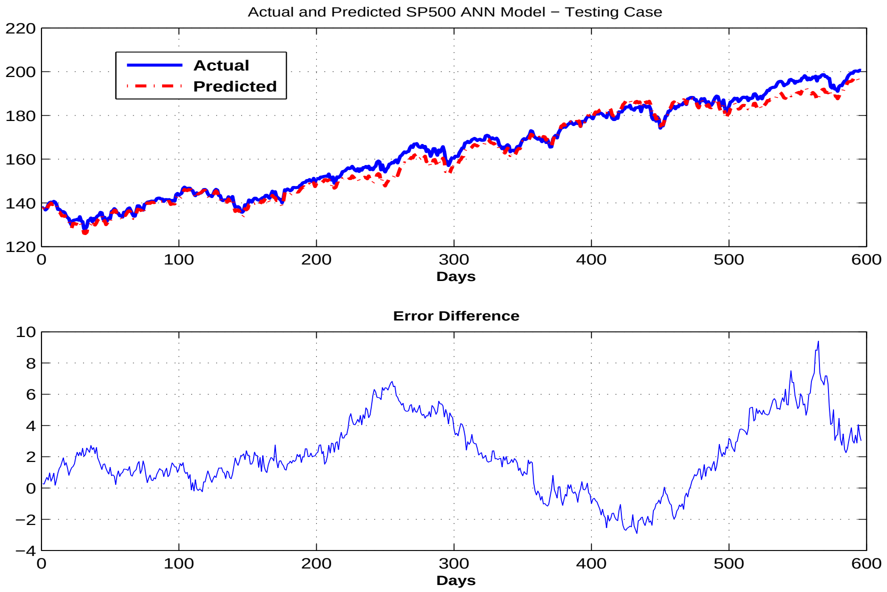

Figure 5.8: MLP Actual and Estimated S&P 500 Index values in Testing Case

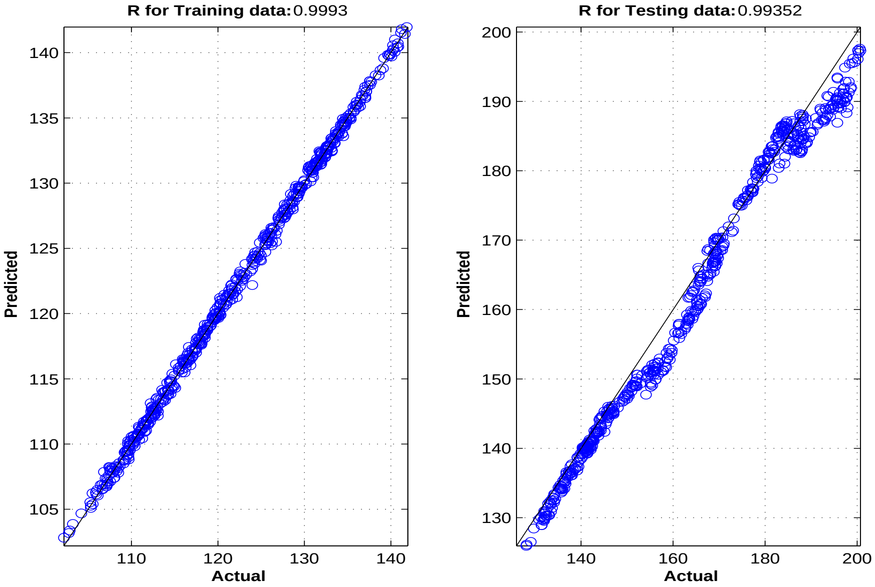

Figure 5.9: MLP Scattered

and Figure 5.14 Show the scatterd The best generated stock market Multigene GP model is given in Equation 5.3. It can be clearly seen that the final model is a simple and compact mathematical model which is easy to evaluate. The performance measurements for the model were computed and summarized in Table 5.30. Figure 5.12 and Figure 5.13 shows the actual and estimated stock market in training and testing.

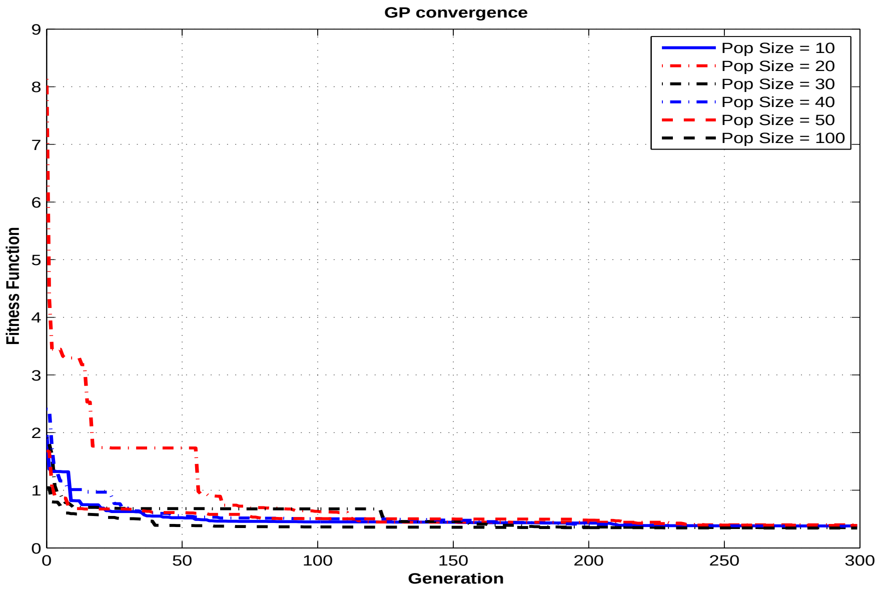

Figure 5.10: Convergence of MGP with various population sizes

high level crossover	0.2
low level crossover	0.8
subtree mutation	0.9
replace input terminal with another random terminal	0.05
standard deviation of Gaussian	0.1

Table 5.5: Default Parameters

$$\begin{aligned} y &= 0.01449 * SPY(t-1) + 0.05648 * OIL + 0.0473 * USD/JPY + 0.2167 * FCHI \\ &+ 0.1183 * XOM + 0.2167 * GE + 1.928e-5 * PG + 0.02365 * SPY \\ &+ 1.928e-5 * x22 * SPY(t-1) + 1.928e-5 * SPY(t-1) * XOM * GE \\ &- 1.928e-5 * SPY(t-1) * FCHI * (FCHI - JNJ) \\ &- 3.856e-5 * SPY(t-1) * GE * (MSFT - PG) + 19.26 \end{aligned} \tag{5.3}$$

Population size	30
Number of generation	1000
Selection mechanism	Tournament size = 2
Max. tree depth	7
Probability of Crossover	0.85
Probability of Mutation	0.1
Max.No.of genes allowed in an individual	6
direct reproduction	0.05
Function node set	plus, minus, times
Ephemeral random constants in the range	[-10 10]

Table 5.6: Tuning Parameters

	MGP MODEL	
	Training	Testing
VAF	99.674	99.321
MSE	0.30626	6.1108
MAE	0.41564	2.0588
RMSE	0.58272	3.0235

Table 5.7: Evaluation Criteria for the MGP Model

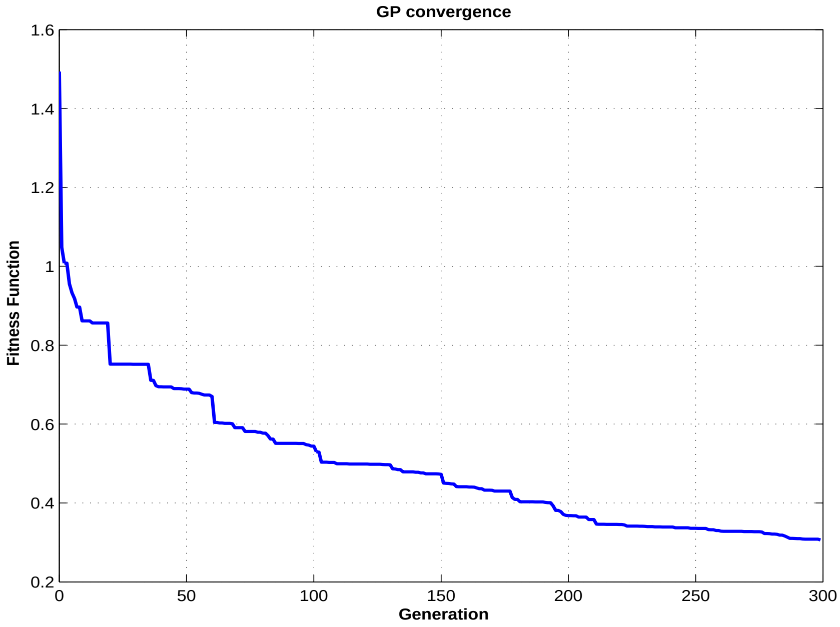

Figure 5.11: MGP Convergence

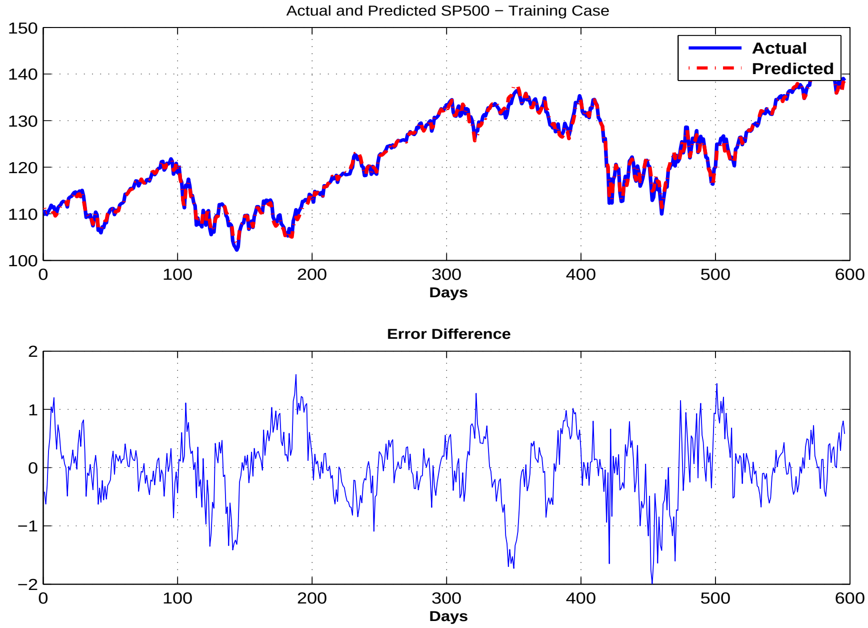

Figure 5.12: MGP Actual and Estimated S&P 500 Index values in Training Case

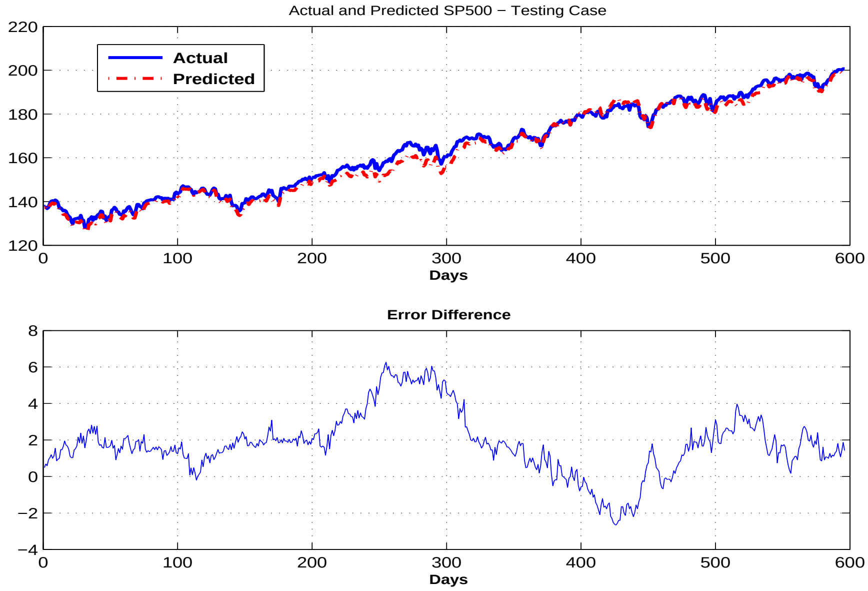

Figure 5.13: MGP Actual and Estimated S&P 500 Index values in Testing Case

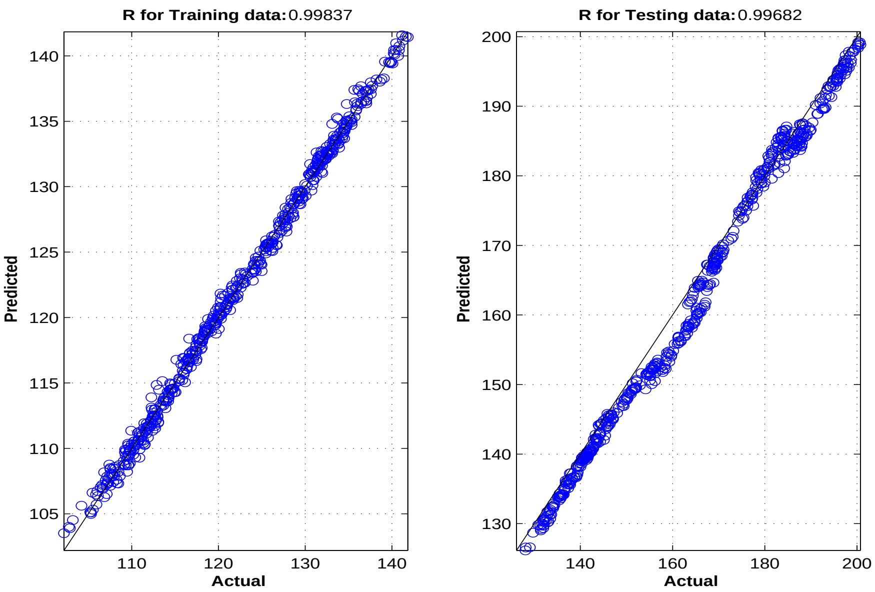

Figure 5.14: MGP Scattered

5.1.4 Feature Selection

we train the GA selection model with different numbers of population size 20,50 and 100 as shown in table 5.8 and chosen the best one.

The parameters of the algorithm were tuned as listed in Table 5.9 The Selected attributes is: 1,2,3,6,7,8,10,12,21,22,23,24,25,27 = 14 attribute as shown in table 5.10

5.1.5 Linear Regression Model with selected attributes

In this section we explain the regression model with 14 attribute that selected by GA. Equation 5.4 show the model of regression and figure 5.15 and 5.16 show the actual and estimated variable. Table 5.11 and 5.12 summarise the evaluation criteria.

$$\begin{aligned} SPY &= 0.3405 * SPYt - 1 + 0.0141 * SPYt - 2 - 0.0282 * SPYt - 3 - 4.7906 * CTB3M \\ &- 7.1253 * CTB6M - 7.1699 * CTB1Y - 1.1133 * CTB10Y - 1.5124 * BBB \\ &+ 0.1039 * XOM + 0.3118 * GE + 0.1016 * MSFT + 0.1789 * PG \\ &+ 0.2159 * JNJ + 0.0158 * IXIC + 14.1416 \end{aligned} \tag{5.4}$$

Attribute Selected		
Population size=20	Population size=50	Population size=100
10(100 %) SPYt-1	10(100 %) SPYt-1	10(100 %) 1 SPYt-1
10(100 %) 2 SPYt-2	7(70 %) 2 SPYt-2	1(10 %) 2 SPYt-2
8(80 %) 3 SPYt-3	8(80 %) 3 SPYt-3	0(0%) 3 SPYt-3
0(0 %) 4 OIL	1(10 %) 4 OIL	0(0 %) 4 OIL
0(0 %) 5 Gold -GLD	0(0 %) 5 Gold -GLD	0(0 %)5 Gold -GLD
2(20 %) 6 CTB3M	7(70 %) 6 CTB3M	5(50 %) 6 CTB3M
5(50 %) 7 CTB6M	5(50 %) 7 CTB6M	7(70 %) 7 CTB6M
7(70 %) 8 CTB1Y	5(50 %) 8 CTB1Y	5(50 %) 8 CTB1Y
0(0 %) 9 CTB5Y	1(10 %) 9 CTB5Y	0(0 %) 9 CTB5Y
1(10 %) 10 CTB10Y	5(50 %) 10 CTB10Y	1(10 %) 10 CTB10Y
0(0 %) 11 AAA	1(10 %) 11 AAA	0(0 %) 11 AAA
1(10 %) 12 BBB	7(70 %) 12 BBB	0(0 %) 12 BBB
0(0 %) 13 USD/JPY	1(10 %) 13 USD/JPY	0(0 %) 13 USD/JPY
4(40 %) 14 USD/GBP	4(40 %) 14 USD/GBP	2(20 %) 14 USD/GBP
2(20 %) 15 USD/CAD	2(20 %) 15 USD/CAD	3(30 %) 15 USD/CAD
0(0 %) 16 HIS	0(0 %) 16 HIS	0(0 %) 16 HIS
0(0 %) 17 FCHI	1(10 %) 17 FCHI	0(0 %) 17 FCHI
0(0 %) 18 FTSE100	0(0 %) 18 FTSE100	0(0 %) 18 FTSE100
0(0%) 19 GDAXI_DAX	0(0 %)19 GDAXI_DAX	0(0 %) 19 GDAXI_DAX
0(0 %) 20 V	0(0 %) 20 V	0(0 %) 20 V
10(100 %) 21 XOM	10(100 %) 21 XOM	10(100 %) 21 XOM
10(100 %) 22 GE	10(100 %) 22 GE	10(100 %) 22 GE
6(60 %) 23 MSFT	5(50 %) 23 MSFT	0(0 %) 23 MSFT
1(10 %) 24 PG	5(50 %) 24 PG	0(0 %) 24 PG
10(100 %) 25 JNJ	10(100 %) 25 JNJ	10(100 %) 25 JNJ
0(0 %) 26 DJI	0(0 %) 26 DJI	0(0 %) 26 DJI
10(100 %) 27 IXIC	10(100 %) 27 IXIC	10(100 %) 27 IXIC

Table 5.8: The result of GA selection model with Three population size

Start set	no attributes
Population size	50
Number of generations	50
Probability of crossover	0.6
Probability of mutation	0.033
Report frequency	20
Random number seed	1

Table 5.9: The Tuning Parameter of GA model

S & P 500 selection attribute
SPYt-1
SPYt-2
SPYt-3
CTB3M
CTB6M
CTB1Y
CTB10Y
BBB
XOM
GE
MSFT
PG
JNJ
IXIC

Table 5.10: The S& P 500 14 Selected attributes

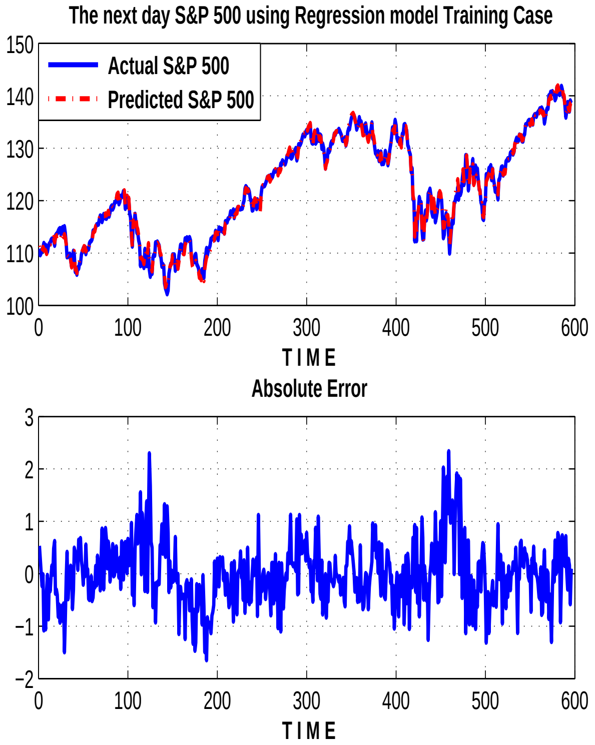

Figure 5.15: MLR Actual and Estimated S&P 500 Index values in Training Case

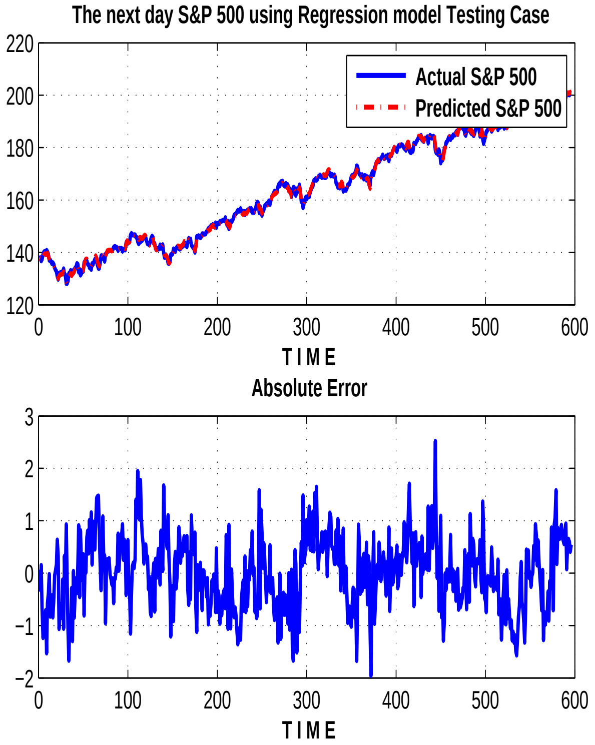

Figure 5.16: MLR Actual and Estimated S&P 500 Index values in Test Case

	MLR	
	Training	Testing
VAF	99.6169	99.8885
MSE	0.3597	0.4725
MAE	0.532	2.8959
RMSE	0.7073	3.6286

Table 5.11: Evaluation Criteria for the developed MLR model in split validation

	MLR
	Cross-validation
MSE	0.86341
MAE	0.7214
RMSE	0.9292

Table 5.12: Evaluation Criteria for the developed MLR model in cross validation

5.1.6 Multilayer Perceptron Neural Network model with selected attribute

In this section we explain the MLP model with 14 attribute that selected by GA. Figures 5.17 and 5.18 show the actual and estimated variable. Table 5.13 and 5.14 summarise the evaluation criteria.

	MLP MODEL	
	Training	Testing
VAF	99.7521	99.9116
MSE	0.2644	0.4234
MAE	0.3776	0.5303
RMSE	0.4962	0.649

Table 5.13: Evaluation Criteria for the developed MLP model in split validation

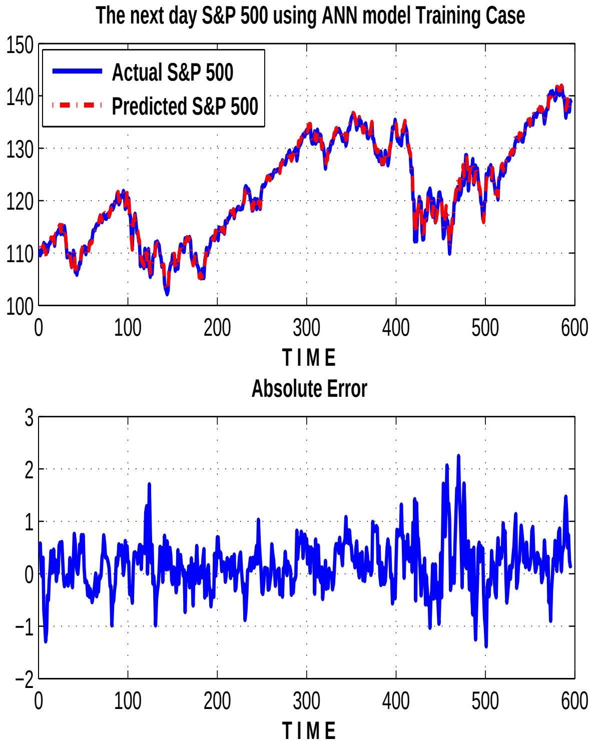

Figure 5.17: MLP Actual and Estimated S&P 500 Index values in Training Case

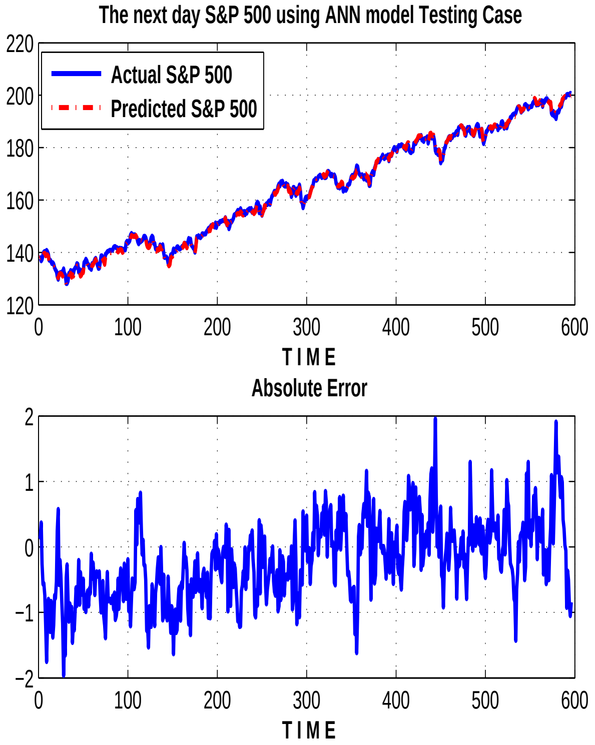

Figure 5.18: MLP Actual and Estimated S&P 500 Index values in Test Case

	MLP MODEL
	Cross-validation
MSE	0.6895
MAE	0.6655
RMSE	0.8304

Table 5.14: Evaluation Criteria for the developed MLP model cross validation

5.1.7 Genetic Programming model with selected attribute

This section show the result of GP model with 14 attribute selected by GA, equation 5.5 show the GP model,Figure 5.20 ,5.21 show the actual and estimated values and Table 5.15 show the evaluation of the model.

$$\begin{aligned}
ypred &= 0.3604 * SPY(t-1) + 1.267 * GE - 0.1505 * MSFT - 0.5017 * JNJ \\
&+ 0.02073 * IXIC - 0.04188 * SPY(t-3) + 63.91 * CTB6M + 1.533 * CTB5Y \\
&+ 0.1585 * XOM - 0.008367 * SPY(t-1) * CTB3M - 0.009367 * IXIC * CTB6M \\
&+ 0.04435 * SPY(t-1) * CTB1Y - 0.005372 * SPY(t-1) * XOM \\
&- 0.008367 * GE * XOM + 0.008367 * JNJ * XOM - 12.15 * CTB6M * CTB5Y \\
&- 0.01673 * CTB6M * XOM - 1.221 * CTB6M^2 * CTB5Y \\
&- 2.443 * CTB6M^2 + 11.4; \qquad (5.5)
\end{aligned}$$

	MGP MODEL	
	Training	Testing
VAF	99.558	99.256
MSE	0.41536	6.4018
MAE	0.4713	1.9753
RMSE	0.64448	2.5302

Table 5.15: Evaluation Criteria for the developed MGP model

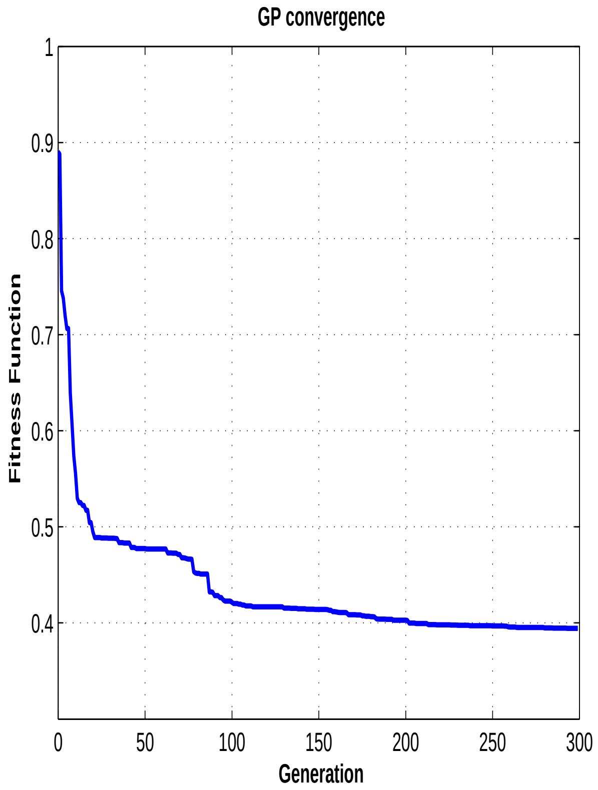

Figure 5.19: MGP Convergence

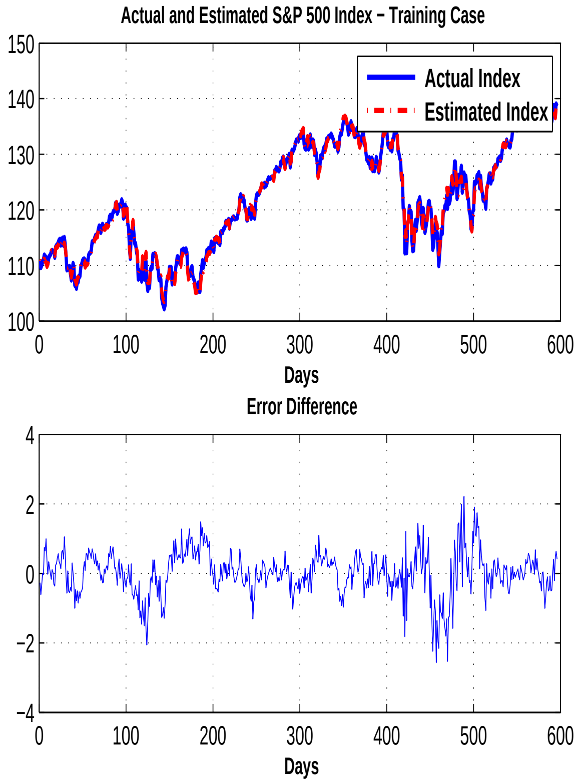

Figure 5.20: MGP Actual and Estimated S&P 500 Index values in Training Case

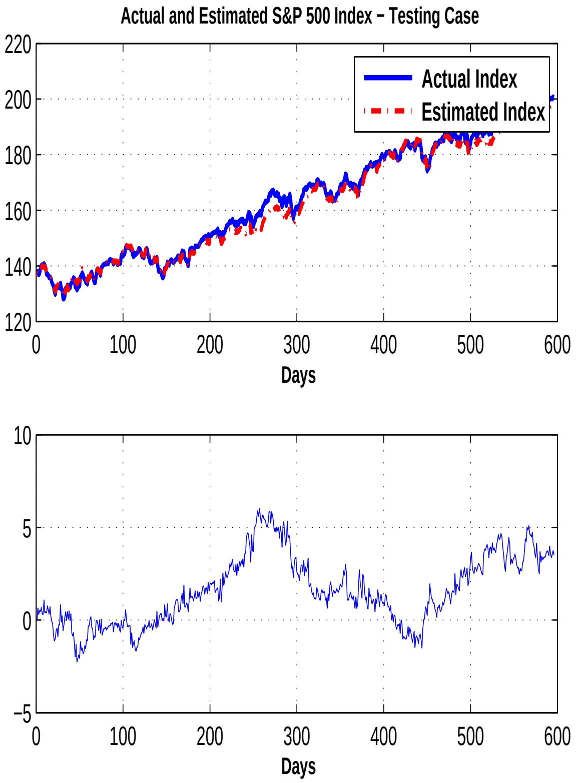

Figure 5.21: MGP Actual and Estimated S&P 500 Index values in Testing Case

5.1.8 Summary and Comparison

This section summaries the result of the the three models. Table 5.19 present the performance of the evaluation criteria with all attribute dataset. table 5.20 show the evaluation criteria for the three models with 14 attribute dataset. A comparison between MLR model used 27 feature and MLR model using 14 feature show in Table 5.16 show that the feature that selected by GA has a good result .
Table 5.17 and Table 5.18 is also show the comparison between the MLP models and MGP models.The GA feature selection enhance the models.

A comparison between the three proposed models for forecasting the S& P 500 is shown in Table 5.19 and 5.20. It was found that all models performing good with respect to the VAF evaluation criteria. we use 27 potential financial and economic factors these feature has impact and influence on stock movement the models provides good prediction . although these feature selected by GA.

Table 5.21 and 5.22 comparing the Correlation Coefficient (CC) of a forecast models

	MLR27		MLR14	
	Training	Testing	Training	Testing
VAF	99.8266	99.9400	99.6169	99.8885
MSE	0.1628	0.2542	0.3597	0.4725
MAE	0.3096	0.4017	0.532	2.8959
RMSE	0.4828	0.5042	0.7073	3.6286

Table 5.16: Comparing MLR27 with MLR14

	MLP27		MLP14	
	Training	Testing	Training	Testing
VAF	99.6895	98.7892	99.7521	99.9116
MSE	0.2223	1.0320	0.2644	0.4234
MAE	0.2774	0.390	0.3776	0.5303
RMSE	0.3564	0.4851	0.4962	0.649

Table 5.17: Comparing MLP27 with MLP14

	MGP27		MGP14	
	Training	Testing	Training	Testing
VAF	99.674	99.321	99.558	99.256
MSE	0.30626	6.1108	0.41536	6.4018
MAE	0.41564	2.0588	0.4713	1.9753
RMSE	0.58272	3.0235	0.64448	2.5302

Table 5.18: Comparing MGP27 with MGP14

	MLR		MLP		MGP	
	Training	Testing	Training	Testing	Training	Testing
VAF	99.8266	99.9400	99.6895	98.7892	99.674	99.321
MSE	0.1628	0.2542	0.2223	1.0320	0.30626	6.1108
MAE	0.3096	0.4017	0.2774	0.390	0.41564	2.0588
RMSE	0.4828	0.5042	0.3564	0.4851	0.58272	3.0235

Table 5.19: Evaluation Criteria for the developed models with all attribute

	MLR		MLP		MGP	
	Training	Testing	Training	Testing	Training	Testing
VAF	99.6169	99.8885	99.7521	99.9116	99.558	99.256
MSE	0.3597	0.4725	0.2644	0.4234	0.41536	6.4018
MAE	0.532	2.8959	0.3776	0.5303	0.4713	1.9753
RMSE	0.7073	3.6286	0.4962	0.649	0.64448	2.5302

Table 5.20: Evaluation Criteria for the developed models with 14 attribute

with a benchmark. The result show that the quotient is nearest to one; that mean the model is good or there is no difference between the new models and benchmark. as done in Equation 5.10.

$$VAF_q = \frac{VAF_f}{VAF_b} \tag{5.6}$$

	MLP		MGP	
	Training	Testing	Training	Testing
VAF_q	0.9986	0.9884	0.9984	0.9938

Table 5.21: Comparing Benchmark (MLR) with developed models

	MLP		MGP	
	Training	Testing	Training	Testing
VAF_q	1.0013	1.00023	0.9994	0.99366

Table 5.22: Comparing Benchmark (MLR) with developed models with 14 attribute

General remarks for the models:

- Cross-validation does not outperform the split-validation in all cases .
- All models are given high performance.

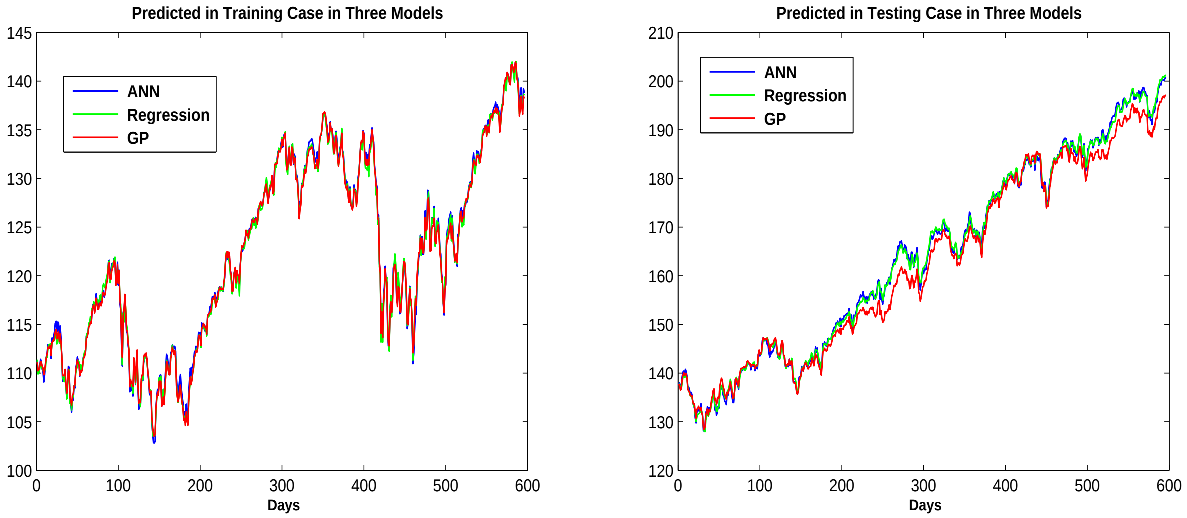

Figure 5.22: Estimated S&P 500 Index values in Three Models (a) Training Case (b) Testing Case

- GA enhance the model by decrease the complexity of the model and given high performance.
- The selected attribute given good performance but not better than all attribute.
- The GP model has better transparency capability.
- The GP model given simple equation to solve the prediction problem.

5.2 Solution 2: Weekly Data prediction

5.2.1 Multiple Linear Regression

In this section, we show the results produced for modelling the S&P500 stock index weekly data to predict next week price using regression model. The developed multiple regression model is given in

$$y = a_0 + a_1x_1 + a_2x_2 + \cdots + a_27x_27 + \epsilon$$

The values of the parameters a shall be estimated using the least square estimation (LSE) method to produce the optimal values of the parameters a_i. LSE is one of the oldest popular technique in statistics. The produced linear regression model can be presented as given in Table 5.23. The actual and Estimated S&P 500 index values based the MLR in both training and testing cases are shown in Figure 5.23. The scattered plot of the actual and predicted responses is shown in Figure 5.24. Evaluation criteria shown in Table 5.30 and 5.25.

$$\begin{aligned}
y_{MLR} &= -0.0234 * SPY(t-1) + 0.13 * SPY(t-2) + 0.021 * SPY(t-3) \\
&+ 0.021 * OIL - 0.021 * GOLD - GLD - 10.303 * CTB3M + 6.0031 * CTB6M \\
&+ 0.7738 * CTB1Y + 0.2779 * CTB5Y - 0.43916 * CTB10Y \\
&- 0.27754 * AAA + 0.12733 * BBB - 0.058638 * USD/JPY + 13.646 * USD/GBP \\
&+ 9.5224 * USD/CAD - 0.0003 * HTS + 0.24856 * FCHS - 0.0016 * FTSE100 \\
&+ 0 * V - 2.334 \times 10^{-9} * V + 0.16257 * XOM + 0.63767 * GE \\
&- 0.14301 * MSFT + 0.08 * PG + 0.074 * JNJ \\
&- 0.0002 * DJI + 0.026301 * IXIC + 6.9312 \qquad (5.7)
\end{aligned}$$

Table 5.23: A Regression Model with Inputs: $x_1, \ldots, x_{27}$ for weekly prediction

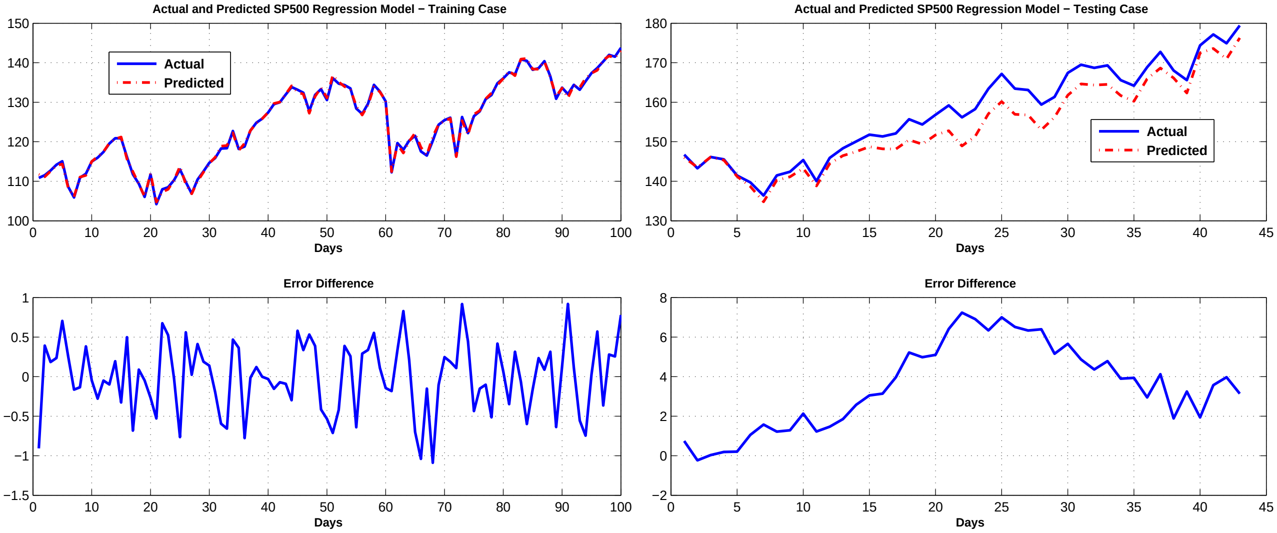

Figure 5.23: Regression: Actual and Estimated S&P 500 Index values (a) Training Case (b) Testing Case

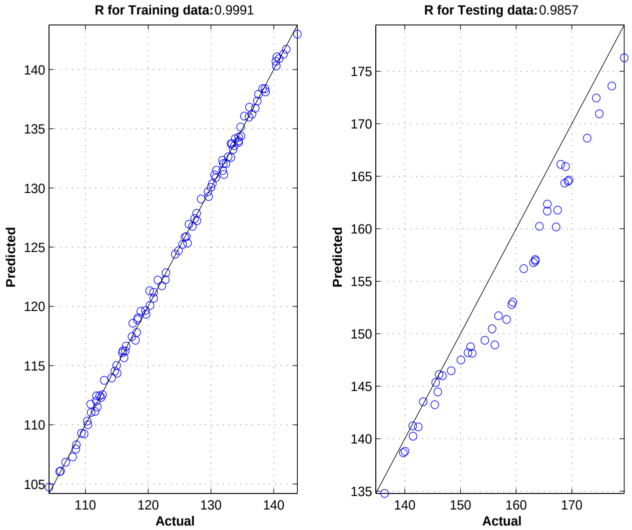

Figure 5.24: Regression Scattered Plot

Criterion	MLR	
	Training	Testing
VAF	99.94	99.91
MSE	0.3054	0.2312
MAE	0.4356	0.378
RMSE	0.5527	0.4809

Table 5.24: Evaluation Criteria for the MLR in split validation

	MLR
	Cross-validation
MSE	0.6938
MAE	0.6362
RMSE	0.833

Table 5.25: Evaluation Criteria for the MLR in cross validation

5.2.2 Multilayer Perceptron Neural Network model

The basic architecture of the MLP Network used to model the stock price prediction problem consists of three layers with single hidden layer. Thus input layer of our neural network model has 27 input nodes while the output layer consists of only one node that gives the predicted next day value. Empirically, we found that 30 neurons in the hidden layer achieved the best performance. The Backpropagation algorithm is used to train the MLP and update its weight.

All evaluation results are shown in Table 5.26 and 5.27. Figures 5.25 depict actual and predicted stock prices for training and testing cases of the developed MLP.

Criterion	MLP	
	Training	Testing
VAF	99.95	99.91
MSE	0.33849	0.26153
MAE	0.4457	0.393
RMSE	0.5818	0.5114

Table 5.26: Evaluation Criteria for the MLP in split validation

5.2.3 Multigene Genetic Programming model

GPTIPS toolbox Searson et al. (2010) adopted to develop the results. In GPTIPS, the initial population is constructed by creating individuals that contain randomly generated GP trees with between 1 and G_{max} genes. During the run, genes are acquired and deleted using a tree

	MLP
	Cross-validation
MSE	1.14319
MAE	0.8388
RMSE	1.0692

Table 5.27: Evaluation Criteria for the MLP in cross validation

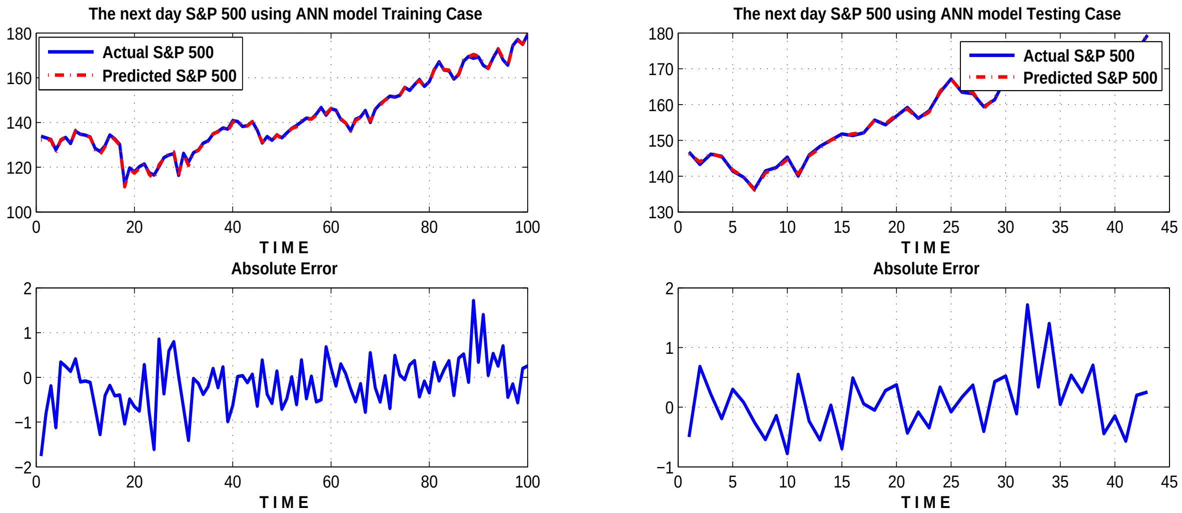

Figure 5.25: MLP Actual and Estimated S&P 500 Index values (a) Training Case (b) Testing Case

crossover operator called two point high level crossover. This allows the exchange of genes between individuals and it is used in addition to the ”standard” GP recombination operators.

Some parameters have to be defined by the user at the beginning of the evolutionary process. They include: population size, probability of crossover, mutation probability and the type of the selection mechanism. User has also to setup the maximum number of genes G_{max} where a model is allowed to have. The maximum tree depth D_{max} allows us to change the complexity of the evolved models. Restricting the tree depth helps evolving simple model but it may also reduce the performance of the evolved model.

A prior knowledge on the problem domain helps in designing a function set which could speed up the evolutionary process for model development. The adopted function set to develop the GP model is given as:

$$F = \{+, -, \times\}$$

Population size	30
Number of generation	300
Selection mechanism	Tournament
Tournament size	7
Max. tree depth	7
Probability of Crossover	0.85
Probability of Mutation	0.1
Number of inputs:	27
Max genes	3
Function set	+, -, ×
Constants range	[-10 10]

Table 5.28: GP Tuning Parameters

Figure 5.26 shows the actual and estimated stock market values based the developed GP model in both training and testing cases. In Figure 5.27, we show the convergence of GP over 500 generations along with the scattered plot. The best generated stock market Multigene GP model is given in Table 5.29. It can be clearly seen that the final model is a simple and

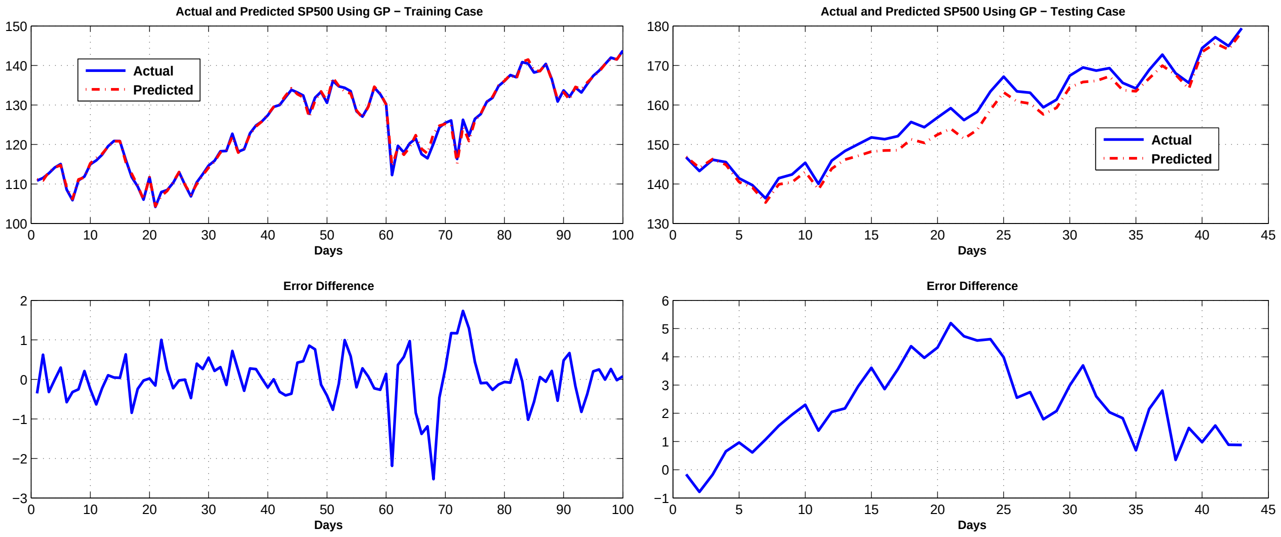

Figure 5.26: GP: Actual and Estimated S&P 500 Index values (a) Training Case (b) Testing Case

compact mathematical model which is easy to evaluate. The performance measurements for the model were computed and summarized in Table 5.30. In our experiments, we limited the max number of genes allowed to three. Thus, we can develop a simplified model structure. It was found that increasing the max number of genes will produce more complex models with high performance (i.e. less error).

$$
\begin{aligned}
y_{GP} &= 0.2206 * SPY(t-1) - 0.3617 * GOLD - GLD - 6.6983 * CTB3M \\
&+ 32.817 * USD/GBP + 0.8029 * USD/CAD + 0.382 * GE \\
&- 0.0556 * JNJ + 0.085 * IXIC - 0.1 * x_1 * USD/GBP \\
&+ 0.4542 * GOLD - GLD * USD/GBP - 0.1 * x_6 * GOLD - GLD \\
&+ 0.51 * AAA * USD/CAD + 0.3617 * USD/GBP * XOM + 0.1325 * USD/GBP * GE \\
&+ 0.302 * USD/GBP * JNJ - 0.1 * USD/GBP * IXIC - 0.1 * USD/GBP^2 * XOM \\
&+ 104.35 * USD/GBP^2 + 0.1 * GOLD - GLD * USD/GBP * USD/CAD \\
&+ 0.1 * AAA * USD/GBP * USD/CAD - 55.494 \qquad (5.8)
\end{aligned}
$$

Table 5.29: A GP model with Inputs: $x_1, \ldots, x_{27}$

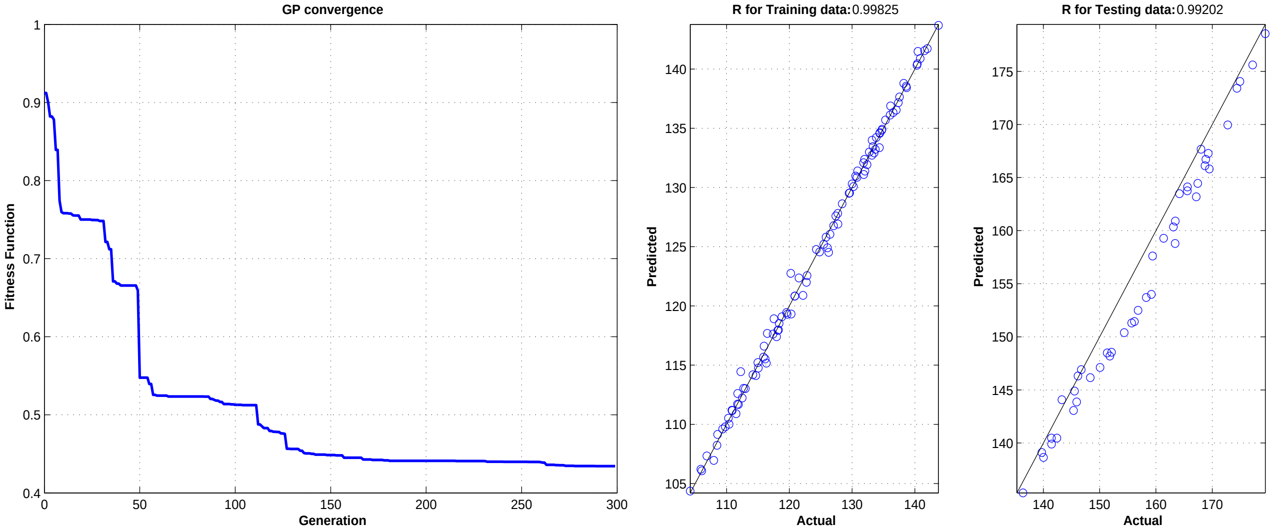

Figure 5.27: (a) Convergence of GP with various population sizes (b) GP Scattered Plot

Criterion	MGP	
	Training	Testing
VAF	99.825	99.202
MSE	0.33956	7.1684
MAE	0.43426	2.2948
RMSE	0.58272	2.6774

Table 5.30: Evaluation Criteria for the MGP model

5.2.4 Feature Selection

The purpose here is to obtain an optimal subset of features which produce the best possible results.as we describe in section 5.1.4.

The result obtain 10 feature given in Table 5.31

Selected attributes: 1,2,3,4,20,21,22,23,25,27 : 10

5.2.5 Multi linear Regression model

In this section we explain the regression model with 10 attribute that selected by GA. Equation 5.32 show the model of regression and figure 5.28 show the actual and estimated variable. Table 5.33 and 5.34 summarise the evaluation criteria.

selection attribute
SPYt-1
SPYt-2
SPYt-3
OIL
V
XOM
GE
MSFT
JNJ
IXIC

Table 5.31: The feature Selected

$$\begin{aligned} SPY &= 0.0626 * SPYt - 1 + 0.1325 * SPYt - 2 + 0.0319 * OIL + 0 * V \\ &+ 0.0788 * XOM + 0.8415 * GE + -0.1016 * MSFT + 0.135 * JNJ \\ &+ 0.0214 * IXIC + 15.3538 \end{aligned} \tag{5.9}$$

Table 5.32: A Regression Model with Inputs: $x_1, \ldots, x_{27}$

	MLR	
	Training	Testing
VAF	99.82	99.08
MSE	0.3937	9. 6137
MAE	0.5071	2.6769
RMSE	0.6275	3.1006

Table 5.33: Evaluation Criteria for the MLR with 10 feature selection in split validation

	MLR
	Cross-validation
MSE	0.8658
MAE	0.6854
RMSE	0.9305

Table 5.34: Evaluation Criteria for the MLR with 10 feature selection in cross validation

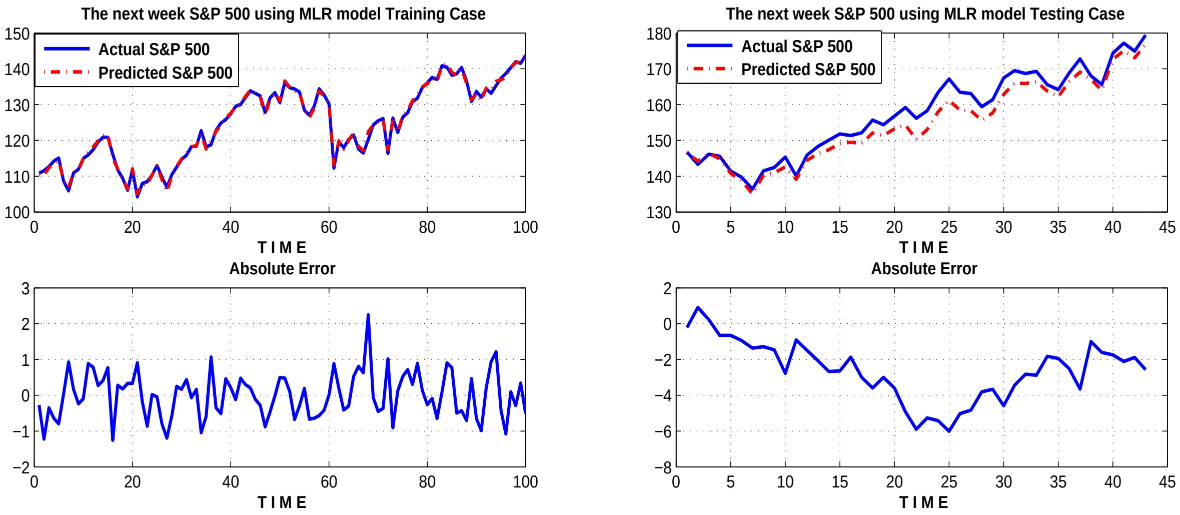

Figure 5.28: Regression: Actual and Estimated S&P 500 Index values (a) Training Case (b) Testing Case

5.2.6 Multilayer Perceptron Neural Network

In this section we explain the MLP model with 10 attribute that selected by GA. Figures 5.29 and show the actual and estimated variable. Table 5.35 summarise the evaluation criteria.

	MLP	
	Training	Testing
VAF	99.84	99.72
MSE	0.7202	1.5129
MAE	0.7057	0.997
RMSE	0.8487	1.23

Table 5.35: Evaluation Criteria for the MLP with 10 feature selection in split validation

	MLP
	Cross-validation
MSE	1.3108
MAE	0.8845
RMSE	1.1449

Table 5.36: Evaluation Criteria for the MLP with 10 feature selection in cross validation

5.2.7 Multigene Symbolic Regression Genetic Programming

In this section we explain the GP model with 10 attribute that selected by GA. Figures 5.30 and show the actual and estimated variable. Table 5.37 summarise the evaluation criteria.

5.2.8 Summary and Comparison

This section summaries the result of the the three models for weekly prediction . Table 5.41 present the performance of the evaluation criteria with all attribute dataset. table 5.42 show the evaluation criteria for the three models with 10 attribute dataset. A comparison between MLR model used 27 feature and MLR model using 10 feature show in Table 5.38 show that the feature that selected by GA has a good result .

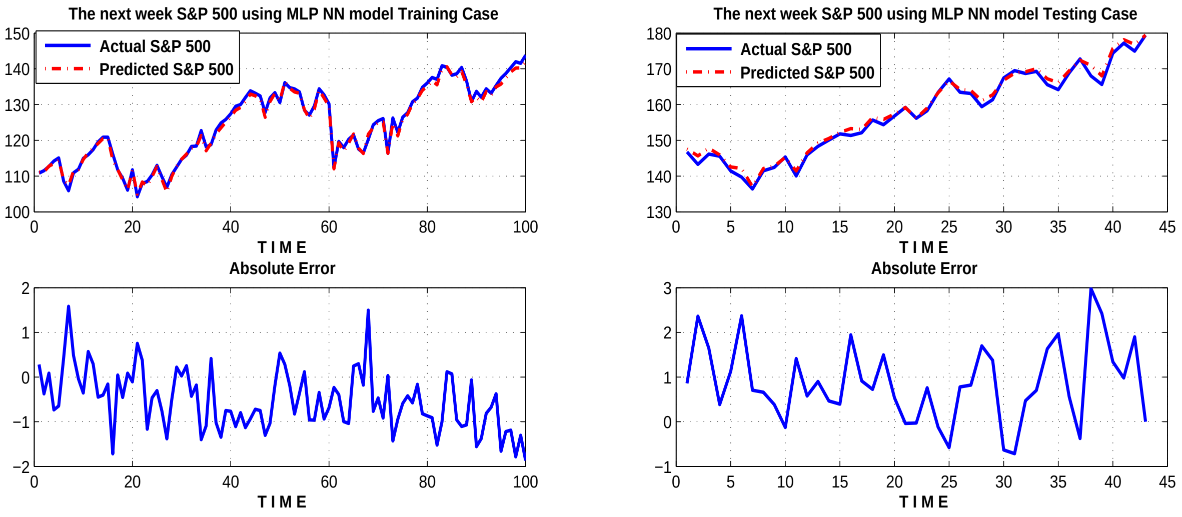

Figure 5.29: MLP: Actual and Estimated S&P 500 Index values (a) Training Case (b) Testing Case

	MGP	
	Training	Testing
VAF	99.827	98.945
MSE	0.38012	16.38
MAE	0.47399	3.4971
RMSE	0.61654	4.0477

Table 5.37: Evaluation Criteria for the MGP with 10 feature selection

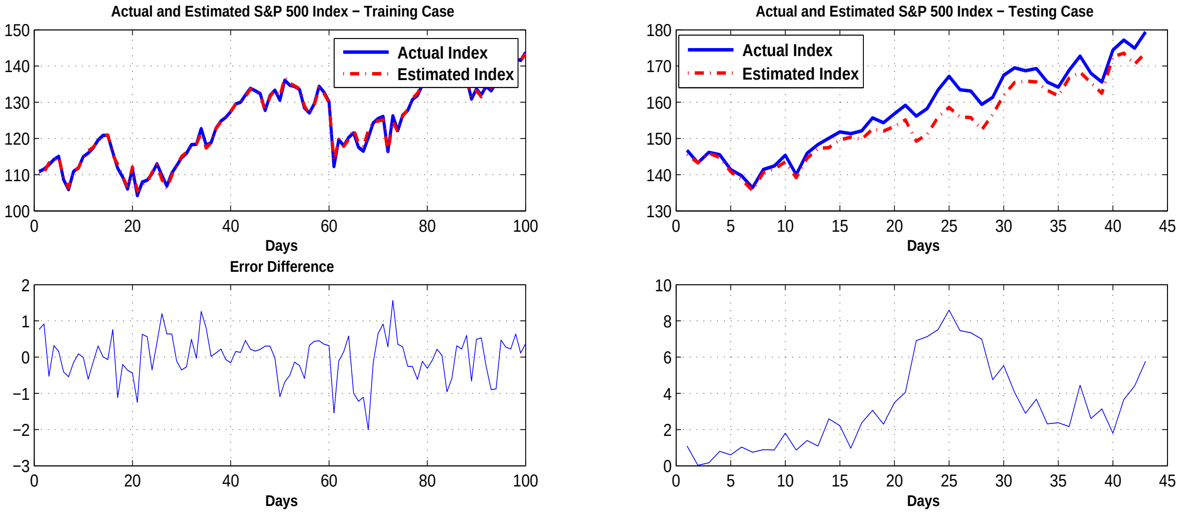

Figure 5.30: MGP: Actual and Estimated S&P 500 Index values (a) Training Case (b) Testing Case

Table 5.39 and Table 5.40 is also show the comparison in the MLP models and MGP models.The GA feature selection enhance the models.

	MLR27		MLR10	
	Training	Testing	Training	Testing
VAF	99.94	99.91	99.82	99.08
MSE	0.3054	0.2312	0.3937	9. 6137
MAE	0.4356	0.378	0.5071	2.6769
RMSE	0.5527	0.4809	0.6275	3.1006

Table 5.38: Comparing MLR27 with MLR10 in weekly prediction

	MLP27		MLP14	
	Training	Testing	Training	Testing
VAF	99.95	99.91	99.84	99.72
MSE	0.33849	0.26153	0.7202	1.5129
MAE	0.4457	0.393	0.7057	0.997
RMSE	0.5818	0.5114	0.8487	1.23

Table 5.39: Comparing MLP27 with MLP10 for weekly prediction

A comparison between the three proposed models for forecasting the S& P 500 is shown in Table 5.41 and 5.42. It was found that all models performing good with respect to the correlation coefficient evaluation criteria. we use 27 potential financial and economic factors these feature has impact and influence on stock movement the models provides good prediction . although these feature selected by GA.

Table 5.43 and 5.44 comparing the Variance-Accounted-For (VAF) of a forecast models with a benchmark. The result show that the quotient is nearest to one; that mean the model is good or there is no difference between the new models and benchmark. as done in Equation 5.10.

$$VAF_q = \frac{VAF_f}{VAF_b} \tag{5.10}$$

	MGP27		MGP10	
	Training	Testing	Training	Testing
VAF	99.825	99.202	99.827	98.945
MSE	0.33956	7.1684	0.38012	16.38
MAE	0.43426	2.2948	0.47399	3.4971
RMSE	0.58272	2.6774	0.61654	4.0477

Table 5.40: Comparing MGP27 with MGP10 for weekly prediction

	MLR		MLP		MGP	
	Training	Testing	Training	Testing	Training	Testing
VAF	99.94	99.91	99.95	99.91	99.825	99.202
MSE	0.3054	0.2312	0.33849	0.26153	0.33956	7.1684
MAE	0.4356	0.378	0.4457	0.393	0.43426	2.2948
RMSE	0.5527	0.4809	0.5818	0.5114	0.58272	2.6774

Table 5.41: Evaluation Criteria for the developed models with all attribute in weekly prediction

	MLR		MLP		MGP	
	Training	Testing	Training	Testing	Training	Testing
VAF	99.82	99.08	99.84	99.72	99.827	98.945
MSE	0.3937	9. 6137	0.7202	1.5129	0.38012	16.38
MAE	0.5071	2.6769	0.7057	0.997	0.47399	3.4971
RMSE	0.6275	3.1006	0.8487	1.23	0.61654	4.0477

Table 5.42: Evaluation Criteria for the developed models with 10 attribute for weekly prediction

	ANN		MGP	
	Training	Testing	Training	Testing
VAF_q	0.9986	0.9884	0.9984	0.9938

Table 5.43: Comparing Benchmark (MLR) with developed models

	ANN		MGP	
	Training	Testing	Training	Testing
VAF_q	1.0013	1.00023	0.9994	0.99366

Table 5.44: Comparing Benchmark (MLR) with developed models with 10 attribute

Chapter Six

Conclusions

Financial Time Series Forecasting is an important area of the knowledge and there are many applications in the real world. Accurate forecasting is an essential element for many management decisions. Stock market prediction is an important task in time series forecasting and there are several methods and techniques to find a good model that can be used to produce accurate forecasting the traditional techniques have your foundations in statistics.

Furthermore, the appropriate dataset used is making the models more accuracy. The most important model statistical methodology is the Autoregressive (AR) models. These methods present some obstacles and complexities to overcome. The major difficulty is to select the good model that can best adjustment for a specific dataset; usually many attempts must be performed until the best model must be found. Because of these difficulties, many researchers have been done several efforts to overcome these problems, such as Artificial Neural Network (ANN), Evolutionary Computation (EC) and in special Genetic Programming (GP) that have been provided good results in predicting stock market. In this research a new dataset based on S& P500, has been created by selecting features from 6 groups (27 influence features) including the delay of the closed (SPY) attribute in the S& P500. Dataset consists of 27 feature and 1192 days of data which cover five-year period starting from December 2009 to September 2014.

Developed and investigated the statistical multiple Linear Regression for forecasting the stock market indexing; and using the model as benchmark.

Developed and investigated the multilayer perceptron neural network for forecasting the stock market indexing.found that 20 neurons in the hidden layer achieved the best performance. The Back propagation algorithm is used to train the MLP and update its weight. The developed forecasting model were trained and tested based on a S&P 500 stock market data set, evaluated and tested based on different evaluation metrics and the VAF is 99.9400 ,99.6895 for training and test showing good model.

Developed and investigated A genetic programming enhance by Multigene GP model for forecasting the stock market indexing.The developed GP model provided good estimation and prediction capabilities in both training and testing cases. The results were validated using number of evaluation criteria.the VAF is 99.674 ,99.321 for training and test showing good model and simple mathematical model is developed .

Developed GA model as feature selection to select the feature that has impact factor in forecasting stock market. Used these feature with previous models to enhance the performance . evaluated and tested based on different evaluation metrics and compared to the statistical multiple Linear Regression model.

The knowledge gained is comprehensible and can enhance the decision making process.

Bibliography

Abhishek, K., A. Khairwa, T. Pratap, and S. Prakash (2012). A stock market prediction model using artificial neural network. pp. 1–5.

Aboueldahab, T. and M. Fakhreldin (2010). Stock market indices prediction via hybrid sigmoid diagonal recurrent neural networks and enhanced particle swarm optimization. *10*(1), 23–30.

Abraham, A., B. Nath, and P. Mahanti (2001). Hybrid intelligent systems for stock market analysis. pp. 337–345.

Agrawal, J., V. Chourasia, and A. Mittra (2013). State-of-the-art in stock prediction techniques. *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering 2*, 1360–1366.

Akintola, K., B. Alese, and A. Thompson (2011). Time series forecasting with neural network: A case study of stock prices of intercontinental bank nigeria. *International Journal of Research & Reviews in Applied Sciences 9*(3).

Aldin, M. M., H. D. Dehnavi, and S. Entezari (2012). Evaluating the employment of technical indicators in predicting stock price index variations using artificial neural networks (case study: Tehran stock exchange). *International Journal of Business and Management 7*(15), p25.

Ali, S. A., A. A. Hammad, M. Samhouri, and A. Al-Ghandoor (2011). Modeling stock market exchange prices using artificial neural network: A study of amman stock exchange. *EDITORIAL BOARD 5*(5), 439.

Allen, F. and R. Karjalainen (1999, February). Using genetic algorithms to find technical trading rules. *Journal of Financial Economics 51*(2), 245–271.

Atsalakis, G. S. and K. P. Valavanis (2009a). Forecasting stock market short-term trends using a neuro-fuzzy based methodology. *Expert Systems with Applications 36*(7), 10696–10707.

Atsalakis, G. S. and K. P. Valavanis (2009b). Forecasting stock market short-term trends using a neuro-fuzzy based methodology. *Expert Syst. Appl. 36*(7), 10696–10707.

Boyacioglu, M. A. and D. Avci (2010, December). An Adaptive Network-Based Fuzzy Inference System (ANFIS) for the prediction of stock market return: The case of the Istanbul Stock Exchange. *Expert Systems with Applications 37*(12), 7908–7912.

Chaigusin, S., C. Chirathamjaree, and J. Clayden (2008). Soft computing in the forecasting of the stock exchange of thailand (set). pp. 1277–1281.

Chaturvedi, D. K. (2008). *Introduction to Soft Computing*, Volume 103 of *Studies in Computational Intelligence*. Springer Berlin Heidelberg.

Chen, S., C. Tao, and W. He (2009). A new algorithm of neural network and prediction in china stock market. pp. 686–689.

Chen, Y., L. Peng, and A. Abraham (2006). Exchange rate forecasting using flexible neural trees. pp. 518–523.

Choudhry, R. and K. Garg (2008). A hybrid machine learning system for stock market forecasting. *World Academy of Science, Engineering and* , 315318.

Cubiles-de-la Vega, M.-D., R. Pino-Mejías, A. Pascual-Acosta, and J. Muñoz-García (2002). Building neural network forecasting models from time series arima models: A procedure and a comparative analysis. *Intelligent Data Analysis 6*(1), 53–65.

Davis, L. et al. (1991). *Handbook of genetic algorithms*, Volume 115. Van Nostrand Reinhold New York.

durofy.com. Machine learning: Introduction to the artificial neural network.

El-Telbany, M. E. (2005). The egyptian stock market return prediction: A genetic programming. *The International Journal of Artificial Intelligence and Machine Learning 5*.

Erkam Guresen, Gulgun Kayakutlu, T. U. D. (2011). Using artificial neural network models in stock market index prediction. *Expert Systems with Applications 38*, 10389–10397.

Fama, E. F. and K. R. French (1995). Size and book-to-market factors in earnings and returns. *The Journal of Finance 50*(1), 131–155.

Fatima, S. and G. Hussain (2008). Statistical models of kse100 index using hybrid financial systems. *Neurocomputing 71*(13), 2742–2746.

Fok, W. W., V. IAENG, W. Tam, and H. Ng (2008). Computational neural network for global stock indexes prediction. *2*.

Grosan, C. and A. Abraham (2006). Stock market modeling using genetic programming ensembles. pp. 131–146.

Guresen, E., G. Kayakutlu, and T. U. Daim (2011). Using artificial neural network models in stock market index prediction. *Expert Systems with Applications 38*(8), 10389–10397.

Hanias, M., P. Curtis, and J. Thalassinos (2007). Prediction with neural networks: The athens stock exchange price indicator. *European Journal of Economics, Finance and Administrative Sciences 9*, 21–27.

Harvey, A. C. and P. Todd (1983). Forecasting economic time series with structural and box-jenkins models: A case study. *Journal of Business & Economic Statistics 1*(4), 299–307.

Hassan, M. R. (2009). A combination of hidden markov model and fuzzy model for stock market forecasting. *Neurocomputing 72*(16), 3439–3446.

Hossain, A. and M. Nasser (2011). Comparison of the finite mixture of arma-garch, back propagation neural networks and support-vector machines in forecasting financial returns. *Journal of Applied Statistics 38*(3), 533–551.

Huang, S.-C. and T.-K. Wu (2008). Integrating ga-based time-scale feature extractions with svms for stock index forecasting. *Expert Systems with Applications 35*(4), 2080–2088.

Huang, W., Y. Nakamori, and S.-Y. Wang (2005). Forecasting stock market movement direction with support vector machine. *Computers & Operations Research 32*(10), 2513–2522.

Hui, A. (2003). Using genetic programming to perform time-series forecasting of stock prices. *Genetic Algorithms and Genetic Programming at Stanford*, 83–90.

Iba, H. and T. Sasaki (1999). Using genetic programming to predict financial data. In *Evolutionary Computation, 1999. CEC 99. Proceedings of the 1999 Congress on*, Volume 1, pp. –251 Vol. 1.

Jain, A. K., J. Mao, and K. Mohiuddin (1996). Artificial neural networks: A tutorial. *Computer 29*(3), 31–44.

Jegadeesh, N. and S. Titman (2001). Profitability of momentum strategies: An evaluation of alternative explanations. *The Journal of Finance 56*(2), 699–720.

Jia, G., Y. Chen, and P. Wu (2008). Menn method applications for stock market forecasting. In *Advances in Neural Networks-ISNN 2008*, pp. 30–39. Springer.

Kaboudan, M. A. (2000). Genetic programming prediction of stock prices. *Computational Economics 16*(3), 207–236.

Kara, Y., M. A. Boyacioglu, and Ö. K. Baykan (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the istanbul stock exchange. *Expert systems with Applications 38*(5), 5311–5319.

Karazmodeh, M., S. Nasiri, and S. M. Hashemi (2013). Stock price forecasting using support vector machines and improved particle swarm optimization. *Journal of Automation and Control Engineering 1*(2).

Khadka, M. S. (2012). Performance analysis of hybrid forecasting model in stock market forecasting. *International Journal of Managing Information Technology (IJMIT) 4*(3).

Kim, M.-J., S.-H. Min, and I. Han (2006). An evolutionary approach to the combination of multiple classifiers to predict a stock price index. *Expert Systems with Applications 31*(2), 241–247.

Koza, J. (1991). evolving a computer program to generate random numbers using the genetic programming paradigm. *Proceedings of the Fourth International Conference on Genetic Algorithms, Morgan Kaufmann, La Jolla,CA,*.

Koza, J. R. (1992). Genetic programming: on the programming of computers by means of natural selection. *1*.

Krollner, B. (2011). *Risk management in the Australian stockmarket using artificial neural networks*. Ph. D. thesis, School of Information Technology Bond University Risk Management in the Australian Stockmarket using Artificial Neural Networks Bjoern Krollner A dissertation submitted in total fulfilment of the requirements of the degree of Doctor of Philosophy for the School of Information Technology, Bond University.

Krose, B. and P. van der Smagt (2009). An introduction to neural networks. 1996. *Amsterdam, Amsterdam: The University of Amsterdam*.

Kumar, K. and J. D. Haynes (2003). Forecasting credit ratings using an ann and statistical techniques. *International journal of business studies 11*(1), 91–108.

Lahmiri, S. (2011). A comparative study of backpropagation algorithms in financial prediction. *International Journal of Computer Science, Engineering and Applications (IJCSEA) 1*(4).

Leonard, J. and M. Kramer (1990). Improvement of the backpropagation algorithm for training neural networks. *Computers & Chemical Engineering 14*(3), 337–341.

Liu, Y. and X. Yao (2001). Evolving neural networks for hang seng stock index forecast. *1*, 256–260.

Madziuk, J. and M. Jaruszewicz (2011). Neuro-genetic system for stock index prediction. *22*, 93123.

Majhi, R., G. Panda, B. Majhi, and G. Sahoo (2009). Efficient prediction of stock market indices using adaptive bacterial foraging optimization (abfo) and bfo based techniques. *Expert Systems with Applications 36*(6), 10097–10104.

Majhi, R., G. Panda, G. Sahoo, and A. Panda (2008). On the development of improved adaptive models for efficient prediction of stock indices using clonal-pso (cpso) and pso techniques. *International Journal of Business Forecasting and Marketing Intelligence 1*(1), 50–67.

Majhi, R., G. Panda, G. Sahoo, A. Panda, and A. Choubey (2008). Prediction of s&p 500 and djia stock indices using particle swarm optimization technique. pp. 1276–1282.

Melek Acar Boyacioglu, D. A. (2010). An adaptive network-based fuzzy inference system (anfis) for the prediction of stock market return: The case of the istanbul stock exchange. *Expert Systems with Applications: An International Journal 37*, 7908–7912.

Menkhoff, L. and M. P. Taylor (2007). The obstinate passion of foreign exchange professionals: technical analysis. *Journal of Economic Literature*, 936–972.

Mitchell, M. (1998). *An introduction to genetic algorithms*. MIT press.

Mohammadi, H. and L. Su (2010). International evidence on crude oil price dynamics: Applications of arima-garch models. *Energy Economics 32*(5), 1001–1008.

Mori, S., K. Hirasawa, and J. Hu (2005). The stock price prediction and sell-buy strategy model by genetic network programming. *IEEJ Transactions on Electronics, Information and Systems 125*, 631–636.

Neelima Budhani, D. C. K. J. (2012). Application of neural network in analysis of stock market prediction. *International Journal of Computer Science and Engineering Technology (IJCSET) 3*(4).

Neely, C. J. and P. A. Weller (2011). Technical analysis in the foreign exchange market. *Federal Reserve Bank of St. Louis Working Paper No*.

Negnevitsky, M. (2005). *Artificial intelligence: a guide to intelligent systems*. Pearson Education.

Nguyen, P. (2003). Fundamental analysis and stock returns: Japan 1993-2003. *WBP Financial Integrator*, 193–210.

Niaki, S. T. A. and S. Hoseinzade (2013). Forecasting s&p 500 index using artificial neural networks and design of experiments. *Journal of Industrial Engineering International 9*(1), 1–9.

Nigrin, A. (1993). *Neural networks for pattern recognition*. MIT Press.

Olatunji, S. O., M. Al-Ahmadi, M. Elshafei, and Y. A. Fallatah (2011). Saudi arabia stock prices forecasting using artificial neural networks. pp. 81–86.

Ou, J. A. and S. H. Penman (1989). Financial statement analysis and the prediction of stock returns. *Journal of accounting and economics 11*(4), 295–329.

Palit, A. K. and D. Popovic (2005). *Computational Intelligence in Time Series Forecasting: Theory and Engineering Applications (Advances in Industrial Control)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Piotroski, J. D. (2000). Value investing: The use of historical financial statement information to separate winners from losers. *Journal of Accounting Research*, 1–41.

Preethi, G. and B. Santhi (2012a). Stock market forecasting techniques: A survey. *Journal of Theoretical & Applied Information Technology 46*(1).

Preethi, G. and B. Santhi (2012b). Stock market forecasting techniques: A survey. *Journal of Theoretical & Applied Information Technology 46*(1).

Pring, M. (2002). *Technical Analysis Explained: The Successful Investor's Guide to Spotting Investment Trends and Turning Points*. McGraw-Hill Education.

Quah, T.-S. (2007). Using neural network for djia stock selection. *Engineering Letters 15*(1), 126–133.

Rahamneh, Z., M. Reyalat, A. Sheta, and S. Aljahdali (2010). Forecasting stock exchange using soft computing techniques. pp. 1–5.

Rajabioun, R. and A. Rahimi-Kian (2008). A genetic programming based stock price predictor together with mean-variance based sell/buy actions. In *Proceedings of the World Congress on Engineering*, Volume 2. Citeseer.

Refat, S., M. El-Telbany, H. Hefny, and A. Bahnasawi (2003). Discovering the classification rules for egyptian stock market using genetic programming. In *Circuits and Systems, 2003 IEEE 46th Midwest Symposium on*, Volume 2, pp. 952–955. IEEE.

Rezaiedolatabadi, H., S. Sayadi, A. Hosseini, M. Forghani, M. Shokhmgar, et al. (2013). Modeling and forecasting stock prices using an artificial neural network and imperialist competitive algorithm. *International Journal of Academic Research in Accounting, Finance and Management Sciences 3*(1), 296–302.

Searson, D. P., D. E. Leahy, and M. J. Willis (2010). Gptips: an open source genetic programming toolbox for multigene symbolic regression. In *Proceedings of the International multiconference of engineers and computer scientists*, Volume 1. Citeseer.

Sermpinis, G., K. Theofilatos, A. Karathanasopoulos, E. F. Georgopoulos, and C. Dunis (2013). Forecasting foreign exchange rates with adaptive neural networks using radial-basis functions and particle swarm optimization. *European Journal of Operational Research 225*(3), 528–540.

Setty, D. V., T. Rangaswamy, and K. Subramanya (2010). A review on data mining applications to the performance of stock marketing. *International Journal of Computer Applications 1*(3), 33–43.

Sheta, A., H. Faris, and M. Alkasassbeh (2013). A genetic programming model for s& p500 stock market prediction. *International Journal of Control and Automation 6*, 303–314.

Soni, S. (2011). Applications of anns in stock market prediction: A survey. *International Journal of Computer Science and Engineering Technology 2*(3), 71–83.

TSE, Y.-K. and W.-S. CHAN (2010). The lead–lag relation between the s&p500 spot and futures markets: An intraday-data analysis using a threshold regression model. *Japanese Economic Review 61*(1), 133–144.

Vafaie, H. and K. De Jong (1992). Genetic algorithms as a tool for feature selection in machine learning. In *Tools with Artificial Intelligence, 1992. TAI'92, Proceedings., Fourth International Conference on*, pp. 200–203. IEEE.

Walczak, S. (2001). An empirical analysis of data requirements for financial forecasting with neural networks. *Journal of management information systems 17*(4), 203–222.

Wijaya, Y. B., S. Kom, and T. A. Napitupulu (2010). Stock price prediction: Comparison of arima and artificial neural network methods-an indonesia stock's case. pp. 176–179.

Wu, Q., Y. Chen, and Q. Z. Liu (2008). Ensemble model of intelligent paradigms for stock market forecasting. In *First International Workshop on Knowledge Discovery and Data Mining (WKDD 2008)*, pp. 205–208.

Yen, M.-f., T.-n. Chou, H.-c. Li, and Y.-y. Ho (2007). Using neural network and genetic programming techniques to forecast inter-commodity spreads. pp. 192–192.

Yen, S. M.-F. and Y.-L. Hsu (2010). Profitability of technical analysis in financial and commodity futures marketsa reality check. *Decision Support Systems 50*(1), 128–139.

Yu, L., S. Wang, and K. K. Lai (2009). A neural-network-based nonlinear metamodeling approach to financial time series forecasting. *Applied Soft Computing 9*(2), 563–574.

Zainab Al Rahamneh, Mohammad Reyalat, A. F. S. S. A. (2010). Forecasting stock exchange using soft computing techniques. pp. 1–5.

Zhang, X., Y. Chen, and J. Y. Yang (2007). Stock index forecasting using pso based selective neural network ensemble. In *IC-AI*, pp. 260–264.

Zhang, Y. and L. Wu (2009). Stock market prediction of s&p 500 via combination of improved bco approach and bp neural network. *Expert Syst. Appl. 36*(5), 8849–8854.

Zhuo, L., J. Zheng, X. Li, F. Wang, B. Ai, and J. Qian (2008). A genetic algorithm based wrapper feature selection method for classification of hyperspectral images using support vector machine. pp. 71471J–71471J. International Society for Optics and Photonics.

Zimmermann, H.-G., R. Grothmann, and C. Stolz (2004). Multi-agent market modeling: A survey. *Zeitreihenanalyse in der empirischen Wirtschaftsforschung: Festschrift für Winfried Stier zum 65. Geburtstag*, 197.