Physica A 551 (2020) 124569

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

An approach to predict and forecast the price of constituents and index of cryptocurrency using machine learning

Reaz Chowdhury [a], M. Arifur Rahman [b], M. Sohel Rahman [c], M.R.C. Mahdy [a,d,*]

[a] Department of Electrical & Computer Engineering, North South University, Bashundhara, Dhaka, 1229, Bangladesh
[b] Department of Accounting & Finance, North South University, Bashundhara, Dhaka, 1229, Bangladesh
[c] Department of Computer Science & Engineering, Bangladesh University of Engineering & Technology, West Palasi, Dhaka 1205, Bangladesh
[d] Pi Labs Bangladesh LTD, Eden Center, 2/1/E, Toyenbee Rd, Dhaka 1000, Bangladesh

ARTICLE INFO

Article history:
Received 1 June 2019
Received in revised form 9 March 2020
Available online 14 April 2020

Keywords:
Cryptocurrency constituents
Cryptocurrency index
Close price
Machine learning
Data mining
Prediction
Forecasting

ABSTRACT

At present, cryptocurrencies have become a global phenomenon in financial sectors, as it is one of the most traded financial instruments worldwide. Cryptocurrency is not only one of the most complicated and abstruse fields among financial instruments but also deemed as a perplexing problem in finance due to its high volatility. This work makes an attempt to apply machine learning techniques on the index and constituents of cryptocurrency with a goal to predict and forecast prices thereof. In particular, the purpose of this article is to predict and forecast the close (closing) price of the cryptocurrency index 30 and nine constituents of cryptocurrencies using machine learning algorithms and models so that it becomes easier for people to trade these currencies. We have used several machine learning techniques and algorithms and compared the models with each other to get the best output. We believe that our work will help reduce the challenges and difficulties faced by people who invest in cryptocurrencies. Moreover, the obtained results can play a major role in cryptocurrency portfolio management and in observing the fluctuations in the prices of constituents of cryptocurrency market. We have also compared our approach with similar state of the art works from the literature, where machine learning approaches have been considered for predicting and forecasting the prices of these currencies. In the sequel, we have found that our best approach presents better and competitive results (especially by using ensemble learning method) than the best works from the literature thereby advancing the state of the art. Using such prediction and forecasting methods, people can easily understand the trend and it would be even easier for them to trade in a difficult and challenging financial instrument like cryptocurrency.

1. Introduction

Cryptocurrencies are a subset of virtual currencies that use cryptography for security. These are decentralized and open source currencies and hence function on a peer-to-peer basis. Cryptocurrencies mostly use a very complex cryptographic algorithm that requires connected network of computers to conduct computationally expensive mathematical operations [1]. Cryptocurrencies have a built-in implementation of cryptography in their design. At present, people are

* Corresponding author.
E-mail address: mahdy.chowdhury@northsouth.edu (M.R.C. Mahdy).

https://doi.org/10.1016/j.physa.2020.124569
0378-4371/

using cryptocurrencies to implement a new form of economy, because of its cheapness, online, and anonymous means of exchange. A list of cryptocurrencies and their prices can be found at https://coinmarketcap.com, which lists more than 2175 cryptocurrencies of varying types. Cryptocurrencies feature certain computer protocols that are out of any government control. These currencies are unregulated and highly volatile [1]. As a result, it can quickly devalue overnight. These currencies have aggressive swings in their prices, as it is largely based on public perception. It is therefore very hard to make related risk assessment at any moment. With the increase of the prices of cryptocurrencies, mining has also turned into a very advantageous business for the people [2]

One of the most valuable and decentralized cryptocurrencies is Bitcoin, which has been introduced by Satoshi Nakamoto on October 31, 2008 [3]. It captures around 35% of the total market capitalization [4]. Bitcoin's greatest innovation is blockchain, which has been introduced to solve the issue of double spending as well as to disrupt the control of centralized parties in the transaction of values. The blockchain is the technology, in which a record of any financial and economic transactions made in any cryptocurrency, is maintained using cluster of computers (which are linked in a peer-to-peer network). In simple terms, it is a powerful technology, which has the capacity to maintain permanent records of commercial transactions, transfer of assets and contracts, financial records, and intellectual property [5]. Blockchain is completely a public ledger that is made up of blocks and any node connected on the Bitcoin network can process and clear a transaction by posting the transaction [6].

While the prices of cryptocurrencies have gone up since 2016 with great fluctuation, the enthusiasm of people to invest more and more in these virtual currencies stays more or less constant. These virtual currencies are nowadays used in official cash flows and exchange of goods. As a result, in recent years, various physical approaches and modeling techniques have been introduced by researchers and scholars to model the price of cryptocurrencies and to analyze the spontaneity of the market for making real decision support systems [4,7]. These techniques include, but are not limited to, various dynamic topic modeling, machine learning, data mining, and text mining approaches. Moreover, to study the cryptocurrency market, agent based artificial financial market and genetic programming for finding attractive technical patterns have also been proposed [8,9]. In addition, as cryptocurrencies are correlated [10], the cross correlation between price changes of various cryptocurrencies using random matrix theory and minimum spanning trees has also been studied [11]. In recent years, different machine learning algorithms and techniques have also been taken into account to generate abnormal profits by exploiting the inefficiency of the cryptocurrency market [12]. The involvement of some cryptocurrencies including Bitcoin in illicit activities can also be accurately measured using machine learning approach [13].

The aim of this work is to predict and forecast the close price of cci30 and constituents of cryptocurrencies thereby helping in minimize the risk in cryptocurrency market. We have collected and analyzed the historical data from https://coinmarketcap.com and have applied various machine learning approaches, employing software RapidMiner [14], to find and analyze the close price of nine cryptocurrencies and for the index, cci30. In particular, we have considered an ensemble learning method, gradient boosted trees model, neural net model, K-Nearest Neighbor (K-NN) model, and have analyzed the performance thereof using the standard measures/metrics from the literature. Out of all four models, the best accuracy we have obtained so far is by using ensemble learning method. As the volatility of these currencies is extremely high, it may not be possible to model these currencies using just one algorithm for either prediction or forecast. This fact actually motivates us to consider and study four different models in this work. Using these models, we can easily observe the behavior of these currencies and decide, which algorithms could be better for prediction and forecasting of the close price thereof.

2. Background and literature review

Cryptocurrency is a new digital asset in finance, which has extremely high volatility as compared to almost all other financial instruments. As a result of its high volatility and price fluctuations, a very limited number of articles, to the best of our knowledge, exists in the literature that deal with predicting the price fluctuations.

Xiaolei et al. [15] have proposed three models; SVM, RF model and Light Gradient Boosting Machine to forecast the price trend of the cryptocurrency market, where the daily data of 42 kinds of cryptocurrencies have been combined with key economic indicators. Light Gradient Boosting Machine stands out to be a better model compared to other models. In the case of forecasting performance in the first category of training set, the test set is the true subset of the training set. In this case, the accuracy obtained using Light Gradient Boosting Machine is 0.776 for 2 months, 0.881 for 2 weeks, 0.762 for 2 days and for the first day of the period, the accuracy is 0.7622 for 2 months, 0.905 for 2 weeks and 0.548 for 2 days. In the case of forecasting performance in the second category of training set, the test set is not a subset of the training set. In this case, the obtained accuracy using Light Gradient Boosting Machine is 0.607 for 2 weeks, 0.476 for 2 days and for the first day of the period, the accuracy is 0.952 for 2 weeks and 0.93 for 2 days.

Lahmiri et al. [16] have proposed two deep learning techniques, namely, deep learning neural network (DLNN) and generalized regression neural networks (GRNN) to forecast the price of Bitcoin, Digital Cash and Ripple. In the case of Bitcoin, RMSE obtained using DLNN and GRNN are 2.75×10^3 and 8.80×10^3, respectively. In case of Digital Cash, these values are 19.2923 and 50.2418 and in case of Ripple, 0.0499 and 0.3115. It is seen that, in the case of Bitcoin and Digital Cash, the RMSE value obtained using DLNN and GRNN are extremely high, whereas in case of Ripple, this value is comparatively lower than others.

Table 1
Names of the constituents and prediction models under consideration in this paper.

Name of the nine constituents	Bitcoin Cash; Bitcoin; Dash; Doge coin (DOGE); Ethereum; IOTA (MIOTA); Litecoin; NEM; NEO.
Predictive models and learners	Gradient boosted trees, Neural net, Ensemble learning, K-NN

Kim et al. [17] have analyzed user comments in online cryptocurrency communities to predict fluctuations in the prices of the cryptocurrency and in the number of transactions thereof. The accuracy achieved for the predicted fluctuation in Bitcoin price and in Bitcoin transaction are 50.538% and 48.387% for 13 days. For Ethereum, the accuracy of price fluctuation and transaction fluctuation are 49.425% and 51.149% for 13 days. In the case of Ripple, the accuracy of Ripple price fluctuation is 63.200 for 13 days. Notably, Ripple transaction fluctuations have not been considered by Kim et al.

Greaves et al. [18] have proposed transaction graph data by collecting Bitcoin transactions to predict the Bitcoin prices. They have used four classification models: baseline, logistic regression, SVM and neural network model. The accuracy obtained using these models are 53.4% for baseline, 54.3% for logistic regression, 53.7% for SVM and 55.1% for neural network respectively.

Barnwal et al. [19] have proposed an approach of stacking with neural network for cryptocurrency investment. Bitcoin data has been used and obtained from quandl to predict the direction of Bitcoin's price. Extreme gradient boosting, K-NN, Light Gradient Boosting Machine are used as discriminative classifiers to create stacks, which are optimized over one layer of neural network to model the direction of the price of the cryptocurrencies. Two time periods are considered. Apr–May 2018 time period has been used to create level-1 data in stacking, and from June–July 2018, stacked generalizer's performance is compared to the rest of the models. Accuracies of extreme gradient boosting, SVM, K-NN, light gradient boosting, Random Forest, Logistic Elastic Net, Naïve Bayes, Linear Discriminant Analysis, Quadratic Discriminant Analysis and stacked Generalization for April–May (June–July) period have been found to be 0.57 (0.46), 0.48 (0.50), 0.59 (0.52), 0.61 (0.52), 0.55 (0.50), 0.52 (0.50), 0.52 (0.50), 0.54 (0.48), 0.55 (0.52), 0.52 (0.54) respectively.

McNanny et al. [20] have used recurrent neural network (RNN), long short time memory (LSTM) network and ARIMA model to predict the direction of Bitcoin price in USD. RNN and LSTM are two deep learning pipelines, which outperformed the ARIMA forecasting model. Root mean squared error (RMSE) is used to evaluate and compare the regression accuracy and an 80/20 holdout validation strategy is used to instrument the validation of models. As a result, the accuracy and RMSE obtained using ARIMA model are 50.05% and 53.74%. Using RNN (LSTM) model, the accuracy and RMSE obtained are 50.25% (52.78%) and 5.45% (6.87%).

Bakar et al. [21] have implemented an ARIMA forecasting method to determine the accuracy of Bitcoin exchange rates in a high volatility environment. The have achieved absolute percentage errors of 1.4% and 9.3% for September 2017 and October 2017 respectively with a mean absolute percentage error between forecast and actual value is 5.36%.

Rebane et al. [22] have presented an ARIMA model and seq2seq recurrent deep multi-layer neural network (sq2seq) utilizing a varied selection of input types. Cumulative error obtained using seq2seqA, seq2seq B, seq2seq c and Arima models are 1.00, 0.89, 0.45, 1.73 respectively. They have also presented visual comparison of Bitcoin prediction over 40 days, where the results of seq2seq A, seq2seq B, seq2seq c and Arima models are mostly seen to be extremely deviated from the true values.

3. Materials and methods

3.1. Dataset

Our dataset includes seven-day week daily data which we have obtained from https://coinmarketcap.com. The constituents and predictive models that we have considered are given in Table 1. We have considered seven attributes [see Table 2] and divided our data into two subsets — testing and training. By creating different models in RapidMiner, we have predicted the close price of the constituents and cci30 for the month January 2019 based on historical data. In order to perform gradient boosted tree model and neural network model, we have divided our data of constituents of cryptocurrencies into two subsets — training and testing datasets [see Table 2]. In the case of ensemble learning method, the dataset is given in Table 4. For K-NN model, the dataset contains only the training phase for all cryptocurrencies as shown in Table 3 with no testing phase because we plan to forecast the values of the month of January 2019.

3.2. Performance metrics

The performance metrics we have used in this paper are root mean squared error (RMSE), prediction trend accuracy, absolute error, relative error, squared error, correlation, and squared correlation, which are achieved through "Performance (Regression)" and "Forecasting Performance" operators.

Root mean squared error (RMSE) is the measure of the differences between values predicted by the model and the actual values. Prediction trend accuracy measures the average of times a regression prediction was able to correctly predict the trend of the regression. Absolute error is the average absolute deviation of the prediction from the actual value, where the values of the label attribute are the actual values. Relative error is the average of the absolute deviation

Table 2
Attributes of the dataset under consideration.

SL.	Attributes of the dataset	Remarks
1.	Date	Date or a trade date is a day at which an order is executed in the market to purchase, sell or otherwise acquire a currency is performed.
2.	Open price	Open (opening) price is the price at which a currency is first traded on a given trading day.
3.	Close price	Close (closing) price is the final price at which a currency is traded on a given trading day.
4.	High	High is the highest price at which a currency is traded on a given trading day.
5.	Low	Low is the lowest price at which a currency is traded on a given trading day.
6.	Volume	Volume or volume of trade is the total quantity of contracts traded for a specified currency.
7.	Market capital	Market capital refers to the total dollar market value of a currency's outstanding contracts.

Table 3
Training, testing and forecasting dataset of the constituents for gradient boosted trees, neural net and K-NN models.

Names of constituents	Training data	Testing and forecasting data
Bitcoin Cash	01.08.2017–31.12.2018	01.01.2019–31.01.2019
Bitcoin	27.12.2013–31.12.2018	01.01.2019–31.01.2019
Dash	14.02.2014–31.12.2018	01.01.2019–31.01.2019
Doge coin (DOGE)	27.12.2013–31.12.2018	01.01.2019–31.01.2019
Ethereum	07.08.2015–31.12.2018	01.01.2019–31.01.2019
IOTA (MIOTA)	13.06.2017–31.12.2018	01.01.2019–31.01.2019
Litecoin	27.12.2013–31.12.2018	01.01.2019–31.01.2019
NEM	01.04.2015–31.12.2018	01.01.2019–31.01.2019
NEO	25.10.2016–31.12.2018	01.01.2019–31.01.2019

Table 4
Training dataset of the companies for ensemble learning method.

Names of constituents	Dataset
Bitcoin Cash	01.08.2017–31.01.2019
Bitcoin	27.12.2013–31.01.2019
Dash	14.02.2014–31.01.2019
Doge coin (DOGE)	27.12.2013–31.01.2019
Ethereum	07.08.2015–31.01.2019
IOTA (MIOTA)	13.06.2017–31.01.2019
Litecoin	27.12.2013–31.01.2019
NEM	01.04.2015–31.01.2019
NEO	25.10.2016–31.01.2019

of the prediction from the actual value divided by actual value. Squared error chooses the model with the smallest average squared error value. Correlation returns the correlation coefficient between the "label" and "prediction" attributes. Squared correlation returns the squared correlation coefficient between the "label" and "prediction" attributes.

3.3. Methodology

Machine learning approaches can play a wide range of critical roles in the finance domain especially when it comes to predict the prices of financial instruments in general and cryptocurrencies in particular. Starting from managing cryptocurrency portfolio [23] to predicting the fluctuations in the prices of cryptocurrencies transactions [17], machine learning stands out to be one of the best approaches and techniques. Machine learning techniques can be integrated into business intelligence systems for making real life decisions [24]. Effort to predict and analyze the price of cryptocurrencies have been a very challenging one due to its high volatility and price fluctuations. Thus, it is hoped that, ML techniques will bring a new dimension in this domain and as it has already been discussed in the literature review, we already have a number of ML based approaches in this domain in the literature.

In this paper, we have used the widely used software called RapidMiner as it supports all steps of a data mining process [12]. In RapidMiner software, for performing data analysis usually graphs, plots, charts and tables are used in which one can easily visualize the output and also compare between one or more attributes and models. For a machine to predict and forecast the future close price of any cryptocurrency, it is essential to train the machine to learn from the given dataset. Based on these datasets, models will be created applying different algorithms and thus the prediction/forecasting task will be accomplished [25].

3.3.1. Predictive model: Gradient boosted trees

Gradient boosted trees model is very advantageous especially in the context of price prediction for a number of reasons as follows. Firstly, it is not required to normalize the data in this case as it is sensitive to arithmetic range of data and

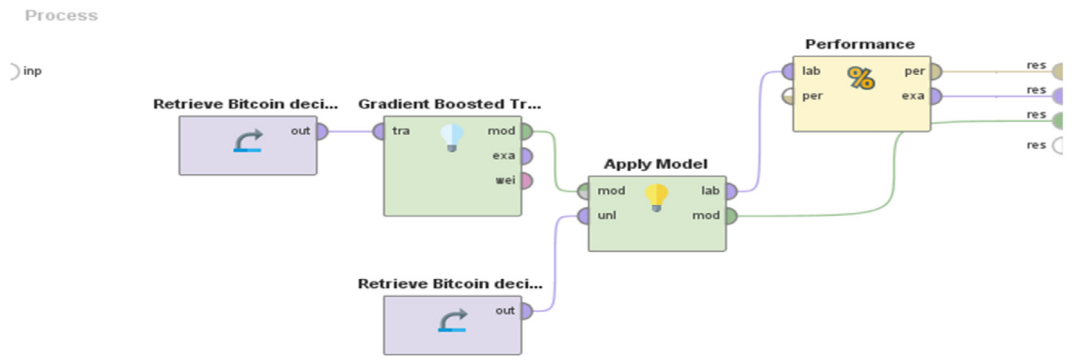

Fig. 1. Gradient boosted trees model for predicting constituents and index of cryptocurrencies.

features. Secondly, it is a very scalable machine learning model due to its construction process and finally, it is also a rule-based learning method [26]. A number of works dealing with prediction and forecasting of sales as well as cryptocurrency prices in the literature have successfully employed gradient boosted trees model [12,15,27].

Gradient boosted trees model is an ensemble of either regression or classification tree models, which is a forward learning ensemble method that obtains predictive results through gradually improved estimations. For predicting the close price of cryptocurrency of the month January 2019, we have considered the attributes of dataset stated in Table 2 and provided the historical price of the constituents as shown in Table 3. We have further used RapidMiner to optimize the parameters of the model, which runs through various permutations to get the best value of its parameters. The parameter, number of trees is tuned to 500 through optimization, while the other parameters are set as default. The "Performance (Regression)" and "Forecasting Performance" operators are used to determine the performance of our testing dataset. The model is given in Fig. 1. Finally, to visualize the comparison between original close price and predicted close price of all nine constituents of the month January 2019, we have created graphs which are shown in Fig. 2(a)–(i). A small part of leaf of gradient boosted tree model is shown in supplement S2.

3.3.2. Predictive model: Neural net

An artificial neural network, also called a neural network, is a mathematical and computational model that is inspired by the structure and functional aspects of biological neural networks. A neural network consists of an interconnected group of artificial neurons, and it processes information using a connectionist approach to computation. Neural networks can be employed to model complex relationships between inputs and outputs or to find patterns in data to predict the price of cryptocurrencies [28,29].

We have considered the same model used in [25], the outlook of which is given in Fig. 3. We have given the historical prices of all seven attributes as input in the training dataset. In the testing dataset, we have provided the attributes of the month January 2019. Set role operator is used to set the attribute name as "Date" and target role as "id". A windowing operator is used, which will create examples from the value series data set by windowing the input data we have provided. The parameters of this operator are changed. The parameters "series representation" is used to represent the series values and it is set as "encode_series_by_examples". The parameter "window size" is the width of the used window and the "step size" is the distance between the first values. Both of these parameters are set to 1. Create single attribute and create label options are checked as we have the close price as our label and we are predicting the future close prices of the cryptocurrencies depending on the close prices of the past. On the other hand, in testing side no label attribute is created on another windowing operator as it will be predicted by the model. Horizon is the distance between the last window value and the value to predict. A sliding window validation operator is used to enclose sliding windows of training and tests in order to estimate the performance of a prediction operator. The parameters of these operator are also changed. The parameter "training window width" is the number of examples in the window, which is used for training and it is tuned to 4. The "training window step size" is the number of examples the window is moved after each iteration and it is tuned to 1. The "test window width" is the number of examples, which are used for testing and it is tuned to 4. Horizon is the increment from last training to first testing example and it is kept at default state i.e. 1. The sliding window validation operator is a subprocess, so it has two phases — training and testing. The training phase have been provided with a neural net model. The parameters of this operator are training cycles, learning rate and momentum. We have trained the model with 500 training cycles. Learning rate determines the change of weights at each step and it is tuned to 0.03. The momentum adds a fraction of the previous weight update to the current one as a result, it avoids local maxima and smoothen the optimization directions; it is tuned to 0.9. We have sorted our data by shuffling the input data before learning. The testing side has a "Apply Model" operator with a "Forecasting Performance" or "Performance (Regression)" operator. The value of the parameter 'Horizon' has been set to 1 and the parameter 'Main Criterion' has been set to 'first'. To visualize the difference between original and predicted close price of the constituents, we have shown comparative graphs in Fig. 4(a)–(i).

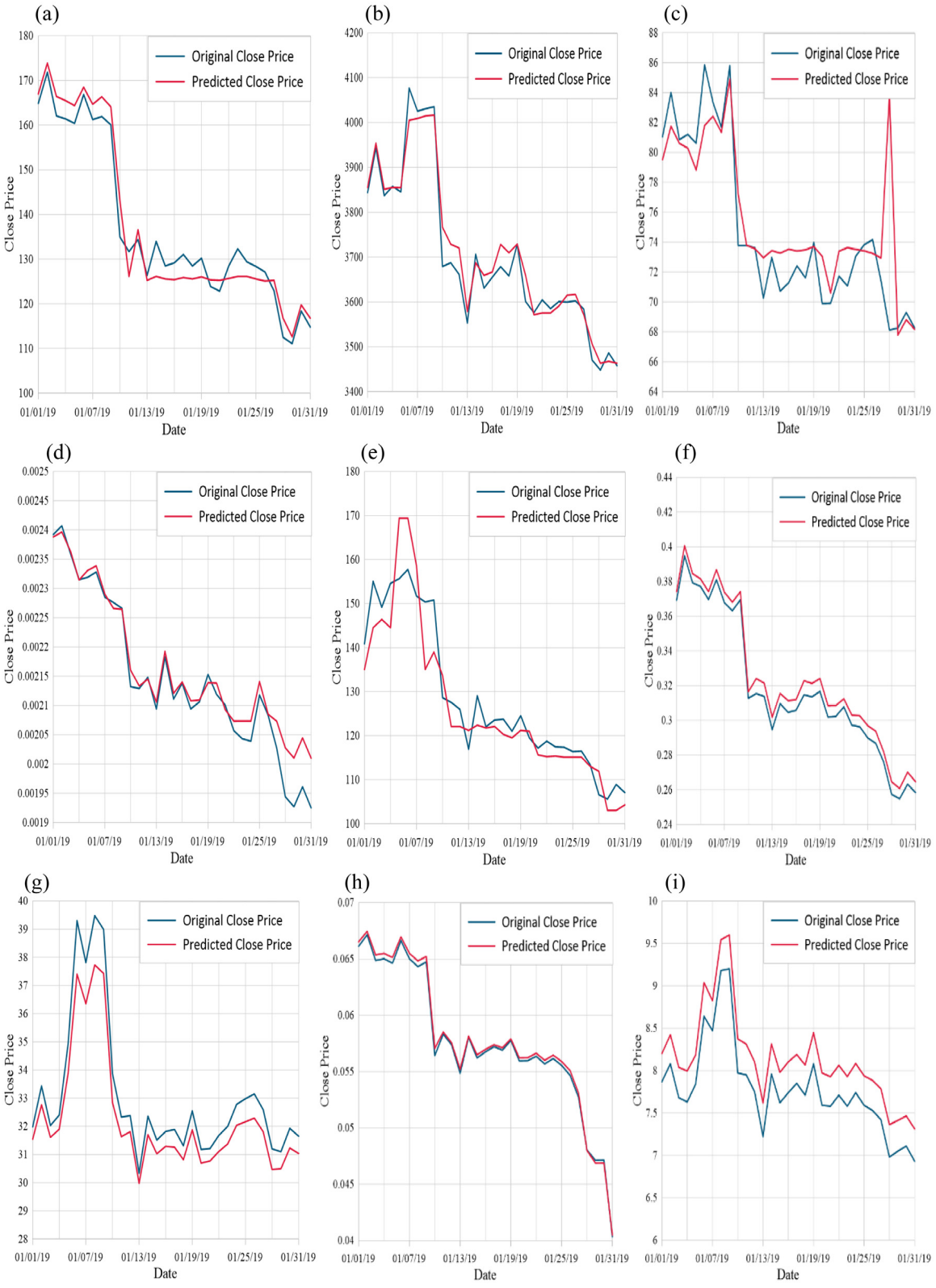

Fig. 2. Comparison between original and predicted close price obtained from RapidMiner using model gradient boosted trees for the month January 2019 (a) of constituent Bitcoin Cash. (b) of constituent Bitcoin. (c) of constituent Dash. (d) of constituent Dogecoin (DOGE). (e) of constituent Ethereum. (f) of constituent IOTA (MIOTA). (g) of constituent Litecoin. (h) of constituent NEM. (i) of constituent NEO.

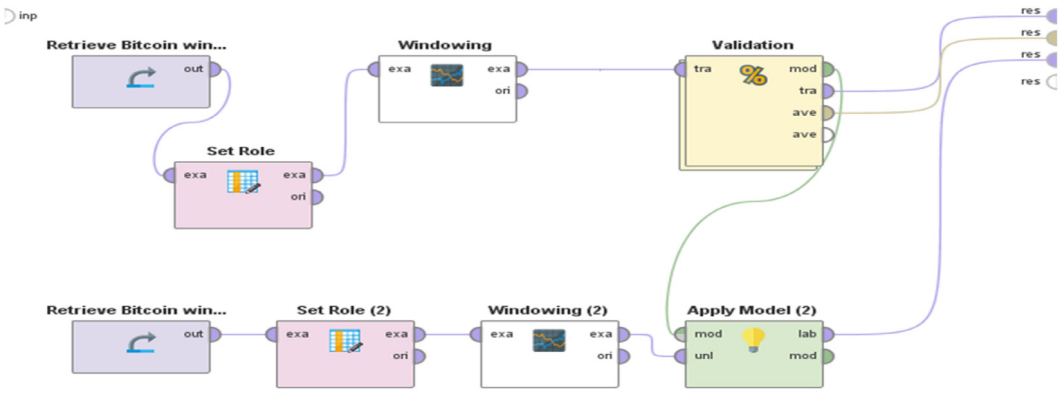

Fig. 3. Neural net model for predicting constituents and index of cryptocurrencies.

3.3.3. Predictive model: Ensemble learning method

In an ensemble learning method multiple machine learning algorithms or learners are strategically generated and combined together in order to solve one particular computational intelligence problem. Using ensemble learning method, we can construct set of models and combine them through voting process, as a result of which, the likelihood of an unfortunate selection of a bad model for prediction can be reduced. This method helps to overcome the biases and error rates of the individual (weak) models by combining them through creating a strong learner by uniting some weak learners. The training dataset plays the role of most effective contributor to the error in the model so we have taken a large training dataset with seven attributes and close price as the label attribute. In recent years, several ensemble learning techniques have been proposed to find and predict the prices of cryptocurrencies [12,30].

In this paper, we have created an ensemble learning method in order to get better results by employing different models together, which is shown in Fig. 5. As the "Split Data" operator produces the desired number of subsets of the given dataset so it is used to partition our data into subsets. The parameters of this operator are "partitions" and "sampling type". The "partitions" parameter is used to split our data for training and testing in the ratio of 0.6 and 0.4. The "sampling type" parameter is set as "linear sampling" to simply divide our dataset into partitions without changing the order of the examples. The "Vote" operator is a subprocess, which uses a majority vote for classification or the average for regression. We have given gradient boosted trees, neural net and relative regression operator inside the "Vote" operator. In the case of gradient boosted trees model, the number of trees parameter is set to 500 and other parameters are kept at default state. For neural net model, there are 500 training cycles, with the learning rate set to 0.3 and momentum set to 0.2. The relative regression operator learns a regression model for predictions relative to another attribute value. It is a meta regression learner and useful for time series predictions of dataset with large trends. Inside this learner, we have chosen a "Linear Regression" model and the parameters of linear regression are kept at default state. It is seen that, using multiple relative regression learner, it is possible to minimize the error and thus the desired output can be obtained through a voting process. Finally, an "Apply Model" operator is used to apply the models through the voting process with a "Performance (Regression)" operator or a "Forecasting Performance" operator to find the performance vectors/metrics. To visualize the difference between original and predicted close price of the constituents, we have shown comparative graphs in Fig. 6(a)–(i).

3.3.4. Predictive model: K-NN

The K-NN (k-Nearest Neighbor) algorithm is a non-parametric method that is based on comparing an unknown dataset with the K training examples, which are the nearest neighbors of the unknown example. This algorithm is basically used for generating a K-Nearest Neighbor model that can be used for either classification or regression. So, it can be easily understood that, K-NN algorithm can play a very important role in forecasting the price of constituents and of cryptocurrencies. Moreover, we have seen its use in the case of detecting covert cryptocurrency mining operations, which causes great losses to the organizations [31]. Besides, reverse K-NN method is also used for managing location data of IoT service providers and users based on blockchain with smart contract [32].

In this paper, we have used a model similar to a forecasting model presented in [33]. For forecasting the close price of these currencies using K-NN model, we have taken into account, the attributes "Date" and "Close Price" only. The model is shown in Fig. 7. We have considered $k = 3$ through optimization and we have also considered the "Weighted Vote" through which the distance values between the Examples are considered for the prediction. We have used the loop operator, which mainly a subprocess and it loops over the subprocess as often as it is specified in the parameter number of iterations. One parameter of this operator is "number of iterations", which specifies how frequently the subprocess will be executed. Number of iterations is tuned to 31. We have also used the "Predict Series" operator, which will create predictions for the ordered time series that we have provided as the label attribute in our dataset. Two parameters of

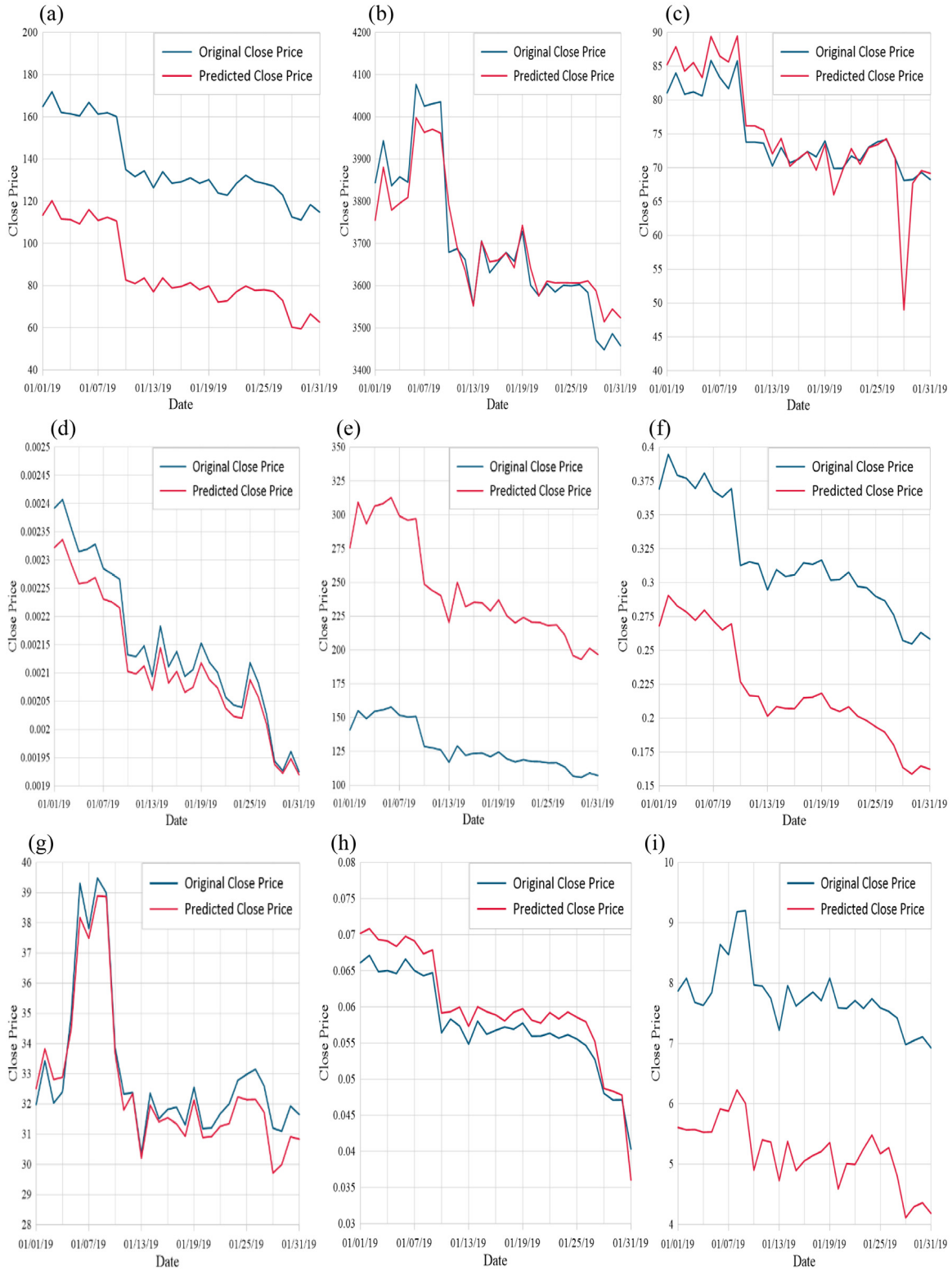

Fig. 4. Comparison between original and predicted close price obtained from RapidMiner using neural net model for the month January 2019 (a) of constituent Bitcoin Cash. (b) of constituent Bitcoin. (c) of constituent Dash. (d) of constituent Dogecoin (DOGE). (e) of constituent Ethereum. (f) of constituent IOTA (MIOTA). (g) of constituent Litecoin. (h) of constituent NEM. (i) of constituent NEO.

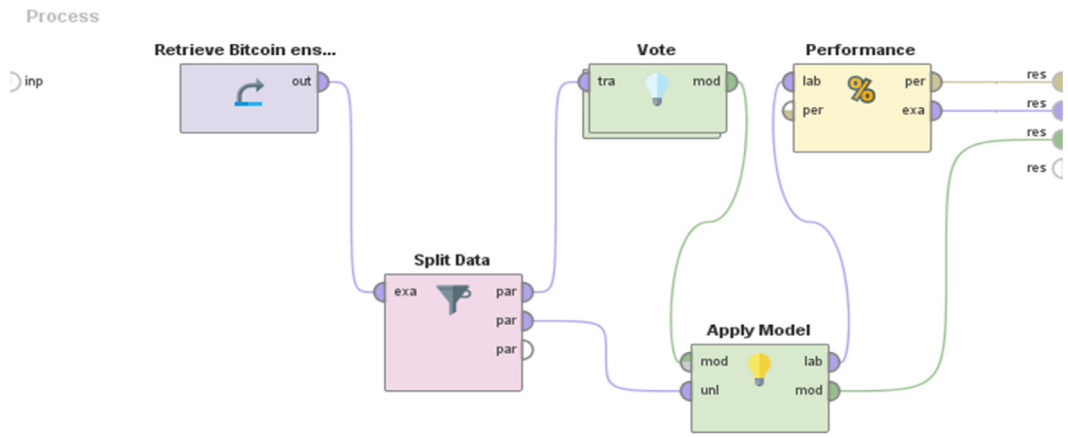

Fig. 5. Ensemble learning model for predicting constituents and index of cryptocurrencies.

Table 5
Residuals (average) of all the constituents and index using K-NN model.

Names of constituents	Residuals (average)
Bitcoin	−68.28510616239669
Bitcoin cash	−78.30352757348386
Dash	−9.896544294902897
Doge coin (DOGE)	3.489083182322858E−5
Ethereum	−26.347865050983437
IOTA (MIOTA)	−0.01377814670751103
Litecoin	0.968223740614247
NEM	−0.030791453489023235
NEO	−3.602597822081795
Index (cci30)	−186.59374626179212

this operator are "window width", "horizon". "Window width" parameter is the number of values used as indicators for predicting the target value, which is "Close" and it is tuned to 30. "Horizon" parameter is the gap size used between training windows and prediction value i.e. the number of days we want to predict in the future. It is equal to the "Number of Iterations" parameter of the loop operator. So, it is tuned to 31. We have used our model to forecast the close prices of the month of January 2019. In order to evaluate a time series model, a residual analysis has to be performed. It indicates the difference between our chosen forecasting method and actuals and its determination helps to observe how well the chosen forecast method fits to our historical data that we have provided. We know that, a good forecasting method will yield residuals which are uncorrelated and having mean near to zero. If the mean is not near to zero then, that means that the forecasts are biased. However, cryptocurrencies are correlated [10]. The list of residuals (average) of all cryptocurrencies and index is shown in Table 5. Fig. 8(a)–(i) illustrates the differences between the original and forecasted close prices of the month January 2019.

3.4. Index

In this paper, we have taken into account the "Cryptocurrencies Index 30 (cci30)", for prediction and forecasting from https://cci30.com. It is a rules-based index designed to objectively measure the overall growth, daily and long-term movement of the blockchain sector [34]. It works by tracking thirty largest cryptocurrencies market capitalization. It was launched on January 1st 2017 and it's starting value was arbitrarily set at 100 on January 1st 2015.

We have downloaded the OHLCV daily values of the index from [34]. We have predicted the daily close price of the index of the month of January 2019 using gradient boosted trees model, ensemble learning method, neural network model and for forecasting the daily close price of the month of January 2019, we have used K-NN algorithm in RapidMiner platform. The training and testing datasets are given in Table 6 for prediction, while we have used only training phase for K-NN model and forecasted the close prices of the month of January 2019. For predicting the close price of the index of the month January 2019 using RapidMiner, we have considered the attributes, which are given in Table 5, except volume and market capital. For forecasting we have used only the "Date" and "Close Price" attributes. The performance vectors of the models are shown in supplement S6 and the comparison graphs are shown in Fig. 9(a)–(d), which illustrates the comparison between original and predicted close price of the month January 2019 obtained from all models. This comparison is also shown including all models together in Fig. 10(j).

Fig. 6. Comparison between original and predicted close price obtained from RapidMiner using ensemble learning method for the month January 2019 (a) of constituent Bitcoin Cash. (b) of constituent Bitcoin. (c) of constituent Dash. (d) of constituent Dogecoin (DOGE). (e) of constituent Ethereum. (f) of constituent IOTA (MIOTA). (g) of constituent Litecoin. (h) of constituent NEM. (i) of constituent NEO.

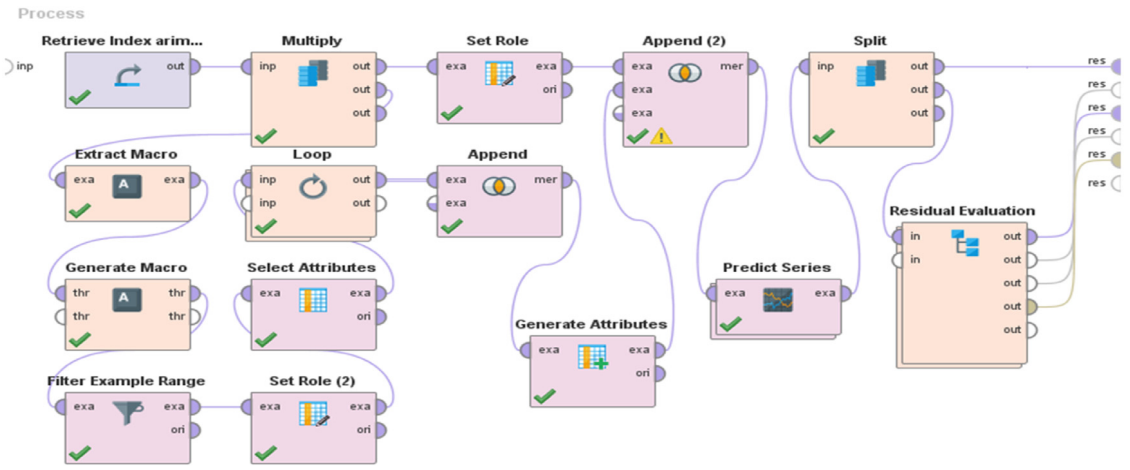

Fig. 7. K-NN model for forecasting constituents and index of cryptocurrencies.

Table 6
Training, testing and forecasting dataset of the index for gradient boosted trees, neural net, ensemble learning method and K-NN models.

Prediction and forecasting models	Training	Testing
Gradient boosted trees	01.01.2015–31.12.2018	01.01.2019–31.01.2019
Neural net	01.01.2015–31.12.2018	01.01.2019–31.01.2019
Ensemble	01.01.2015–31.01.2019	01.01.2019–31.01.2019
K–NN	01.01.2015–31.12.2019	01.01.2019–31.01.2019

Table 7
Comparison between state-of-the-art model in [15] and our model in this paper.

State of the art work							Work of this paper	
Forecasting performance in the first category of training sets								
Forecasting model	Period (Accuracy)			1st day of the period (accuracy)			Model	Accuracy
	2 months	2 weeks	2 days	2 months	2 weeks	2 days	Ensemble learning method	0.924
LightGBM model	0.776	0.881	0.762	0.762	0.905	0.548		
Forecasting performance in the second category of training sets				1st day of the period (accuracy)				
Forecasting model	2 weeks	2 days		2 weeks	2 days			
LightGBM model	0.607	0.476		0.952	0.93			

3.5. Comparison with previous results

This section presents a detailed comparison among the results obtained by the four models in this paper and other state of the art methods.

A Gradient Boosting Decision Tree (GBDT) algorithm, Light Gradient Boosting Machine (LightGBM) is adopted in [15] to forecast the price trend of cryptocurrency market, where they have argued that, the robustness of LightGBM model works better than SVM and RF model. Using LightGBM model, maximum forecasting performance in the first category of training sets has accuracy of 0.905 for two weeks and in second category of training sets, the accuracy is 0.952 for two weeks. In Table 7, we have shown a comparison between the results obtained in this paper and that in [15], which shows that our accuracy is 0.924 using ensemble learning method, whereas in [15], the highest accuracy is 0.952 using LightGBM model. However, they have forecasted the price trend (falling, not falling) by combining daily data of 42 kinds of cryptocurrencies with key economy indicators, but we have predicted and forecasted the price of 9 cryptocurrencies and cci30. In paper [15], only six months daily trading data from January 1, 2018 to June 30, 2018 were collected and considered for forecasting, but in our paper, we have taken yearly data, as a result of which our result is slightly less than in [15].

Deep learning techniques have been employed to forecast the price of Bitcoin, Digital Cash and Ripple in [16], where they have demonstrated that long-short term memory neural network (LSTM) performs better than generalized regression neural networks. The RMSE obtained using deep learning LSTM networks for Bitcoin, Digital Cash and Ripple are 2.75×10^3 19.2923 and 0.0499. In Table 8, we have shown a comparison between the results obtained in this paper and that in [16], which depicts that the models and learners we have used for prediction and forecast, yielded much better results than

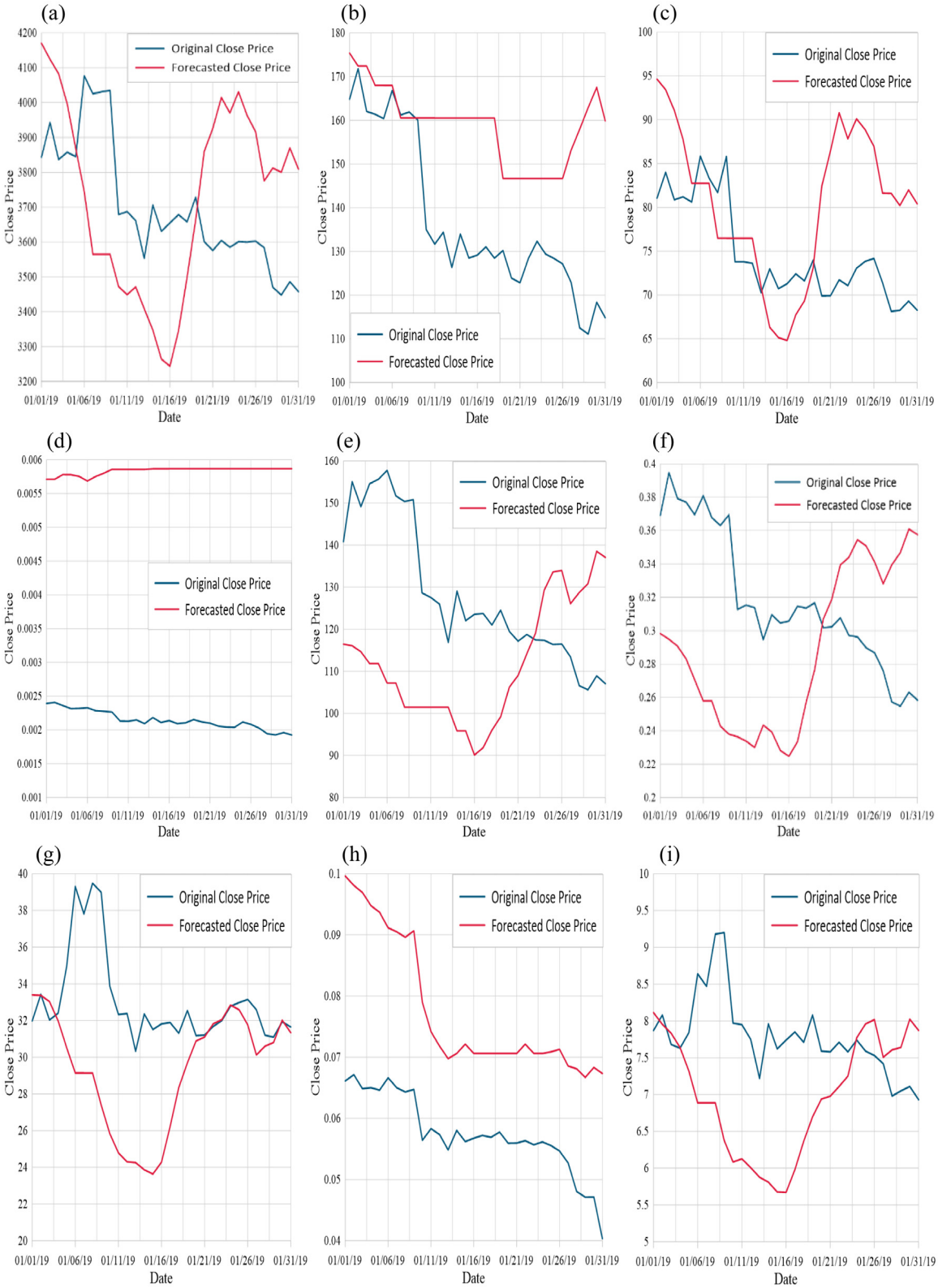

Fig. 8. Comparison between original and forecasted close price obtained from RapidMiner using K-NN model (a) of constituent Bitcoin. (b) of constituent Bitcoin Cash. (c) of constituent Dash. (d) of constituent Dogecoin (DOGE). (e) of constituent Ethereum. (f) of constituent IOTA (MIOTA). (g) of constituent Litecoin. (h) of constituent NEM. (i) of constituent NEO.

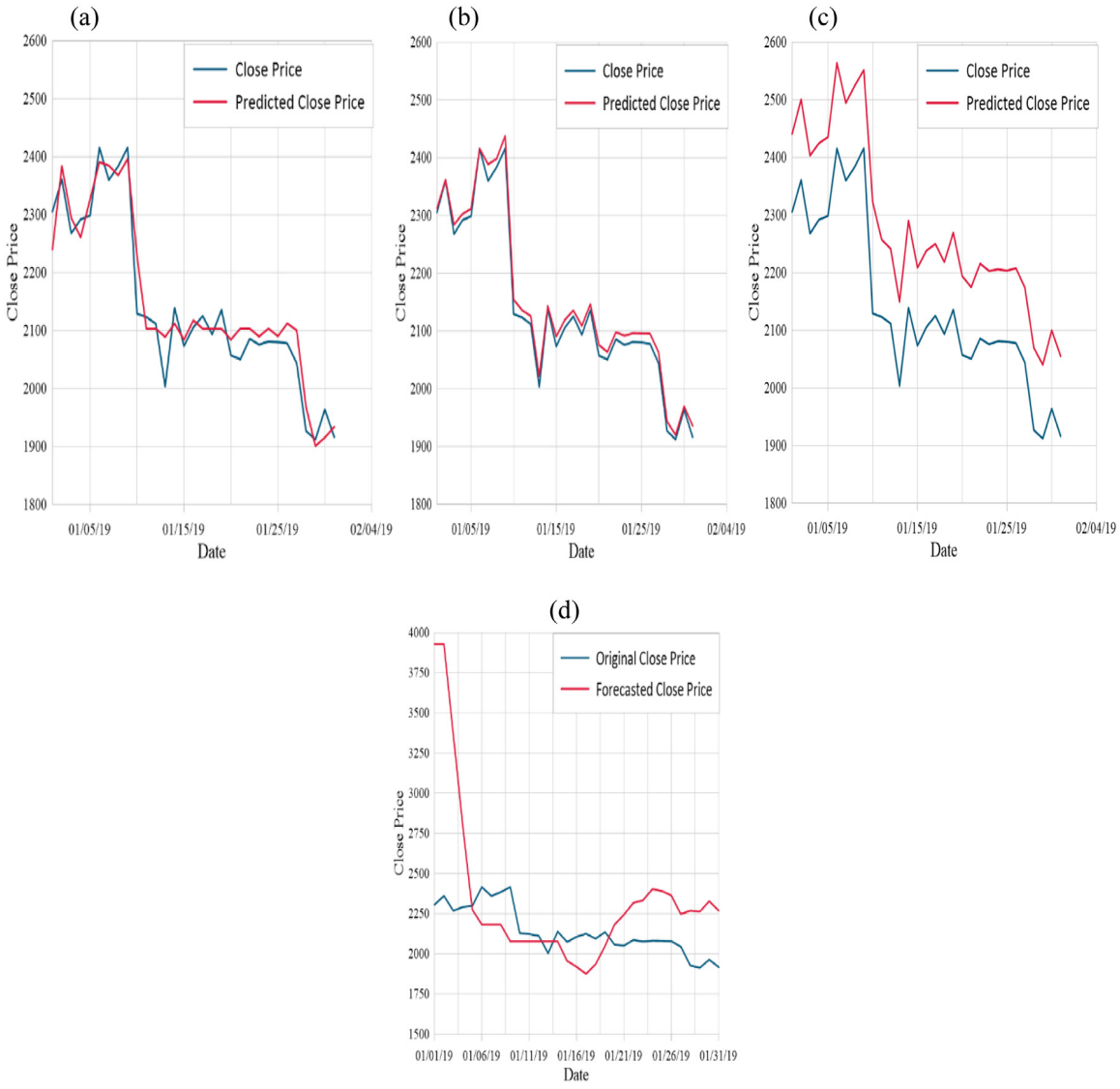

Fig. 9. Comparison between original and forecasted close price obtained from RapidMiner of index cci30 (a) using gradient boosted trees model. (b) using ensemble learning method. (c) using neural net model (d) using K-NN model.

Table 8
Comparison between state-of-the-art model in [16] and our model in this paper.

State of the art work			Work of this paper	
Currency	RMSE		Currency	RMSE
	DLNN	GRNN		Gradient boosted trees
Bitcoin	2.75 × 10^3	8.80 × 10^3	Bitcoin	32.863
Digital Cash	19.2923	50.2418	Doge coin	0.000
Ripple	0.0499	0.3115	NEM	0.001

the work done in [16]. The RMSE we have obtained by Gradient Boosted Trees for Bitcoin, Doge coin and NEM are 32.863, 0.000 and 0.001.

Non-linear deep learning methods (LSTM and RNN) and ARIMA model are employed in [20] to predict the direction of Bitcoin prices in USD with good accuracy. It has been demonstrated that, LSTM and RNN have performed better than ARIMA model. LSTM gives accuracy and RMSE of 52.78% and 6.87%, while RNN gives accuracy and RMSE of 50.25% and 5.45%; for ARIMA model, the values are 50.05% and 53.74% respectively. In Table 9, we have shown a comparison between the results obtained in this paper and that in [20]. It is seen by comparing that, our maximum accuracy and RMSE are

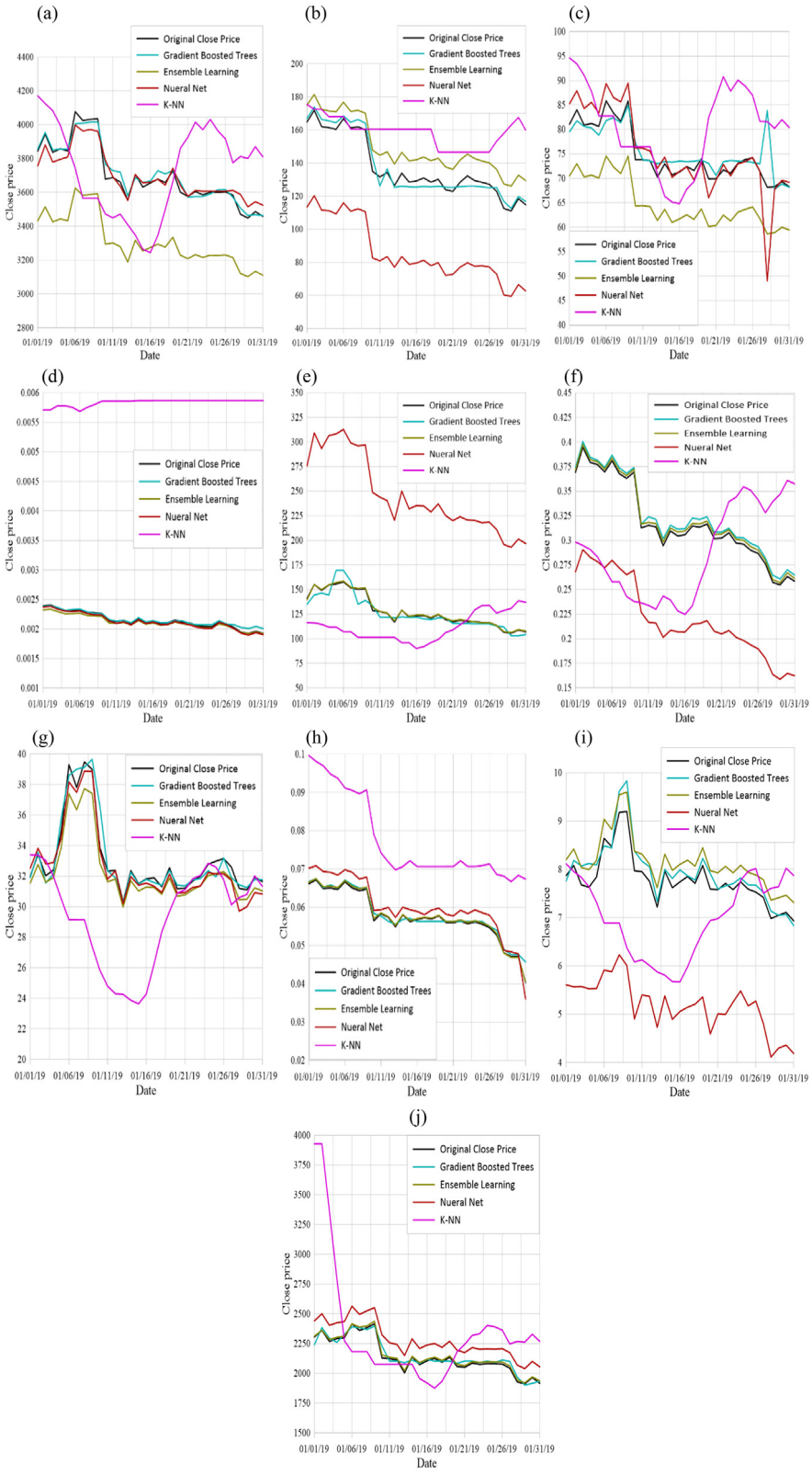

Fig. 10. Comparison between original and forecasted close price obtained from RapidMiner using all models and methods of the month January 2019 (a) of constituent Bitcoin. (b) of constituent Bitcoin Cash. (c) of constituent Dash. (d) of constituent Dogecoin (DOGE). (e) of constituent Ethereum. (f) of constituent IOTA (MIOTA). (g) of constituent Litecoin. (h) of constituent NEM. (i) of constituent NEO. (j) of index (cci30).

Table 9
Comparison between state-of-the-art model in [20] and our models in this paper.

State of the art work			Work of this paper		
Model	Accuracy	RMSE	Model	Accuracy	RMSE
LSTM	52.78%	6.87%	Gradient boosted trees	0.900	0.001
RNN ARIMA	50.25% 50.05%	5.45% 53.74%	Ensemble learning method	0.924	0.002

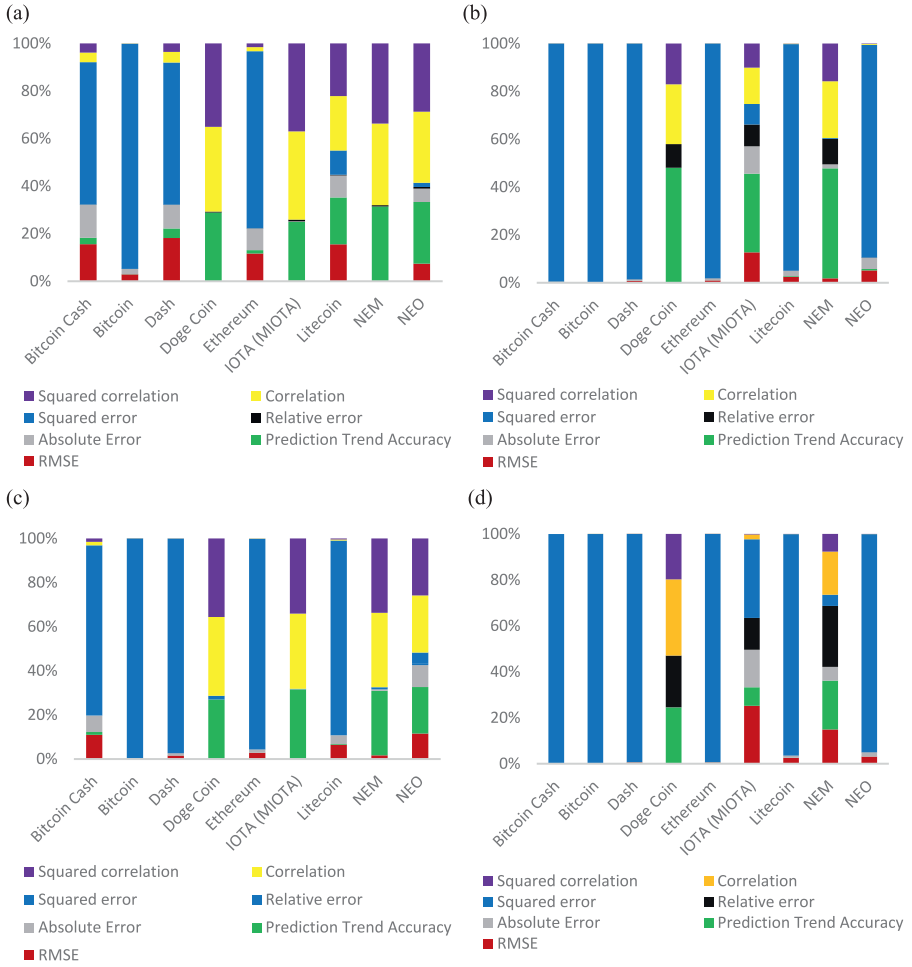

Chart 1. Performance measures of (a) Gradient Boosted Trees model (b) Neural Network model (c) Ensemble learning method (d) K-NN model.

much better than it is in paper [20]. Our highest accuracy and RMSE using gradient boosted trees model are 0.900 and 0.001, and 0.924 and 0.002 using ensemble learning method.

4. Results

Chart 1(a), (b), (c), (d) illustrates the performance of gradient boosted trees model, neural net model, ensemble learning method, and K-NN model using all the metrics mentioned in Section 3.2 for all nine constituents; the actual values for all four models have been provided in tabular format in supplement S1, S3, S4, and S5. Fig. 10(a)–(j) compares the predicted and forecasted values by all models for all nine constituents and cci30.

5. Conclusion

In this article, we have presented four different models to predict and forecast the close prices of nine constituents and cci30 using machine learning approaches. Our models exhibit a very good performance in overall prediction of the close

price of cryptocurrencies, which can be extremely useful for all including public, private, and government organizations. The reason for this is that through our models, the trends and patterns of these currencies can be well-understood. However, in the case of forecasting, the K-NN model has not worked very effectively, unlike other models; which has happened due to the presence of noisy random features and extreme volatility. We have compared our work with state-of-the-art models from the literature and demonstrated that our models' performance seems better and competitive. We have obtained 92.4% accuracy using ensemble learning method, which is considered as the best among all the models used in this paper. We believe and hope that our model will be beneficial for people to observe, understand, and choose their own desired currency from cryptocurrency market.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Reaz Chowdhury: Conceptualization, Formal analysis, Investigation, Methodology, Visualization, Software, Validation, Writing - original draft. **M. Arifur Rahman:** Conceptualization, Formal analysis, Investigation, Methodology, Visualization, Writing - review & editing. **M. Sohel Rahman:** Formal analysis, Methodology, Investigation, Software, Validation, Writing - review & editing. **M.R.C. Mahdy:** Conceptualization, Funding acquisition, Investigation, Visualization, Supervision, Project administration, Writing - original draft, Writing - review & editing.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.physa.2020.124569.

References

[1] Kim-Kwang Raymond Choo, Cryptocurrency and virtual currency: Corruption and money laundering/ terrorism financing risks? in: Handbook of Digital Currency Bitcoin, Innovation, Financial Instruments, and Big Data, Elsevier, 2015, pp. 283–307.
[2] L. Deng, J. Che, H. Chen, L. Zhang, Research on the pricing strategy of the cryptocurrency miner's market, Lecture Notes in Comput. Sci. (2018) 228–240.
[3] S. Nakamoto, Bitcoin: A peer-to-peer electronic cash system, 2008, Consulted.
[4] W. Zhang, P. Wang, X. Li, D. Shen, Some stylized facts of the cryptocurrency market, Appl. Econ. 50 (55) (2018) 5950–5965.
[5] Sally M. Gainsbury, Alex Blaszczynski, Gaming Law Review. Vol. 21, Mary Ann Liebert, Inc., 2017.
[6] J. Barkatullah, T. Hanke, Goldstrike 1: CoinTerra's first-generation cryptocurrency mining processor for Bitcoin, IEEE Micro 35 (2) (2015) 68–76.
[7] X. Li, C. Wang, The technology and economic determinants of cryptocurrency exchange rates: The case of Bitcoin, Decis. Support Syst. 95 (2017) 49–60.
[8] L. Cocco, G. Concas, M. Marchesi, Using an artificial financial market for studying a cryptocurrency market, J. Econ. Interact. Coord. 12 (2) (2015) 345–365.
[9] S. Ha, B. Moon, Finding attractive technical patterns in cryptocurrency markets, Memet. Comput. 10 (3) (2018) 301–306.
[10] K. Gkillas, S. Bekiros, C. Siriopoulos, Extreme correlation in cryptocurrency markets, SSRN Electron. J. (2018).
[11] D. Stosic, D. Stosic, T. Ludermir, T. Stosic, Collective behavior of cryptocurrency price changes, Phys. A 507 (2018) 499–509.
[12] L. Alessandretti, A. ElBahrawy, L. Aiello, A. Baronchelli, Anticipating cryptocurrency prices using machine learning, Complexity 2018 (2018) 1–16.
[13] H. Sun Yin, R. Vatrapu, A first estimation of the proportion of cybercriminal entities in the bitcoin ecosystem using supervised machine learning, in: 2017 IEEE International Conference on Big Data (Big Data), 2017.
[14] J. Han, J.C. Rodriguez, M. Beheshti, Diabetes data analysis and prediction model discovery using rapidminer, in: 2008 Second International Conference on Future Generation Communication and Networking, Hainan Island, 2008, pp. 96–99.
[15] X. Sun, M. Liu, Z. Sima, A novel cryptocurrency price trend forecasting model based on LightGBM, Finance Res. Lett. (2018).
[16] S. Lahmiri, S. Bekiros, Cryptocurrency forecasting with deep learning chaotic neural networks, Chaos Solitons Fractals 118 (2019) 35–40.
[17] Y.B. Kim, J.G. Kim, W. Kim, J.H. Im, T.H. Kim, S.J. Kang, C.H. Kim, Predicting fluctuations in cryptocurrency transactions based on user comments and replies, PLOS ONE 11 (8) (2016) e0161197.
[18] A. Greaves, B. Au, Using the Bitcoin Transaction Graph to Predict the Price of Bitcoin, 2019.
[19] A. Barnwal, H.P. Bharti, A. Ali, V. Singh, Stacking with neural network for cryptocurrency investment, (n.d.).
[20] S. McNally, J. Roche, S. Caton, Predicting the price of bitcoin using machine learning, in: 2018 26th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP), 2018.
[21] N.A. Bakar, S. Rosbi, Autoregressive integrated moving average (ARIMA) model for forecasting cryptocurrency exchange rate in high volatility environment: A new insight of Bitcoin transaction, Int. J. Adv. Eng. Res. Sci. 4 (11) (2017) 130–137.
[22] J. Rebane, I. Karlsson, S. Denic, P. Papapetrou, Seq2Seq RNNs and ARIMA models for cryptocurrency prediction: A comparative study, in: SIGKDD Fintech'18, Association for Computing Machinery (ACM), London, 2018.
[23] Zhengyao Jiang, Jinjun Liang, Cryptocurrency portfolio management with deep reinforcement learning, in: 2017 Intelligent Systems Conference (IntelliSys), 2017.
[24] C. Kai-Sang Leung, R. Kyle Mackinnon, Y. Wang, A machine learning approach for stock price prediction, in: IDEAS '14 Proceedings of the 18th International Database Engineering & Applications Symposium, ACM, Porto, New York, NY, USA, 2014, pp. 274–277, ©2014.
[25] T. k, M. Wadhawa, Analysis and comparison study of data mining algorithms using rapid miner, Int. J. Comput. Sci. Eng. Appl. 6 (1) (2016) 9–21.
[26] Haohua Sun Yin, Ravi Vatrapu, A first estimation of the proportion of cybercriminal entities in the bitcoin ecosystem using supervised machine learning, in: 2017 IEEE International Conference on Big Data (Big Data), 2017.

[27] T. Guo, A. Bifet, N. Antulov-Fantulin, Bitcoin volatility forecasting with a glimpse into buy and sell orders, in: 2018 IEEE International Conference on Data Mining (ICDM), 2018.
[28] H. Jang, J. Lee, An empirical study on modeling and prediction of Bitcoin prices with Bayesian neural networks based on blockchain information, IEEE Access 6 (2018) 5427–5437.
[29] N. Indera, I. Yassin, A. Zabidi, Z. Rizman, Non-linear autoregressive with exogeneous input (narx) bitcoin price prediction model using PSO-optimized parameters and moving average technical indicators, J. Fundam. Appl. Sci. 9 (3S) (2018) 791.
[30] T. Fischer, C. Krauss, A. Deinert, Statistical arbitrage in cryptocurrency markets, J. Risk Financ. Manage. 12 (1) (2019) 31.
[31] R. Tahir, M. Huzaifa, A. Das, M. Ahmad, C. Gunter, F. Zaffar, N. Borisov …, Mining on someone else's dime: Mitigating covert mining operations in clouds and enterprises, in: Research in Attacks, Intrusions, and Defenses, 2017, pp. 287–310.
[32] B.Y. Kim, S.S. Choi, J.W. Jang, Data managing and service exchanging on IoT service platform based on blockchain with smart contract and spatial data processing, in: Proceedings of the 2018 International Conference on Information Science and System - ICISS '18, 2018.
[33] Building a sales forecasting model with RapidMiner [Video file], 2017, Retrieved from https://www.youtube.com/watch?v=_mSRr2aIZuw.
[34] Cryptocurrency Index 30 | CCi30. (n.d.). Retrieved from https://cci30.com/.