Article

A Robust Regression-Based Stock Exchange Forecasting and Determination of Correlation between Stock Markets

Umair Khan [1], Farhan Aadil [1] , Mustansar Ali Ghazanfar [2], Salabat Khan [1], Noura Metawa [3,4], Khan Muhammad [5], Irfan Mehmood [6,*] and Yunyoung Nam [7,*]

[1] Department of Computer Science, COMSATS University Islamabad, Attock Campus, Punjab 43600, Pakistan; umair_khan@ciit-attock.edu.pk (U.K.); farhan.aadil@ciit-attock.edu.pk (F.A.); salabat.khan@ciit-attock.edu.pk (S.K.)
[2] Department of Software Engineering, U.E.T Taxila, Punjab 47080, Pakistan; mustansar.ali@uettaxila.edu.pk
[3] Anderson College of Business, Regis University, Denver, CO 80221-1099, USA; nmetawa@regis.edu
[4] Faculty of Commerce, Mansoura University, Mansoura 1101, Egypt
[5] Intelligent Media Laboratory, Digital Contents Research Institute, Sejong University, Seoul 143-747, Korea; khan.muhammad@ieee.org
[6] Department of Software, Sejong University, Seoul 143-747, Korea
[7] Department of Computer Science and Engineering, Soonchunhyang University, Asan 31538, Korea
* Correspondence: irfanmehmood@ieee.org (I.M.); ynam@sch.ac.kr (Y.N.); Tel.: +82-010-9222-8794 (I.M.)

check for updates

Abstract: Knowledge-based decision support systems for financial management are an important part of investment plans. Investors are avoiding investing in traditional investment areas such as banks due to low return on investment. The stock exchange is one of the major areas for investment presently. Various non-linear and complex factors affect the stock exchange. A robust stock exchange forecasting system remains an important need. From this line of research, we evaluate the performance of a regression-based model to check the robustness over large datasets. We also evaluate the effect of top stock exchange markets on each other. We evaluate our proposed model on the top 4 stock exchanges—New York, London, NASDAQ and Karachi stock exchange. We also evaluate our model on the top 3 companies—Apple, Microsoft, and Google. A huge (Big Data) historical data is gathered from Yahoo finance consisting of 20 years. Such huge data creates a Big Data problem. The performance of our system is evaluated on a 1-step, 6-step, and 12-step forecast. The experiments show that the proposed system produces excellent results. The results are presented in terms of Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

Keywords: financial management; stock exchange prediction; regression; forecasting; correlation

1. Introduction

The stock market is a strong indication for economic conditions of a country. Stock exchange provides a neutral ground for brokers and companies to invest. People can invest their money and can get a huge profit if they invest sensibly. Stock markets provide a better platform to people as compared to traditional banking investments. Stock investments return more profit than bank deposits and bonds. However, the higher profits come with higher risks involved with stock exchange rates. The stock exchange is associated with non-linear and highly fluctuating factors [1]. These factors include the economic conditions of a country, public sentiment and political conditions of a country. These factors cause stock rates to fluctuate after short time intervals. For this reason, investors and brokers purchase and sell stocks within the short time interval. Predicting stock exchange prices by

considering all dynamic factors is an important part of the business investment plan. Many researchers have explored time series analysis, machine learning methods, and technical analysis. Therefore, to assist investors by providing stock price prediction by effectively using available huge Big Data information, remains a key research area [2].

Stock exchange prediction is meant to reduce risk and provide better investment plans. These stock exchange prediction and forecasting methods are categorized into two groups, namely computationally intelligent (AI) based methods and statistical methods. The first category includes Adaptive-Network-based Fuzzy Inference Systems (ANFIS) [3], autoregressive conditional heteroskedasticity (ARCH) [4], AutoRegressive Integrated Moving Average (ARIMA) [5] and Generalized Autoregressive Conditional Heteroskedasticity (GARCH) [6]. These methods work on a strong assumption that data is linearly distributed. However, uncertainty and data complexity make it difficult to create a model based on the strict linear distribution of data. There are some factors such as investor social network, policy changing and economic factors. However, internal rules for stock exchange data can be represented by historical data. The second category suggests artificial intelligence based prediction methods can learn the internal rules for stock exchange data. These methods can predict the results without considering any strict data assumption because of their non-linear underlying capabilities. Various machine learning classifiers have been used to predict stock exchange prices. These classifiers include regression [3], Support Vector Machines (SVM) and Neural Networks [7].

Stock market time series forecasting is an interesting and open research area. Computationally, intelligent artificially intelligent algorithms are now mostly used to forecast time series. However, a highly efficient stock exchange prediction model is yet to be designed. There is another limitation in existing work, namely that they do not consider the effect of top international stock markets on each other. The rise or fall in top international stock markets impacts on other stock exchanges. Usually, the rise or fall in an international stock market is due to some external factors. So, the stock exchange prediction depends upon local factors and also on these international stock exchange markets. The robustness of prediction models remains an open research area.

In this paper, we evaluate the robustness of a simple efficient regression-based stock exchange prediction model. Regression-based models are more flexible and computationally efficient as compared to statistical methods. A linear regression-based prediction approach is used to predict stock exchange indices and companies. We have done time series analysis for 4 stock exchanges—New York, London, NASDAQ, and Karachi stock exchange to evaluate the effectiveness of our regression-based model using Big Data covering 20 years. We have also evaluated our model to forecast time series analysis for 3 companies—Google, Microsoft, and Apple. We also calculate the effect of one stock exchange on another using a correlation factor. The results show that the proposed model provides very good results. In general, this paper offers the following contributions:

1.　We propose and evaluate the robustness of regression-based time series analysis and forecasting;
2.　We forecast the future values for 4 stock exchanges and 3 international companies;
3.　We calculate the correlation between 4 stock exchanges for rising and fall of stock indices.

The remaining paper organization is as follows: related work is presented in Sections 2–4 present methodology and results, respectively, followed by a conclusion.

2. Related Work

To analyze the stock market, various techniques have been used. These techniques fall under the categories of artificial intelligence systems, hybrid of artificial systems with the trading rules and machine learning techniques. Work done in each of the categories is thoroughly discussed in the section below.

2.1. Artificial Intelligence Systems

To estimate the stock market indices, various artificial intelligence techniques have been devised over recent years. Kodogiannis and Lolis in 2002 for the first time proposed the Artificial Neural Network (ANN) to predict stock markets [8]. Later, Thirunavukarasu et al. [9] in 2009 also proposed a stock market prediction system using ANN. In 2014, Xi, Muzhou et al. proposed an ANN to predict the stock market indices [10].

Support Vector Machines has also been of high interest in the research community and it was first introduced for the prediction of the stock market in 2009 by Zhang and Shen [11]. It was utilized by Wen, Yang, Song and Jia in 2010 as the artificial intelligence technique for stock market prediction [12]. Support Vector Machine was utilized by, Lin Guo and Hu in 2013 [13] and Yu, Chen and Zhang in 2014 [14] respectively. Recently in the year 2016, it was proposed by Gong et al. to estimate the stock market [4].

In the year 2002 rough set theory was first utilized by Wang and Wang for the stock market estimation and prediction [15]. Later Wang in 2003 also utilized it in their work [16]. It was then proposed in the methodology by Nair et al. in 2010 [17].

In the year 2007 Ives and Scandol proposed the utilization of Bayesian Analysis [18] which was then used in the methodology by Su and Peterman in 2012 [19], Ticknor in 2013 [20], and later in 2015 it was utilized as part of the methodology by Miao, Wang and Xu [21], Wang et al. [22], and Peng et al. [5].

Other artificial intelligence techniques such as K-Nearest Neighbors (KNN) was proposed by Li, Sun and Sun in 2009 [23] which was later-on utilized by Teixeira and De Oliveira in 2010 [24]. Techniques such as Particle Swarm Optimization (PSO) has been developed by Fu-Yuan in 2008 [25] and Shen, Zhang and Ma in 2009 [11]. Sorensen et al. in the year of 2000 [26], Wu, Lin and Lin in 2006 [27] and Hu, Feng et al. in 2015 [28] proposed Decision Tree in their methodology. The use of evolutionary learning algorithms such as the Genetic Algorithm has also been seen in the work of Hassan et al. in 2007 [29], followed by Huang and Wu in 2008 [30], and Rahman et al. in 2015 [6].

2.2. Artificial Intelligence Systems with Trading Rules

For most of the time, it has been seen that the artificial intelligent systems are accompanied by trading rules in search of development of autonomous and smart decision support systems. Following the chronological distribution of literature in 2015, Cervell'o-Royo et al. proposed a trading rule that was not only beneficial, but also adaptive to risk [31]. The rule was based on technical analysis and a combinational pattern, which provides the information about selling and buying, amount of profit earned and the maximum loss that can be tolerated. Kim and Enke in 2016 devised a heuristic based change trading system (RTCS) compromised of various historical values generated using rough set analysis [32]. The proposed methodology is developed to cater to diverse market conditions. Podsiadlo and Rybinski in 2016 set out to experimentally determine the feasibility of rough sets to build productive prediction models [33]. In 2016 Chiang et al. proposed a dynamic stock prediction system using Predicted Square Error (PSE) and neural network [34]. The proposed method incorporated the shortcomings due to individual application of ANN.

2.3. Artificial Intelligence Systems with Artificial Neural Network

In recent years, hybrid implementation of artificial intelligence systems with ANN has been a trend among the research community. The reason for utilizing ANN has been due to computational complexity introduced by the large dimensionality of neurons. Zhong and Enke in 2017 applied the principal component analysis (PCA) along with its different variants such as Fast Robust Principal Component Analysis (FRPCA) and Kernel Principal Component Analysis (KPCA) for the simplification and re-arrangement of data [3]. The reformed data is then classified to predict the daily market returns using ANN. Gocken et al. in 2016 [7] devised a combination of genetic algorithm (GA) and ANN as a hybrid model to improve the stock market estimation. Majhi et al. in 2009 [35] proposed a neural

network variant for prediction of Dow Jones Industrial Average (DJIA) and Standard & Poor's (S&P) 500. It was concluded that the Functional Link Artificial Neural Network (FLANN) is a compatible model with other ANN models requiring less time during testing and training phase.

Considering the FLANN, Chakravarty and Dash [36] in 2012 developed a system to predict stock prices for DIJA, Bombay stock market and S&P 500. The results show that the fuzzy neural network based system produced better results compared to other systems. Dash and Bisoi in 2014 [37] proposed a hybrid approach based on the search optimization technique. This hybrid approach used a single layer neural network.

2.4. Artificial Intelligence Systems with Support Vector Machines

Hybrid systems with ANN are quite successful in estimating the stock market; however, they seem to undergo limitations of over-fitting, local maxima, and convergence problem. To handle such issues, SVMs are utilized by researchers in a hybrid with other techniques for stock market indices estimation. The application of SVM not only minimizes the likelihood of over-fitting but also provides a globalized solution. In 2012 Huang proposed a hybrid methodology using both GA and SVR for stock prediction [38]. Genetic Algorithm (GA) is mainly used for parameters optimization of the model and to perform feature selection to achieve optimal parameters as an input to the SVR model. The use of GA for feature selection is vital and helps to significantly outperform the benchmark schemes. Liu and Wang in 2013 [39] utilized a combinational model of SVM and Decision Trees (DT) to forecast the stock prediction aiming to achieve an increase in precision, recall, and F-One rate. The proposed methodology was tested against techniques such as Bootstrap-SVM, Bootstrap-DT, and Back Propagation Neural Network (BPNN).

In 2015 Nayak et al. [40] proposed a hybrid framework utilizing SVM with KNN. The proposed methodology was used to predict the Indian stock exchange market. SVM was utilized to predict future or loss. It also estimated the stock value over a time for one day, week and month. The model performed well for high dimensional feature vector and handled the error and the performance of the classification methods. The SVM-KNN model outperformed the mentioned models by removing the need to tune multiple parameters for ANN and fuzzy-based model.

3. Proposed Methodology

The proposed methodology is step-wise explained in the sections below.

3.1. Data Collection

Historical data for stock exchange indices and different companies can be fetched from Yahoo finance and Google finance. Yahoo finance provides the facility to download this historical data into. between any two dates. To evaluate our proposed system, we have trained and tested our technique over the publically available stock exchange dataset from Yahoo finance https://finance.yahoo.com/. The site hosts repositories of multiple stock exchanges such as Karachi, London, and New York Stock Exchange. Stock market data from multiple top-ranked technology companies such as Microsoft, Google and Apple have also been utilized to test our proposed system. The collected data includes weekly stock market trends over a time of 20 years from 10 July 1998 to 10 July 2018, which is Big Data. As we also want to study the effect of the dependency between stock markets, therefore, we have used the same dates to download the dataset for each stock market. The attributes for Yahoo finance dataset are given in Table 1 and details of historical data is presented in Table 2.

3.2. Data Pre-Processing

To avoid spurious regression time series data, such as stock prices, these need to be pre-processed to check for stationarity data. Most of the forecasting methods process the data with an assumption of data being stationary. A stationary time series is the one whose statistical properties such as mean and standard deviation does not change with time. Time series data such as stock

prices are checked for stationarity using unit root tests such as Augmented Dickey–Fuller (ADF) test. Collected stock market data is tested using the ADF test to find the unit root. The results from the ADF test determine whether to accept or reject the null hypothesis of data being non-stationary or if it has unit root based on the significance value *p*. Value of *p* less than 5% leads to the rejection of the null hypothesis.

Table 1. Attributes of Yahoo Finance dataset.

Feature	Description
Date	Corresponding Date for stock values
Open	Opening price of a stock on a particular day
High	Highest selling stock value for a day
Low	The lowest value of the selling price of a stock on a given day
Close	Contains closing value of a stock on a given day
Volume	The number of shares traded or bought on a given day
Adjusting Close	The closing price of a stock after paying dividends to the investors

Table 2. Training data information.

Name	Historical Data	
	From	To
NASDAQ stock exchange	7 October 1998	7 October 2018
New York stock exchange	7 October 1998	7 October 2018
London stock exchange	7 October 1998	7 October 2018
Karachi stock exchange	7 October 1998	7 October 2018
Companies data		
Microsoft	7 October 1998	7 October 2018
Apple	7 October 1998	7 October 2018
Google	7 October 2004	7 October 2018 (max available data)

3.3. Linear Regression

Trends in the stock market can be estimated under different regression models given below [41], linear regression models, neural network based model and SVM based regression.

Among the given models, linear regression is utilized due to its simplicity and robustness. In this methodology, we have considered modeling between a single dependent, and multiple independent variables. The regression model that deals with multiple variables is known as multiple linear regression model [1]. The multiple linear regression is a generalization of simple linear regression in a couple of ways. Multiple linear regression allows the dependence of multiple explanatory variables rather than one and allows for having multiple shapes rather than a single straight line.

Let *y* represents the dependent variable that is in a linear relationship with the *k* independent variables $X_1, X_2, X_3 \ldots X_k$ through parameters $\beta_1, \beta_2, \beta_3 \ldots \beta_k$ and is given as,

$$y = X_1\beta_1 + X_2\beta_2 \ldots X_k\beta_k + \varepsilon \tag{1}$$

where the parameters $\beta_1, \beta_2, \beta_3 \ldots \beta_k$ are the regression coefficients which are having an association with $X_1, X_2, X_3 \ldots X_k$ respectively and ε represents the random error component depicting the difference between the observed and fitted linear relationship.

The jth regression coefficient given as β_j shows the anticipated change in *y* per unit change in j^{th} independent variable X_j. Assuming $E(\varepsilon) = 0$,

$$\beta_j = \frac{\partial E(y)}{\partial X_j} \tag{2}$$

3.4. Stock Exchange Interdependency

Stock exchange interdependency is another research problem addressed in this paper. Stock markets affect one another in multiple ways for multiple reasons, majorly due to the effect of currencies on stock markets, similar listed products in different stock markets, dependencies of economies on one another etc. The objective is to find the correlation between different stock exchange markets. The effect of international markets on each other is evaluated using Pearson correlation and results are shown in tabular as well as graphical form. The correlation between two-time series can be calculated as:

$$correlation = \frac{N \sum xy - (\sum x)(\sum y)}{\sqrt{\left[N \sum x^2 - (\sum(x)^2 \right] \left[N \sum y^2 - (\sum(y)^2 \right]}} \tag{3}$$

Here N, $\sum xy$, $\sum x$, $\sum y$, $\sum x^2$, $\sum y^2$ represents the number of pairs of scores, the sum of the products of paired scores, the sum of the x scores, the sum of the y scores, the sum of the squared x scores and the sum of squared y scores respectively.

Similarly, autocorrelation is a method to determine the correlation between the successive values in the same data to calculate the randomness within the data. In a time series data, autocorrelation is calculated for different lags estimating the data dependency between the instances separated by the respective lag. The value of autocorrelation lies between +1 and −1 where the extremes represent a strong correlation between the values of the dataset. To calculate the randomness of the data autocorrelation is calculated for lag set as 1, i.e., successive values in the dataset.

The autocorrelation coefficient for N observations is calculated as,

$$r_1 = \frac{\sum_{t=1}^{N-1} (x_t - \overline{x}_{(1)}) \left(x_{t+1} - \overline{x}_{(2)} \right)}{\left[\sum_{t=1}^{N-1} (x_t - \overline{x}_{(1)})^2 \right]^{1/2} \left[\sum_{t=1}^{N-1} (x_{t+1} - \overline{x}_{(2)})^2 \right]^{1/2}} \tag{4}$$

The values $\overline{x}_{(1)}$ and $\overline{x}_{(2)}$ are the mean values of the first $N - 1$ and last $N - 1$ observations respectively.

Autocorrelation coefficient for the multiple stock market data of all companies was calculated using the above equation.

4. Experimental Methodology:

4.1. Data Description, Preparation, and Multi-Step Prediction

Stock exchange data acquired from the Yahoo finance provided the stock prices from different stock exchanges such as Karachi Stock Exchange (KSE), London Stock Exchange (LSE), New York Stock Exchange (NYSE) and American Stock Exchange (NASDAQ). They provide stock prices of multiple companies such as Microsoft (MSFT), Apple, and, Google characterized by the opening, closing, highest, and lowest values of the stock along with their number of share trade during the day. These characteristic attributes serve as multiple independent variables which are used to predict and forecast the closing values of stock as the dependent variable in the stock market. The dataset is comprised of historical stock market data for more than 20 years. The dataset is publicly available and covers the stock market trends from regions of Asia, Europe and America elevating its geographical and economic significance.

Considering the null hypothesis of data being non-stationary, all the stock market data were tested for stationarity using the ADF test and were found to be stationary at level using the first difference. Table 3 shows the *p*-values of stock market data showing the unit root and the *p*-value for the first difference of the data. As per the ADF test, the significance value p was less than 5% at first difference level, thus leading to the rejection of the null hypothesis. The stock market data was then used for further processing.

Table 3. Augmented Dickey–Fuller (ADF) test for stationarity of stock data.

Company/Stock Market	*p*-Value of Stock Data (%)	*p*-Value of First Difference of Stock Data (%)
MSFT	99.9	0.01
APPLE	99.3	0.01
LSE	99.9	0.01
NASDAQ,	99.9	0.01
GOOGLE	99.9	0.01
KSE	99.9	0.01
NYSE	99.9	0.01

The close stock values are estimated using multi-step prediction implying the measurement of the accuracy of predicted outcomes at multiple steps in the future. A step represents the time unit for which stock data is forecasted in the future. Then, the 1, 6 and 12 steps represent the forecasting of data for 1, 6 and 12-time units in future. The stock market data comprises weekly recorded prices, thus 1, 6 and 12 step predictions predict stock prices for 1, 6 and 12 weeks ahead in time.

4.2. Evaluation Metrics

The prediction performance of our proposed system is measured through multiple metrics [42–48]. The comparison is mainly drawn based on the difference between the actual value and the predicted one [2]. These evaluation metrics are explained here:

4.2.1. Root Mean Squared Error—RMSE

It is a quadratic score principle used to determine the average magnitude of estimation error in stock market trends [49]. The mathematical representation of RMSE is given as below,

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{n}(forecast(t) - actual(t))^2}{n}} \tag{5}$$

4.2.2. Mean Absolute Error—MAE

It is an average measure of errors in the prediction of stock market indices [49]. The average error is calculated without considering the directions of the set of predictions and each set of difference is having equal weight.

$$\text{MAE} = \frac{\sum_{t=1}^{n}|forecast(t) - actual(t)|}{n} \tag{6}$$

In the equations above n represents the number of estimated values, $forecast(t)$. and $actual(t)$ represent the estimated value and the actual value w.r.t time t respectively.

5. Results and Discussion

5.1. Results for NASDAQ

In this section, we present the results for the stock forecasting prediction. In our first experiment, we carried out the experiment for the NASDAQ stock exchange. In Figure 1 the prediction is done on "Close" for 12 steps ahead. The complete dataset consists of 20 years and we have trained our model on 70% data and tested our model on 30% of the data. The results in the following figure show very good results for prediction. The prediction is very good at the start and varies somewhat at the end of the data. We also carried out the 1, 6 and 12 steps ahead prediction. The results are presented in Figure 1. The first figure shows the results of the training data and the results are extremely good. The results are shown for original data for the 1, 6 and 12 step prediction. The prediction is so good that the original values exceed all predicted values. The results for testing data for 1, 6 and 12 step prediction are shown in Figure 1c. The results on the testing data are also very good. These excellent results

show the performance of our proposed model. There is some error in the prediction for testing data because the error propagates. The results show that the 1 step ahead prediction is closest to the original values. The error becomes more for the 6 step and 12 step ahead predictions. In the last step, the future values are predicted for the NASDAQ stock exchange. The next 12 steps are predicted that can help investors to check the future patterns of the stock market.


(**a**)

(**b**)

(**c**)
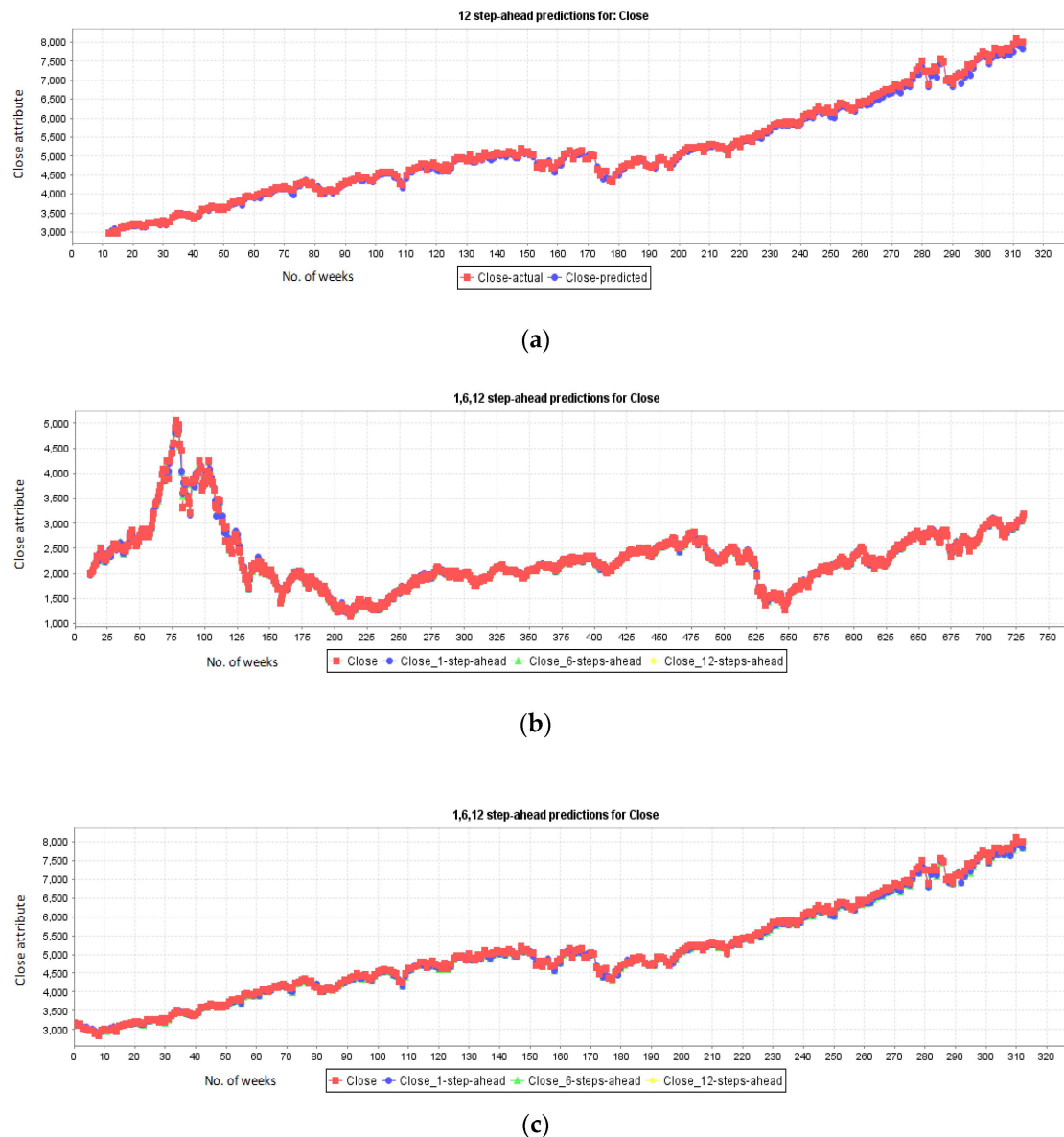
Figure 1. (**a**) NASDAQ stock 12 steps prediction on test data (**b**) NASDAQ stock 1-step, 6-step, and 12-step prediction on training data (**c**) NASDAQ stock 1-step, 6-step and 12-step prediction on test data.

5.2. Results for New York Stock Exchange

In our second experiment, we carried out an experiment for the NYSE. In Figure 2 the prediction is done on "Close" for 12 steps ahead. The complete dataset consists of 20 years and we have trained our model on 70% data and tested our model on 30% of the data. The results in the following figure show even better results for a prediction on test data. There is the only a deviation from the original data and shows almost similar results to training data. The experiment is also carried out for 1, 6 and 12 step ahead predictions. The results are presented in Figure 2. The first figure shows the results of

the training data and the results are extremely good. The results are shown for the original data, 1 step, 6 step, and 12 step predictions. The prediction is so good that the original values exceed all prediction values. The results for testing data for the 1, 6 and 12 step predictions are shown in Figure 2c. The results on the testing data are excellent too. There are very little deviations from the original data for all 1, 6 and 1 step forecasting. All the predictions are almost very similar to original data. The results show that 1, 6 and 12 step ahead predictions are closest to the original values. In the last step, the future values are predicted for the NYSE. The next 12 steps are predicted that can help investors to check the future patterns of the stock market.
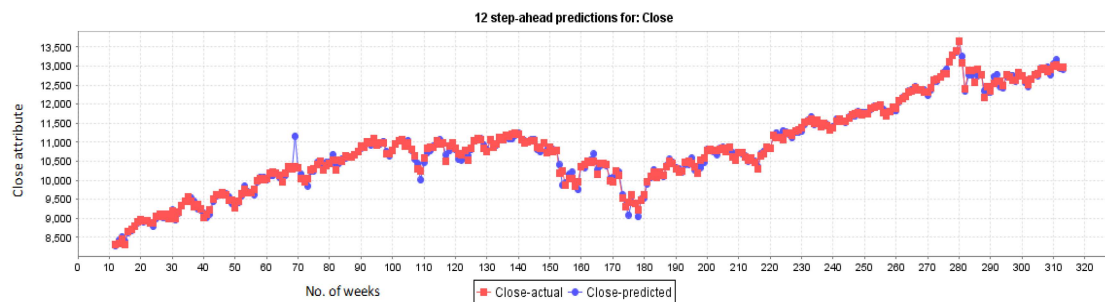
(**a**)
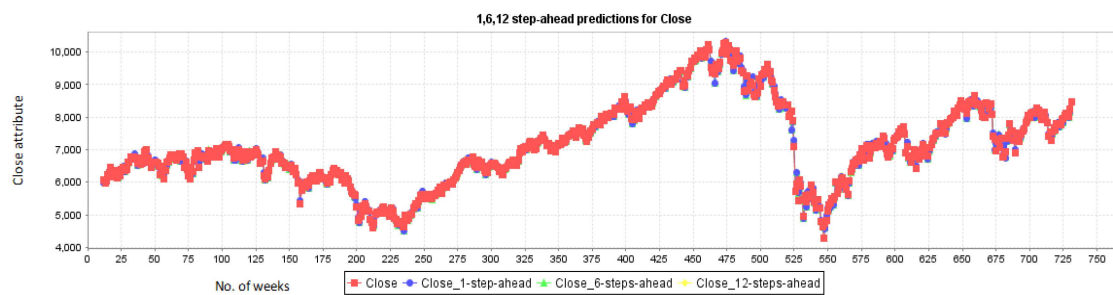
(**b**)
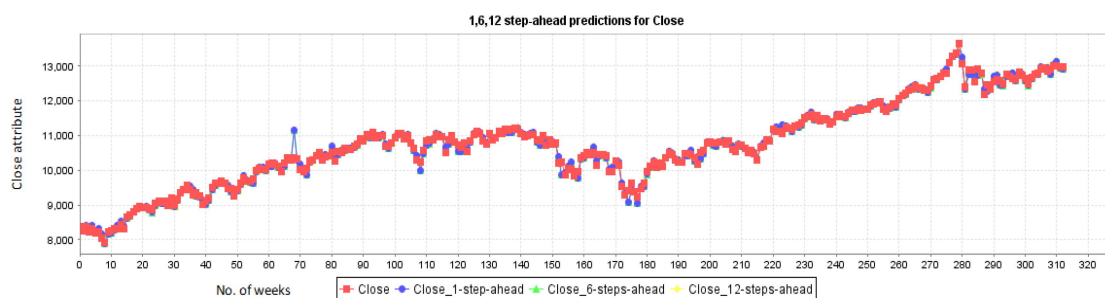
(**c**)

Figure 2. (**a**) New York stock 12 steps prediction on test data (**b**) New York stock 1, 6 and 12-step prediction on training data (**c**) New York stock 1, 6 and 12-step prediction on test data.

5.3. Results for London Stock Exchange

In our third experiment, we carried out an experiment for the LSE. In Figure 3 the prediction is done on "Close" for 12 steps ahead. The complete dataset consists of 20 years and we have trained our model on 70% data and tested our model on 30% of the data. The results in the following figure show even better results for a prediction on test data. There is only a small deviation from the original data and shows almost similar results to training data. There is one spike in the training data but the

prediction is not similar to training for that specific data point. The reason is that prediction is made based on the previous pattern and the previous data is uniform. The experiment is also carried out for 1, 6 and 12 ahead predictions. The first figure shows the results of the training data and the results are extremely good. The results are shown for original data, 1, 6 and 12 step prediction. The prediction is so good that the original values exceed all prediction values. The results for the testing data for 1, 6 and 12 step predictions are shown in Figure 3c. The results on testing data are also good. There are very little deviations from the original data for all 1, 6 and 12 step forecasting. The predictions for 1, 6 and 12 step ahead predictions are almost similar. The deviations from the original data are almost similar for all forecasting. In the last step, the future values are predicted for the NYSE. The next 12 steps are predicted that can help investors to check the future patterns of the stock market.
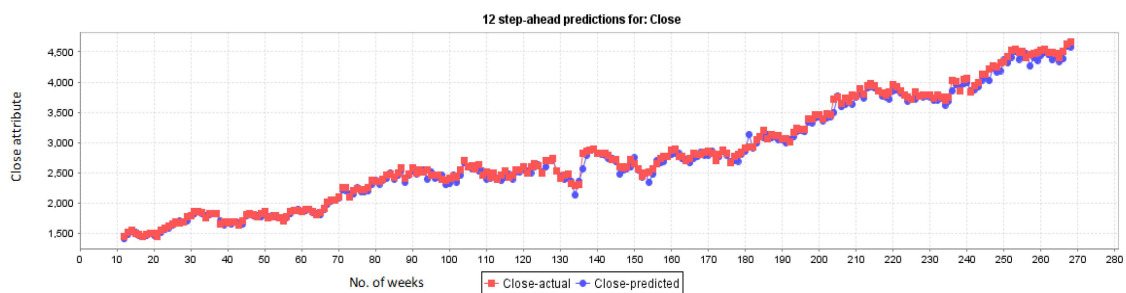
(**a**)

(**b**)
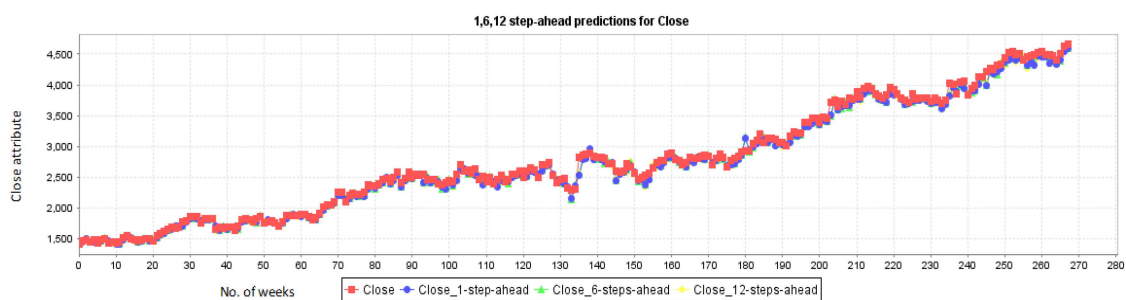
(**c**)

Figure 3. (**a**) London stock 12 steps prediction on test data (**b**) London stock 1, 6 and 12-step prediction on training data (**c**) London 1, 6 and 12-step prediction on test data.

5.4. Results for Karachi Stock Exchange

In our fourth experiment, we carried out an experiment for the KSE. In Figure 4 the prediction is done on "Close" for 12 steps ahead. The complete dataset consists of 20 years and we have trained our model on 70% data and tested our model on 30% of the data. There is only a small deviation from

the original data and shows almost similar results to training data. The experiment is also carried out for 1, 6 and 12step ahead prediction. The results are presented in Figure 4b for 1, 6 and 12 step on training and testing data respectively. The deviations from the original data are almost similar for all forecasting. The results show that error is extremely small for all types of forecasting. In the last step, the future values are predicted for the LSE. The next 12 steps are predicted that can help investors to check the future patterns of the stock market.
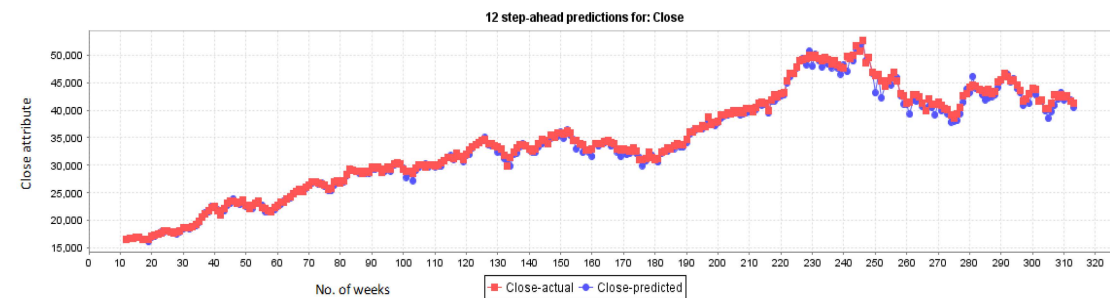
(**a**)
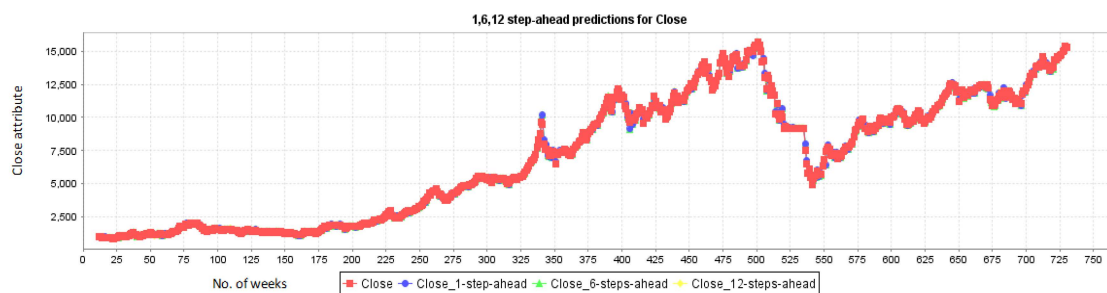
(**b**)
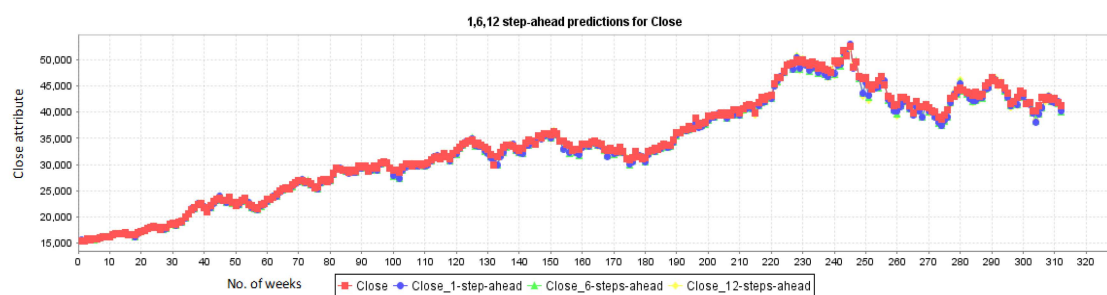
(**c**)

Figure 4. (**a**) Karachi Stock Exchange (KSE) 12 steps prediction on test data (**b**) KSE 1, 6 and 12-step prediction on training data (**c**) KSE 1, 6 and 12-step prediction on test data.

5.5. Stock Prediction for Companies:

In our fifth experiment, we carried out an experiment for the top three companies—Microsoft, Apple, and Google. In this section, only the prediction results on test data and 1, 6 and 12 ahead of data is presented. Figures 5–7 show the results for Microsoft, Apple, and Google (14 years maximum available data) for 12 step prediction respectively. The results are excellent for Microsoft and Apple and face some minor deviations in Google forecasting. The experiment is also carried out for 1, 6 and 12 step ahead predictions. The results are presented in the figure for 1, 6 and 12 steps on training and testing data respectively. The Figures 5–7 show the results for Microsoft, Apple, and Google for 12 step

prediction respectively. The deviations from the original data are almost similar for all forecasting. The results show that error is extremely small for all types of forecasting. In the last step, the future values are predicted for the LSE. Prediction errors in terms of MAE and RMSE are presented in Tables 4 and 5. Since the stock market values are predicted for one week, six weeks and twelve weeks ahead, the predicted error increases as the forecasting use the already predicted values for future prediction. Still, this method achieves this performance robustly.

Table 4. Mean Average Error (MAE) for all companies and stock markets for last 20 years of data (14 years of maximum available for Google).

	1 Step	2 Step	3 Step	4 Step	5 Step	6 Step	7 Step	8 Step	9 Step	10 Step	11 Step	12 Step
Google	3.2	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.4	3.4	3.4	3.4
MSFT	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.37	0.37	0.37	0.37	0.37
APP	0.23	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24
KSE	63.5	64.3	64.5	64.5	65.0	65.1	65.1	65.3	65.3	65.7	65.7	65.7
NSDQ	26.6	26.4	26.5	26.4	26.5	26.7	26.8	26.8	26.9	26.8	26.9	26.9
NY	54.1	54.1	54.2	54.2	54.2	54.3	54.5	54.3	54.3	54.3	54.3	54.5
LSE	14.1	14.2	14.6	14.6	14.7	14.6	14.7	14.6	14.6	14.5	14.6	14.7

Table 5. Root Mean Squared Error (RMSE) for all stock markets and companies for last 20 years of data (14 years of maximum available for Google).

|	1 Step	2 Step	3 Step	4 Step	5 Step	6 Step	7 Step	8 Step	9 Step	10 Step	11 Step	12 Step
Google	4.4	4.6	4.6	4.6	4.6	4.7	4.7	4.7	4.7	4.7	4.7	4.7
MSFT	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.3	0.5	0.5	0.5
APP	0.41	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.43	0.43	0.43	0.43
KSE	109.3	110.6	110.6	110.9	111.1	111.2	111.4	111.5	111.7	111.7	111.8	111.8
NSDQ	40.6	40.5	40.4	40.5	40.5	40.9	41.2	41.2	41.4	41.4	41.4	41.5
NY	75.9	75.9	75.8	75.9	76	76.1	76.1	75.9	75.9	76	76	76
LSE	20.1	20.5	20.8	20.7	20.1	20.9	20.9	20.5	20.5	20.4	20.5	20.7
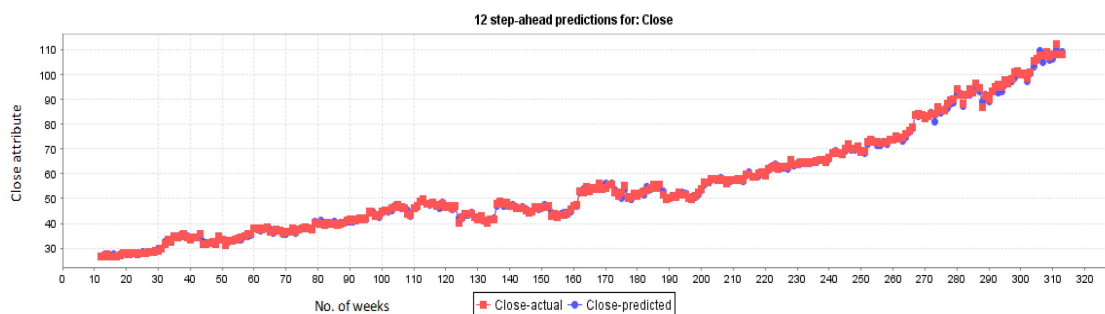
(**a**)
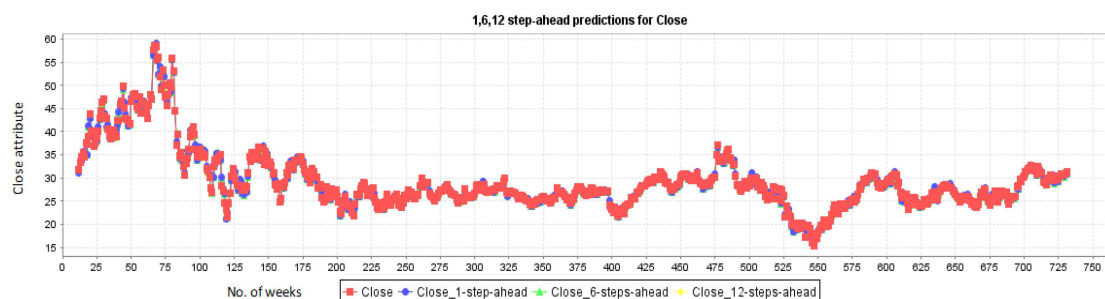
(**b**)

Figure 5. *Cont.*
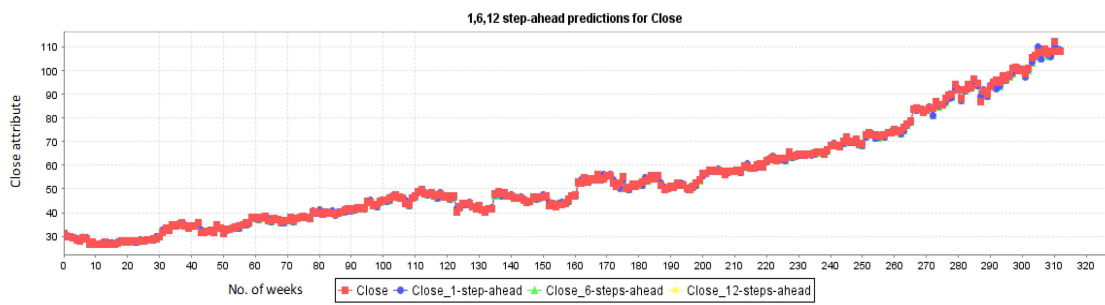
(**c**)

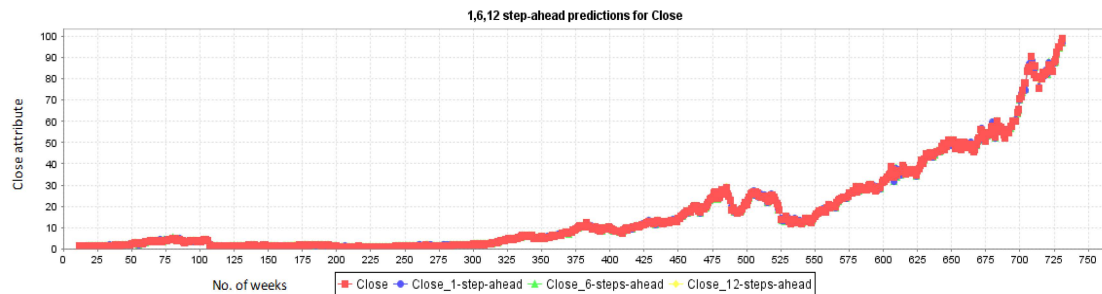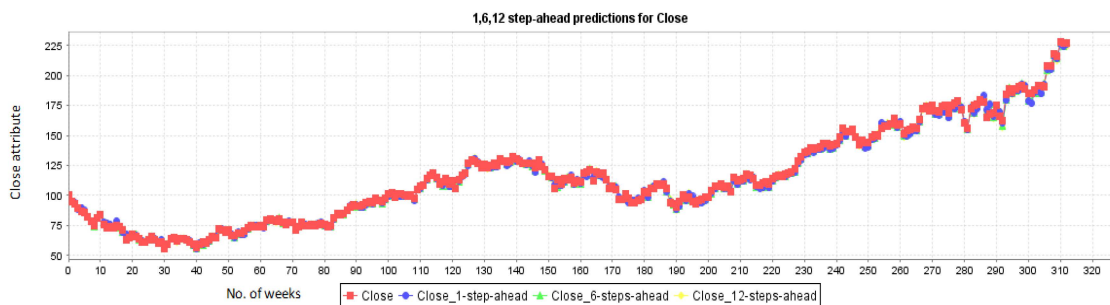**Figure 5.** (**a**) Microsoft 12 steps prediction on test data (**b**) Microsoft 1, 6 and 12-step prediction on training data (**c**) Microsoft 1-step, 6-step and 12-step prediction on test data.
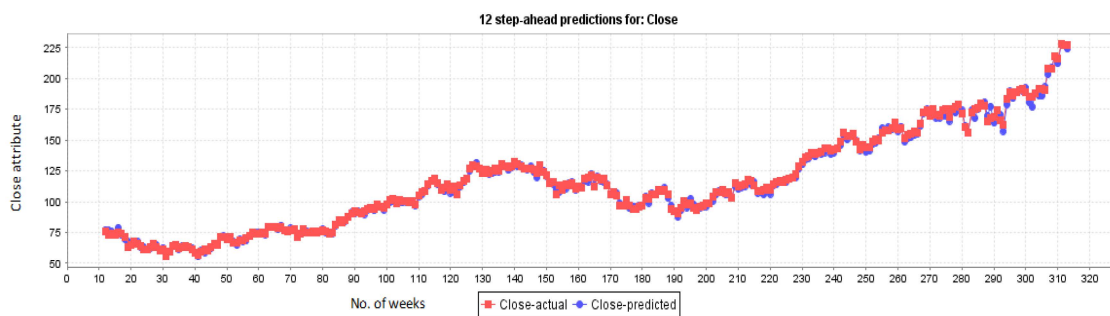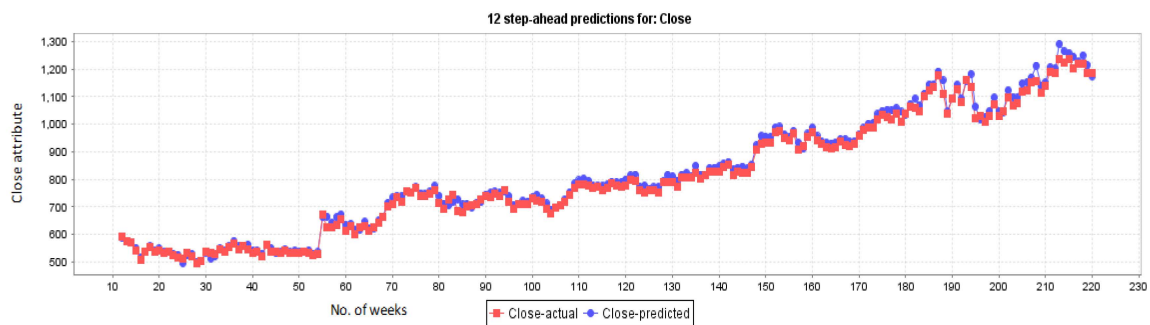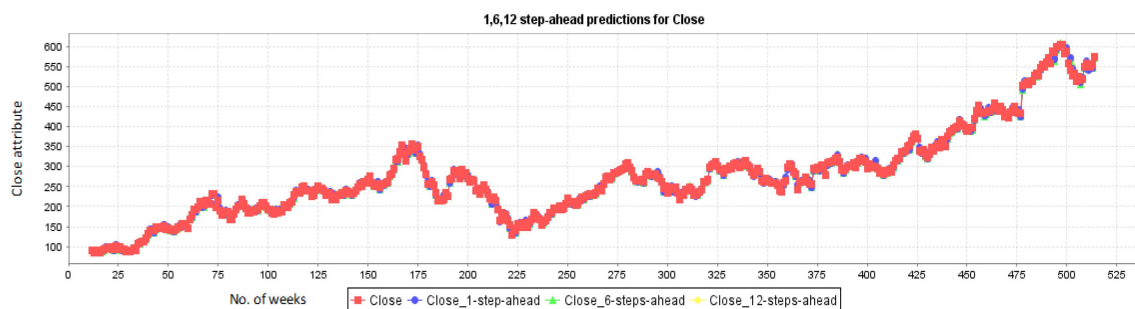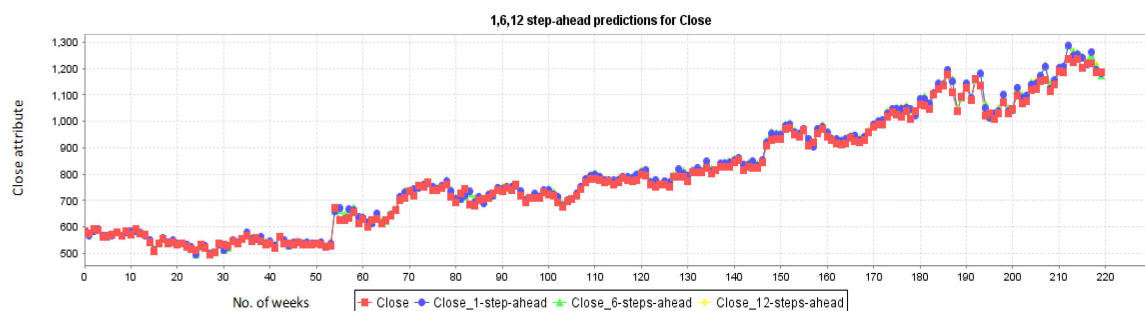

(**a**)

(**b**)

(**c**)

Figure 6. (**a**) Apple 12 steps prediction on test data (**b**) Apple 1, 6 and 12-step prediction on training data (**c**) Apple 1-step, 6-step and 12-step prediction on test data.

(**a**)

(**b**)

(**c**)

Figure 7. (**a**) Google 12 steps stock prediction on test data (**b**) Google stock 1, 6 and 12-step prediction on training data (**c**) Google stock 1, 6 and 12-step prediction on test data.

In our last experiment, we find the correlation between different stock exchange markets. Other researchers have checked the different factors such as political events and sentiment analysis. In this paper, we checked the correlation between different stock exchange companies. All possible combinations for 4 stock exchange markets we evaluated in this paper and the results are presented in Table 6. There is a negative correlation between the KSE and NASDAQ, NY stock exchange and LSE. The correlation values for NASDAQ, New York, and LSE are −0.02, −0.019 and −0.025 respectively. Interestingly, there is no correlation between the KSE and these top 3 stock exchange markets. On the other hand, the NASDAQ and LSE have a positive correlation value of 0.57. NY and LSE have a 0.522 correlation. Interestingly, New York and NASDAQ stock exchange have a very strong positive correlation of 0.829. This is because both are top stock exchanges situated in the USA. The correlation results are shown in graphical form in Figure 8.

Table 6. Correlation results for stock markets.

SR #	Stock Market Pair	Correlation
1	KSE, NASDAQ	−0.02
2	KSE, NY	−0.019
3	KSE, LSE	−0.025
4	NASDAQ, LSE	0.57
5	NY, LSE	0.522
6	NY, NASDAQ	0.829

(**a**)

(**b**)

(**c**)

(**d**)

(**e**)

(**f**)
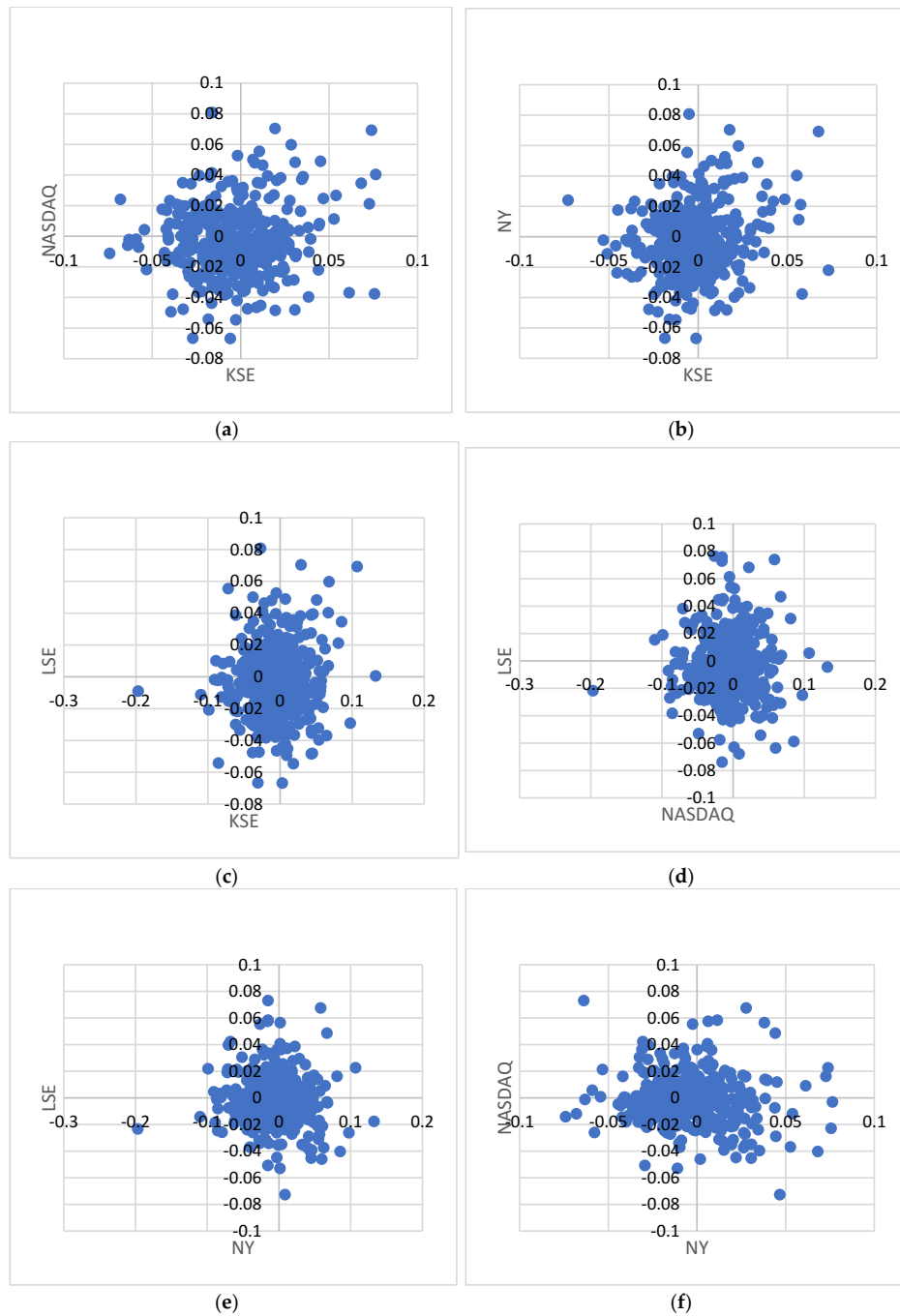
Figure 8. (**a**) Correlation between KSE NASDAQ (**b**) Correlation between KSE and New York (NY) (**c**) Correlation between KSE and London Stock Exchange (LSE) (**d**)Correlation between NASDAQ and LSE (**e**) Correlation between NY and LSE (**f**) Correlation between NY and NASDAQ.

5.6. Robustness Analysis for the Proposed Method

Stock Exchange prediction using linear regression is performed over the historical dataset of about 20 years. The proposed method is tested for robustness varying the distribution of data in terms of years and corresponding computational time is recorded for both training and testing. The test included the regression-based forecasting over the percentage-based data distribution. We compared the training results and testing results for linear regression and SVM-regression. Figure 9a,b shows the training and testing results respectively for stock markets. Similarly, Figure 9c,d represents results for training and testing respectively for different companies. The results are plotted against the increasing order of data distribution. The results show linear regression performs way better as compared to SVM-regression in terms of computational time.
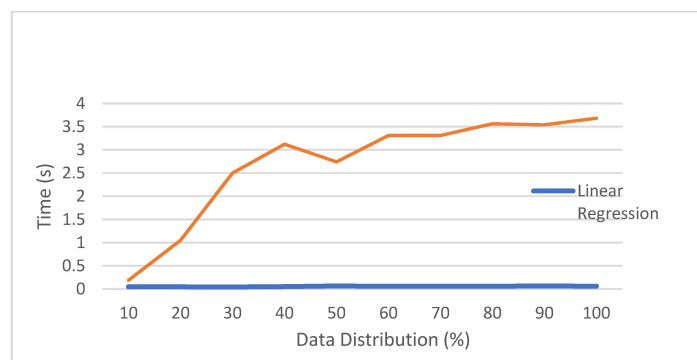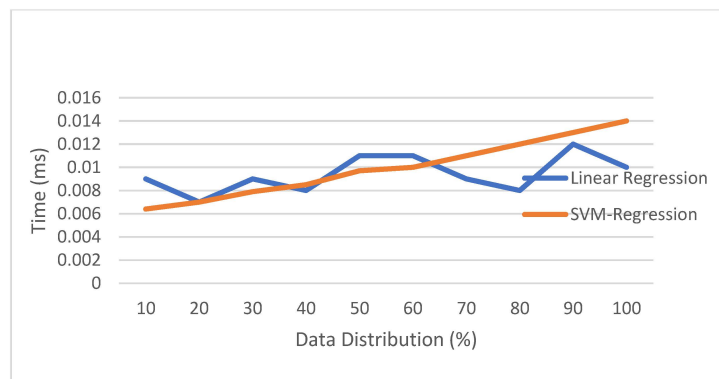

(**a**)
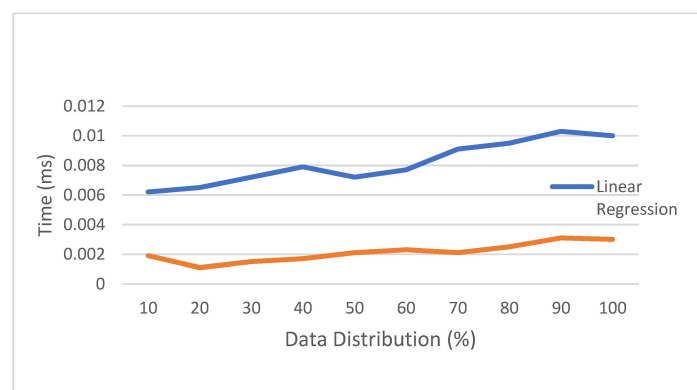
(**b**)
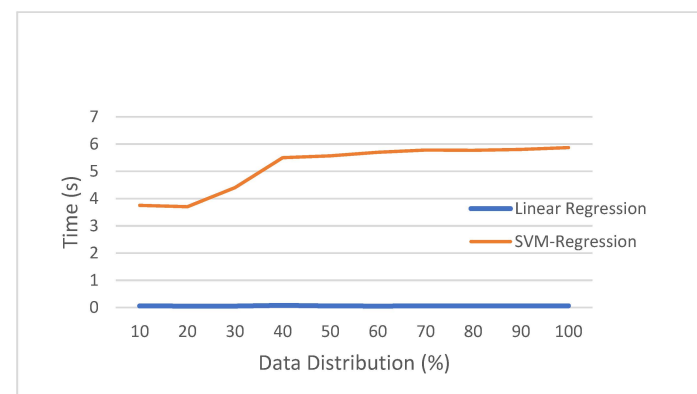
(**c**)

Figure 9. *Cont.*

(**d**)

Figure 9. (**a**) Average Computational Training Times using Linear Regression and Support Vector Machines (SVM)–Regression for Stock Markets (**b**) Average Computational Testing Times using Linear Regression and SVM–Regression for Stock Markets (**c**) Average Computational Training Times using Linear Regression and SVM–Regression for Companies (**d**) Average Computational Testing Times using Linear Regression and SVM–Regression for Companies.

It is noted that the training and testing time falls closely with one another over the range of increasing data distribution, thereby showing its robustness.

6. Conclusions

Intelligent stock exchange prediction is an important aspect for business investment plans. Non-linear and complex factors make it difficult to predict stock exchange indices. We propose a regression-based model to predict stock exchange indices. The proposed model is trained over a historical data of 20 years for 4 stock exchange markets—NASDAQ, New York, London, and KSE. The model is also evaluated for top 3 stock companies—Microsoft, Apple, and Google. The results show that forecasting for a different step ahead is very close to the original data. The time series forecasting is also presented in the paper. The dynamic correlation between different stock markets is also calculated and presented. The results show that there is no effect of NASDAQ, London and New York on KSE, while the other 3 stock exchanges share a positive correlation with each other. The highest correlation is between NASDAQ and the NYSE which is found to be 0.829. In the future, we are planning to design a hybrid deep learning based model for stock exchange prediction.

References

1.   Klein, M.D.; Datta, G.S. Statistical disclosure control via sufficiency under the multiple linear regression model. *J. Stat. Theory Pract.* **2018**, *12*, 100–110. [CrossRef]

2. Guan, H.; Dai, Z.; Zhao, A.; He, J. A novel stock forecasting model based on high-order-fuzzy-fluctuation trends and back propagation neural network. *PLoS ONE* **2018**, *13*, e0192366. [CrossRef] [PubMed]

3. Zhong, X.; Enke, D. Forecasting daily stock market return using dimensionality reduction. *Expert Syst. Appl.* **2017**, *67*, 126–139. [CrossRef]

4. Gong, X.; Si, Y.-W.; Fong, S.; Biuk-Aghai, R.P. Financial time series pattern matching with extended UCR Suite and Support Vector Machine. *Expert Syst. Appl.* **2016**, *55*, 284–296. [CrossRef]

5. Xi, Y.; Peng, H.; Qin, Y.; Xie, W.; Chen, X. Bayesian analysis of heavy-tailed market microstructure model and its application in stock markets. *Math. Comput. Simul.* **2015**, *117*, 141–153. [CrossRef]

6. Rahman, H.F.; Sarker, R.; Essam, D. A genetic algorithm for permutation flow shop scheduling under make to stock production system. *Comput. Ind. Eng.* **2015**, *90*, 12–24. [CrossRef]

7. Göçken, M.; Özçalıcı, M.; Boru, A.; Dosdoğru, A.T. Integrating metaheuristics and artificial neural networks for improved stock price prediction. *Expert Syst. Appl.* **2016**, *44*, 320–331. [CrossRef]

8. Kodogiannis, V.; Lolis, A. Forecasting financial time series using neural network and fuzzy system-based techniques. *Neural Comput. Appl.* **2002**, *11*, 90–102. [CrossRef]

9. Ravichandran, K.; Thirunavukarasu, P.; Nallaswamy, R.; Babu, R. Estimation of return on investment in share market through ANN. *J. Theor. Appl. Inf. Technol.* **2005**, *3*, 44–54.

10. Xi, L.; Muzhou, H.; Lee, M.H.; Li, J.; Wei, D.; Hai, H.; Wu, Y. A new constructive neural network method for noise processing and its application on stock market prediction. *Appl. Soft Comput.* **2014**, *15*, 57–66. [CrossRef]

11. Shen, W.; Zhang, Y.; Ma, X. Stock return forecast with LS-SVM and particle swarm optimization. In Proceedings of the International Conference on Business Intelligence and Financial Engineering (BIFE'09), Beijing, China, 24–26 July 2009; IEEE: Piscataway, NJ, USA, 2009.

12. Wen, Q.; Yang, Z.; Song, Y.; Jia, P. Automatic stock decision support system based on box theory and SVM algorithm. *Expert Syst. Appl.* **2010**, *37*, 1015–1022. [CrossRef]

13. Lin, Y.; Guo, H.; Hu, J. An SVM-based approach for stock market trend prediction. In Proceedings of the 2013 International Joint Conference on Neural Networks (IJCNN), Dallas, TX, USA, 4–9 August 2013; IEEE: Piscataway, NJ, USA, 2013.

14. Yu, H.; Chen, R.; Zhang, G. A SVM stock selection model within PCA. *Procedia Comput. Sci.* **2014**, *31*, 406–412. [CrossRef]

15. Wang, X.-Y.; Wang, Z.-O. Stock market time series data mining based on regularized neural network and rough set. In Proceedings of the 2002 International Conference on Machine Learning and Cybernetics, Beijing, China, 4–5 November 2002; IEEE: Piscataway, NJ, USA, 2002.

16. Wang, Y.-F. Mining stock price using fuzzy rough set system. *Expert Syst. Appl.* **2003**, *24*, 13–23. [CrossRef]

17. Nair, B.B.; Mohandas, V.; Sakthivel, N. A decision tree—Rough set hybrid system for stock market trend prediction. *Int. J. Comput. Appl.* **2010**, *6*, 1–6. [CrossRef]

18. Ives, M.C.; Scandol, J.P. A Bayesian analysis of NSW eastern king prawn stocks (*Melicertus plebejus*) using multiple model structures. *Fish. Res.* **2007**, *84*, 314–327. [CrossRef]

19. Su, Z.; Peterman, R.M. Performance of a Bayesian state-space model of semelparous species for stock-recruitment data subject to measurement error. *Ecol. Model.* **2012**, *224*, 76–89. [CrossRef]

20. Ticknor, J.L. A Bayesian regularized artificial neural network for stock market forecasting. *Expert Syst. Appl.* **2013**, *40*, 5501–5506. [CrossRef]

21. Miao, J.; Wang, P.; Xu, Z. A Bayesian dynamic stochastic general equilibrium model of stock market bubbles and business cycles. *Quant. Econ.* **2015**, *6*, 599–635. [CrossRef]

22. Wang, L.; Wang, Z.; Zhao, S.; Tan, S. Stock market trend prediction using dynamical Bayesian factor graph. *Expert Syst. Appl.* **2015**, *42*, 6267–6275. [CrossRef]

23. Li, H.; Sun, J.; Sun, B.-L. Financial distress prediction based on OR-CBR in the principle of *k*-nearest neighbors. *Expert Syst. Appl.* **2009**, *36*, 643–659. [CrossRef]

24. Teixeira, L.A.; De Oliveira, A.L.I. A method for automatic stock trading combining technical analysis and nearest neighbor classification. *Expert Syst. Appl.* **2010**, *37*, 6885–6890. [CrossRef]

25. Fu-Yuan, H. Forecasting stock price using a genetic fuzzy neural network. In Proceedings of the International Conference on Computer Science and Information Technology (ICCSIT'08), Singapore, 12 September 2008; IEEE: Piscataway, NJ, USA, 2008.

26. Sorensen, E.H.; Miller, K.L.; Ooi, C.K. The decision tree approach to stock selection. *J. Portf. Manag.* **2000**, *27*, 42–52. [CrossRef]

27. Wu, M.-C.; Lin, S.-Y.; Lin, C.-H. An effective application of decision tree to stock trading. *Expert Syst. Appl.* **2006**, *31*, 270–274. [CrossRef]

28. Hu, Y.; Feng, B.; Zhang, X.; Ngai, E.W.T.; Liu, M. Stock trading rule discovery with an evolutionary trend following model. *Expert Syst. Appl.* **2015**, *42*, 212–222. [CrossRef]

29. Hassan, M.R.; Nath, B.; Kirley, M. A fusion model of HMM, ANN and GA for stock market forecasting. *Expert Syst. Appl.* **2007**, *33*, 171–180. [CrossRef]

30. Huang, S.-C.; Wu, T.-K. Integrating GA-based time-scale feature extractions with SVMs for stock index forecasting. *Expert Syst. Appl.* **2008**, *35*, 2080–2088. [CrossRef]

31. Cervelló-Royo, R.; Guijarro, F.; Michniuk, K. Stock market trading rule based on pattern recognition and technical analysis: Forecasting the DJIA index with intraday data. *Expert Syst. Appl.* **2015**, *42*, 5963–5975. [CrossRef]

32. Kim, Y.; Enke, D. Developing a rule change trading system for the futures market using rough set analysis. *Expert Syst. Appl.* **2016**, *59*, 165–173. [CrossRef]

33. Podsiadlo, M.; Rybinski, H. Financial time series forecasting using rough sets with time-weighted rule voting. *Expert Syst. Appl.* **2016**, *66*, 219–233. [CrossRef]

34. Chiang, W.-C.; Enke, D.; Wu, T.; Wang, R. An adaptive stock index trading decision support system. *Expert Syst. Appl.* **2016**, *59*, 195–207. [CrossRef]

35. Majhi, R.; Panda, G.; Sahoo, G. Development and performance evaluation of FLANN based model for forecasting of stock markets. *Expert Syst. Appl.* **2009**, *36*, 6800–6808. [CrossRef]

36. Chakravarty, S.; Dash, P.K. A PSO based integrated functional link net and interval type-2 fuzzy logic system for predicting stock market indices. *Appl. Soft Comput.* **2012**, *12*, 931–941. [CrossRef]

37. Dash, R.; Dash, P.K.; Bisoi, R. A self-adaptive differential harmony search based optimized extreme learning machine for financial time series prediction. *Swarm Evol. Comput.* **2014**, *19*, 25–42. [CrossRef]

38. Huang, C.-F. A hybrid stock selection model using genetic algorithms and support vector regression. *Appl. Soft Comput.* **2012**, *12*, 807–818. [CrossRef]

39. Wang, D.; Liu, X.; Wang, M. A DT-SVM strategy for stock futures prediction with big data. In Proceedings of the 2013 IEEE 16th International Conference on Computational Science and Engineering (CSE), Sidney, Australia, 3–5 December 2013; IEEE: Piscataway, NJ, USA, 2013.

40. Nayak, R.K.; Mishra, D.; Rath, A.K. A Naïve SVM-KNN based stock market trend reversal analysis for Indian benchmark indices. *Appl. Soft Comput.* **2015**, *35*, 670–680. [CrossRef]

41. Nunno, L. *Stock Market Price Prediction Using Linear and Polynomial Regression Models*; Computer Science Department, University of New Mexico: Albuquerque, NM, USA, 2014.

42. Muhammad, K.; Ahmad, J.; Mehmood, I.; Rho, S.; Baik, S.W. Convolutional neural networks based fire detection in surveillance videos. *IEEE Access* **2018**, *6*, 18174–18183. [CrossRef]

43. Mehmood, I.; Sajjad, M.; Muhammad, K.; Shah, S.I.S.; Sangaiah, A.K.; Shoaib, M.; Baik, S.W. An efficient computerized decision support system for the analysis and 3D visualization of brain tumor. *Multimedia Tools Appl.* **2018**, 1–26. [CrossRef]

44. Ateeq, T.; Majeed, M.N.; Anwar, S.M.; Maqsood, M.; Rehman, Z.; Lee, J.W.; Muhammad, K.; Wang, S.; Baik, S.W.; Mehmood, I. Ensemble-classifiers-assisted detection of cerebral microbleeds in brain MRI. *Comput. Electr. Eng.* **2018**, *69*, 768–781. [CrossRef]

45. Aadil, F.; Raza, A.; Khan, M.F.; Maqsood, M.; Mehmood, I.; Rho, S. Energy aware cluster-based routing in flying ad-hoc networks. *Sensors* **2018**, *18*, 1413. [CrossRef] [PubMed]

46. Muhammad, K.; Sajjad, M.; Mehmood, I.; Rho, S.; Baik, S. A novel magic LSB substitution method (M-LSB-SM) using multi-level encryption and achromatic component of an image. *Multimedia Tools Appl.* **2016**, *75*, 14867–14893. [CrossRef]

47. Muhammad, K.; Hussain, T.; Baik, S.W. Efficient CNN based summarization of surveillance videos for resource-constrained devices. *Pattern Recognit. Lett.* **2018**. [CrossRef]
48. Khan, S.; Khan, A.; Maqsood, M.; Aadil, F.; Ghazanfar, M.A. Optimized gabor feature extraction for mass classification using cuckoo search for big data e-healthcare. *J. Grid Comput.* **2018**. [CrossRef]
49. Willmott, C.J.; Matsuura, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim. Res.* **2005**, *30*, 79–82. [CrossRef]