2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT) | 978-1-7281-5791-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICAECT49130.2021.9392629

A Comparative Study of Different Machine Learning Algorithms on Bitcoin Value Prediction

Mayukh Samaddar*1, Rishiraj Roy*1, Sayantani De*1 and Raja Karmakar#2

*1*Department of Computer Science and Engineering, Techno International New Town, Kolkata, India, 700156*
2#Department of Information Technology, Techno International New Town, Kolkata, India, 700156
E-mail: mayukhmmm@gmail.com, riiishi3@gmail.com, sayantani.de@tict.edu.in, rkarmakar.tict@gmail.com

Abstract: **Machine learning is growing rapidly and has made many theoretical breakthroughs which find its application in many fields. Bitcoin is a very secure, decentralized, peer to peer currency with no third-party involvement. The price prediction of Bitcoin in the following years is a difficult task. The objective is to take a dig in the prediction of the future prices, dealing with real world data. A comparative study of the results produced by different machine learning models, along with graphs for epoch vs price, error and accuracy for each model using both linear and non-linear functions is done. We are using both neural network algorithms, such as artificial neural network (ANN), recurrent neural network (RNN) and convolutional neural network (CNN), as well as some famous supervised learning algorithms such as Random Forest (RF) and k-nearest neighbors (k-NN), to form the analysis. The time price prediction graphs and the epoch loss accuracy graphs are used for the analysis of each algorithm working on the same data and produces different results. Finally, the best suited algorithm are used for the prediction of future Bitcoin price.**

Keywords—Machine learning; Bitcoin; artificial neural network; convolutional neural network

I. Introduction

Bitcoin is a cryptocurrency invented in 2008 by a group under the alias of Satoshi Nakamoto. Its implementation was released as an open source software. It is a digital decentralized currency; thus, transactions are done without the need of an intermediary. Transactions are made via cryptography and a publicly distributed ledger called Blockchain is used to record transactions. The Blockchain computing is a technology where an entire set of records can be securely contained. The most unique thing about Blockchain is that copies of the entire set can be contained in millions of systems and any modifications done to the set are immediately reflected in each and every system which contains the set simultaneously (used to confirm transactions). Bitcoin is a very secure way of payment as compared to traditional means (credit cards) [5]. It is also a decentralized peer-to-peer currency, which means there are no third-party involvements. The prediction of future Bitcoin values has an important purpose. It allows an investor to make sound investments and even know the percentage of risk that is undertaken by making such an investment.

Before going into any further depth into Bitcoin value analysis, we should talk about the concept of machine learning. Machine learning is a study of computer algorithms which get better and more accurate with time. For example, let us talk about ourselves, we as humans, in some part of our life were null and void (had no formal education). We were taught by examples and our learning was tested through examinations. Machine learning is kind of similar in this regard. Here, we train the machine (for a particular job) through two sets of data. These are known as training data set and testing data set. We use mathematical models to operate on the "training dataset" so that it can learn from the data and can respond on its own to future data (for particular job) without explicit programming or user interface. There are some different approaches to machine learning, namely supervised learning, unsupervised learning, and reinforcement learning. Supervised learning is the way of making the computer learn by processing input data (along with the outputs) and using mathematical models to map the input to the output. Unsupervised learning is the way of providing the computer with some input data but it is the computer which has to read it and classify the data own its own (maybe find patterns in the data). Reinforcement learning deals with a dynamic environment, where it is given a specific job and it has to maximize its performance through feedbacks given in the particular program (while it tries to navigate through its program space). Some particular machine learning algorithms (MLs) are linear regression, logistic regression, k-nearest neighbors (k-NN), Random Forest (RF) and support vector regression [11].

In deep learning algorithms (DLs), several layers of neural networks are present. Data is simplified as it passes through each layer. Deep learning has many uses. The most common uses are image classification, visual recognition, natural language processing and so on. There are many deep learning algorithms in use. Some of them are artificial neural networks (ANN), convolutional neural networks (CNN) and recurrent neural network (RNN). Linear regression is a straightforward algorithm where we try to fit data points in a straight line and use the slope of the line to predict future values. It is very easy and useful. In Logistic regression algorithm, we use a logistic function to model a binary dependent variable. It can also be said as "binary regression". k nearest neighbor's algorithm takes up data and classifies it into a particular class depending on its k nearest neighbors [13]. It is mainly used in unsupervised learning and it uses distance function to perform the measurement. In RF, a large number of independent decision trees operating as an ensemble gives a class prediction (each tree gives a different prediction). The tree with the greatest number of votes is taken up as the classifier [10][14].

In regression method, we are given some data and the job is to train our computer using the data provided to get close to the actual results. For example, let us take the most common example of wind speed forecasting. In this case, there is no discrete value but a particular range of continuous values. Our work is to generate the wind speed from data as close as possible to actual data. We even use the difference between the actual value and predicted value (known as the error) to improve our algorithm. Regression analysis finds its use in every field like business and medicine. Meanwhile, classification is a process where the output has a discrete value. For example, data is processed, and we get a particular value (0,1 or similar values). Classification has a wide variety of applications starting from spam email classification to image segmentation, speech recognition to DNA sequence classification. Some widely used supervised learning algorithms are linear regression, k-nearest neighbor, gaussian naive bayes, decision trees, support vector machine (SVM) and Random Forest. Unsupervised learning is a method by which no supervision of the model is required. The model discovers patterns on its own and even finds out previously undiscovered information. Unsupervised learning allows users to perform much more complex operations than supervised learning. Clustering is the most important classification of unsupervised learning. Clustering is the task of grouping a set of data or objects into one or more groups in such a format so that the data or objects in a group is more similar to each other in ways and behavior than the data or objects in other groups. Some applications are, grouping of data (grouping newspaper articles according to their topics and displaying them) or finding anomaly points (finding outliers or other fraudulent transactions and activities). Some widely used unsupervised learning algorithms are k-means for clustering problems, Apriori algorithm for association rule learning problems, principal component analysis, singular value decomposition and independent component analysis.

Comparative study is the process of demonstrating ability to examine, compare and contrast one or more subjects or ideas. This type of analysis answers to how a system would react to perturbations of its parameters. Comparative studies are conducted everywhere with a wide range of parameters to understand which gives the most profit. Starting from comparative history to economics, it gives us an insight about which features of a particular thing are best so that they can be inculcated into one. Moreover, it is widely used in decision making. A simple “pros and cons list” used in business strategy making is also somewhat related to comparative analysis. A comparative study of machine learning on the same data is important as it gives us an insight of how a particular algorithm works on the same given data and it also helps us to decide the best suited algorithm for predicting Bitcoin price. In a comparative analysis, we tend to use a number of graphs to understand the benefits of each algorithm. The graphs that we use are dependent on the certain data variables but properties like epoch and error are used to compute and understand the general functioning of the algorithm. By studying those graphs, we can understand and decide which particular algorithm is best suited to be used. Every machine learning algorithm tends to generate a unique result for the same data set. So, in that case we have to examine all the results comparatively, get into the details and figure out why the algorithms give such results.

In this paper, we predict future Bitcoin values by using “real world” data and some proper machine learning algorithms as well as deep learning algorithms. The algorithms that we will be using for this analysis are random forest, artificial neural network, convolutional neural network, recurrent neural network and k-nearest neighbors [5]. Detailed analysis of Bitcoin data using these algorithms is performed in this paper. We generate a time-price graph for each machine learning algorithm and a united epoch vs loss and epoch vs accuracy graph for all the algorithms combined. This paper gives the details of each machine learning algorithm that we are using for the analysis of the result through graphs of each algorithm. A detailed analysis of the result is also done in this paper. Also, we analyze all the graphs (all the time vs price graphs and the epoch vs loss graph). These graphs tell us about the best suited algorithm for the prediction of future Bitcoin value.

II. Related Works

In [1], a good amount of details was given regarding CNN and it further elaborates us about the efficiency of neural network though it was only based on image classification. [4] gives an insight towards how we can use CNN and Long Short Term Memory (LSTM) for text classification problem but again it was based on classification not on regression. In [2], we get an idea of how to approach the problem we have with Bayesian and through RF though. It does not touch the neural networks but it does a good research on Bitcoin. In [1], they present how CNN is better than the traditionally used ANN for image recognition purpose, so its proves the superiority of CNN over ANN. As stated previously, [1] has not focused on the 1d part of CNN but more on the 2D part of CNN which is not much useful for determining future Bitcoin value. The [2] tries to predict the Bitcoin price by using Bayesian neural networks. In this paper, they compared several different compilers using Bayesian approach but for our comparative study we are not using Bayesian approach. In another paper [3], it uses data analysis to predict stock market value in both Python and R. Here, they compared the values of different types of loss calculating functions. The approach [3] uses similar to the approach we use, but here instead of comparing different loss functions we compare different machine learning algorithms. In case of [4], CNN and LSTM have been compared for their respective results for text classification. It proposes a text classification model namely NA-CNN-LSTM or NA-CNNCOIF-LSTM, which is kind of a combination of CNN and LSTM Layers. But [4] does not work with any other algorithms like RF. In paper [5], an attempt was made to predict Bitcoin prices accurately taking into consideration various parameters. For the first phase of this investigation the aim was to understand and identify daily trends in the Bitcoin market and for that bayesian regression and GLM RF regression were made. But this paper does not discuss about Deep learning algorithms. In [6], air quality was predicted using LSTM and Kalman filtering. This work gives an insight to Kalman filtering and how LSTM can be used of the prediction.

III. Overview of Different Training Algorithms

This section provides a brief description about ANN, CNN, RNN, RF and k-NN.

Artificial Neural Network:

The concept of ANN is inspired from the nervous system of our body [14]. It has a dense interconnection of simple elements (the neurons/nodes) which operate in parallel to achieve high performance speed [15]. The strength of connection between two nodes is given a value: inhibition (-1) or excitation (+1). There are three types of nodes: input, hidden and output. The input node takes numerical information and the input is normalized by min-max normalization. The nodes add the weighted input from other nodes in nearby layers and do a nonlinear transformation on that sum. The next layer nodes connected to this node get the output. The value flows through the network and the result is displayed in a meaningful way by the output layer [21]. It also supports backpropagation, where the error will be propagated backwards by adjusting to each node's weights.

Convolutional Neural Network:

CNN is being used in many fields for pattern or image processing. CNN reduces the parameters of ANN, and thus, it is useful in solving complex tasks. The features (of the problem) should be spatially independent [1]. First, low level features are identified which are then combined to find higher level patterns. The different layers in the CNN algorithm are Convolution layer [16], RELU layer [17], POOL layer [18] and SOFTMAX.

Recurrent Neural Network:

RNN is a special type of neural network where the output of the previous state is used as a feedback for the current state, rather than in normal neural networks where the input and output are independent of each other. In feed forward neural networks, there is just a single input layer, some hidden layers and an output layer [19]. In case of RNN, the hidden layer consists of a loop which feedbacks the current state. So, the state of hidden layer at any given time is dependent on previous and current input simultaneously [20].

k-Nearest Neighbour:

k-NN is an algorithm which can be used to solve both classification and regression problems. The basic function of the algorithm is to find the k number of nearest data points to the sample and use its neighbors' data to classify or predict the value of that sample [12]. k-NN is a non-parametric algorithm, so it does not have any presumptions on the data set. The choice of the value of "k" is an important factor as it determines the result. Greater the value of k, the boundary gets smoother [13].

RF:

RF is an algorithm to build a classification or regression ensemble. It is very useful and provides accurate results for data of high dimensions too [9]. RF algorithm produces a set of individual classifiers (learning to classify using individual randomly selected data set) and then sums up their predictions [8]. This algorithm is very beneficial in both classification and regression problems. It randomly selects different data points from the data sample and uses it to create a number of decision trees and getting their votes. For each observation, every tree predictor votes, and the majority is taken as the prediction [10].

III. Performance Analysis

Experiment Description:

Our experiment was carried out on Bitcoin Historical Data, available in Kaggle Our goal was to compare the accuracy with which it predicts the outcome for different machine learning algorithms. The following algorithms are used to test the dataset, CNN, RNN, RF, ANN and k-nearest neighbor. While doing the experiment, 73% of the data was taken as training data and rest 27% of the data was taken as testing data. In case of the neural networks, three different results were computed: 1) loss, 2) Accuracy and 3) Predicted Outcome. Other algorithms were run for 5 to 10 epochs, with the batch-size of 64. The activation function was Rectified Liner Unit (ReLU). The optimizer was Adam and the loss function was Mean Squared Error (MSE). The machine which was used for this experiment consists of 16 GB of RAM and Intel i5 8300h as CPU. From the data, the bitcoin price was predicted using the algorithms. While setting up the dataset, all the null value rows were dropped out for better prediction of the data. The time taken for each algorithm has been given below in the Table I for deep learning algorithms and Table II for other algorithms.

TABLE I. Model training time of DL algorithms

EPOCHS	CNN	RNN	ANN
	TIME (MINUTES)	*TIME (MINUTES)*	*TIME (MINUTES)*
5	30	425	7.5
6	36	510	9
7	42	595	10.5
8	48	680	13
9	54	765	14.5
10	60	850	16

TABLE II. Model training time of ML algorithms

Algorithm	Time (Minutes)
RF	20
k-NN	5

Dataset Analysis:

In this paper, we have used the dataset is Bitcoin Historical Data and it is available in Kaggle. The dataset has eight columns: timestamp, open, high, low, close, volume_(BTC), volume_(Currency) and weighted price. The dataset happens to have many null values, so before proceeding further, first the rows with null values were dropped for better prediction of the data and finally a set of data without any null value is obtained.

Preparing the Dataset:

After cleaning the dataset, it was prepared for fitting it into machine learning algorithms. At first, two more columns were removed that is weighted price and timestamp for it to be fitted to the algorithms. Then the data was scaled for better fitting with MinMaxScaler. The following rows were used for training the data: open, high, low, close, volume_ (Bitcoin price), and volume_(currency).

Training of the Model:

Analysis of CNN, RNN and ANN models for predicting Bitcoin price:

For training the Deep Learning algorithms, the prepared dataset was fed into two different sort of layers. First was the LSTM layers which are feeding the data into RNN model. Then for a different test, it was fed into maxpool, flatten and conv1D layer and the data is fed into the CNN model. Thus, after feeding the data to the neural networks, we found the following Bitcoin price vs time (timestamp) graphs. The different timestamp vs Bitcoin price predictions graphs are included in this section.

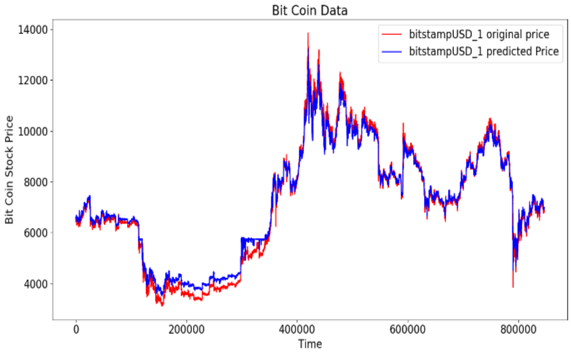

Fig. 1. Bitcoin price prediction by CNN

The Figure 1 shows the graph of CNN where we can see the price vs time graph. From this graph, it is quite clear that the model is quite sufficiently trained and has shown very good results. If we look at Figure 6, we find that the prediction of the accuracy by the CNN is around 99.7%, which means that it has shown an extremely good prediction, as well as the loss of only 0.000162046.

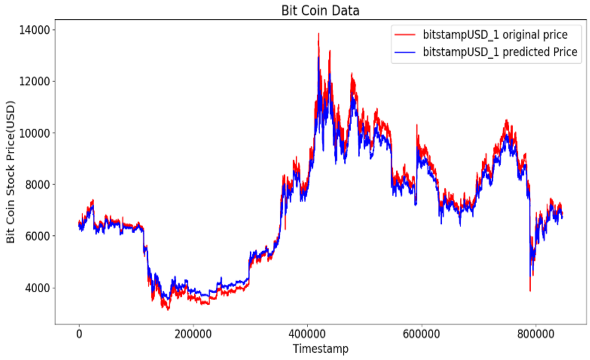

Fig. 2. Bitcoin price prediction by RNN

In Figure 2, the time price graph of RNN has been shown. It also predicts the data that can be seen in Fig. 6.

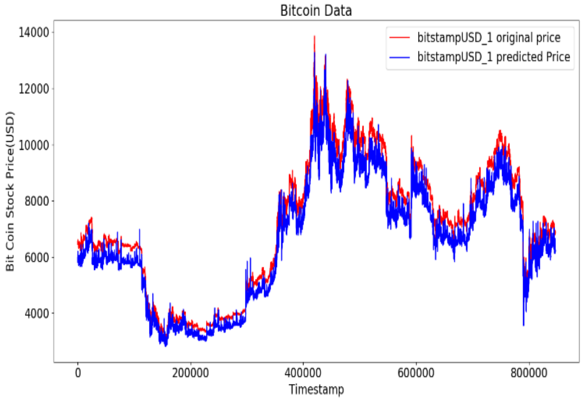

Fig. 3. Bitcoin price prediction by ANN

In Fig. 3, the data predicted by ANN is very good as well but if we compare it with the previous Fig. 1 and fig. 2, it shows a lot of miscalculations and wrong predictions, the accuracy shown by it is 91.8395%, which is relatively less than both CNN and RNN but never the less it also predicted very well with a loss of around 0.0740 which is higher than both CNN and RNN and can be seen at Figure 9.

Analysis of k-NN and RF models for predicting Bitcoin price:

For training RF and k-NN, the dataset was separated similarly but since there was no epoch required, Random Forest was trained in radial basis kernel function and k-NN was trained by taking the value of k as 5. The learning rate value is taken as the default (0.01).

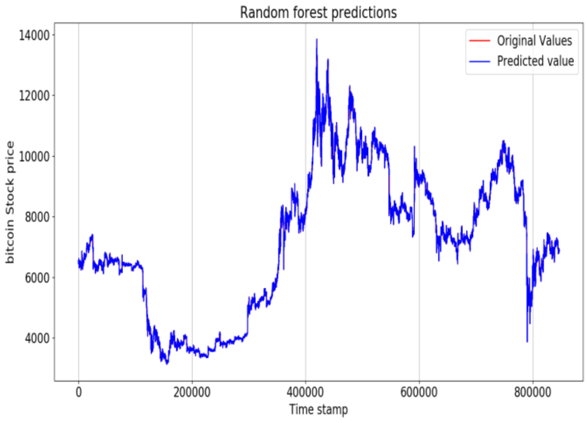

Fig. 4. Bitcoin price prediction by Random Forest

Now in Figure 4, we have seen that the prediction of the data is extremely well. From figure 6 it is seen that RF gives the accuracy of around 99.9957. The reason it predicted very well is because the open value itself was passed as a training sample, but it is not a compatible algorithm since it showed around 0.55797 loss, which is extremely high and can be seen in Figure 9.

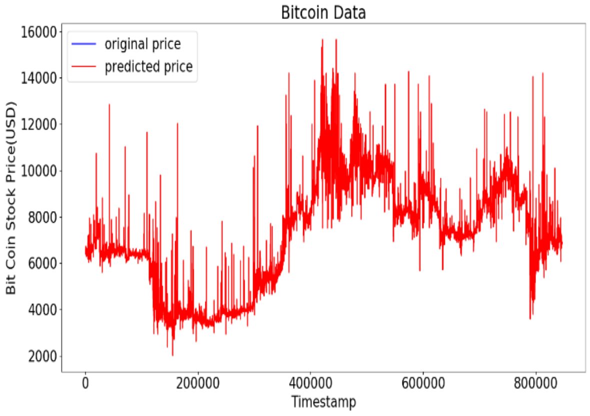

Fig. 5. Bitcoin price prediction by k-NN

In Figure 5, we find that the k-NN model has also predicted the data quiet well and the accuracy was over predicted with 100.02331% but from figure 9 it is seen that the loss was 0.6374 which is extremely higher and similar to Random Forest model.

Analysis of accuracy and loss:

The comparison between the different machine learning models were done by comparing two things, number one is the accuracy and number two is the loss. Accuracy is represented as the ratio between the sum of predicted values and the sum of original values multiplied by 100. The loss we have got in terms of mean squared error. We are representing the accuracy and loss comparisons through graphs and histograms. The above graphs help us about how well fitted the predicted data is with respect to the testing data.

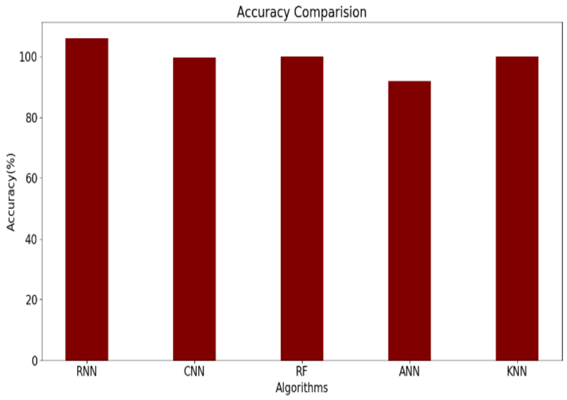

Fig. 6. Accuracy comparison histogram

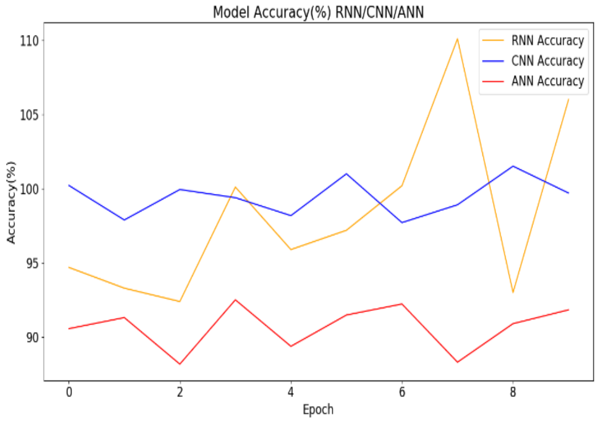

Fig. 7. Accuracy Comparison of artificial neural networks

In Figure 7, we see the comparison of the accuracies between CNN (blue), RNN (yellow) and ANN (red) over 5 to 10 epochs and we find that CNN is the most stable one as well as the most accurate one.

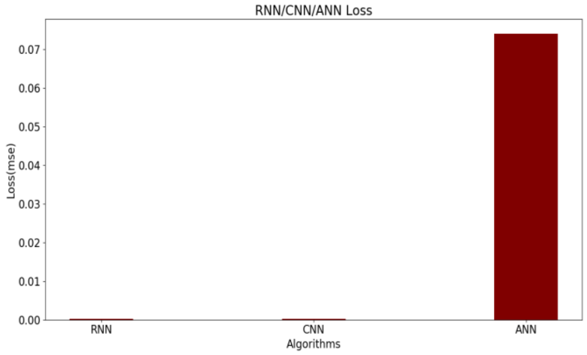

Fig. 8. Loss comparison histogram (RNN/CNN/ANN)

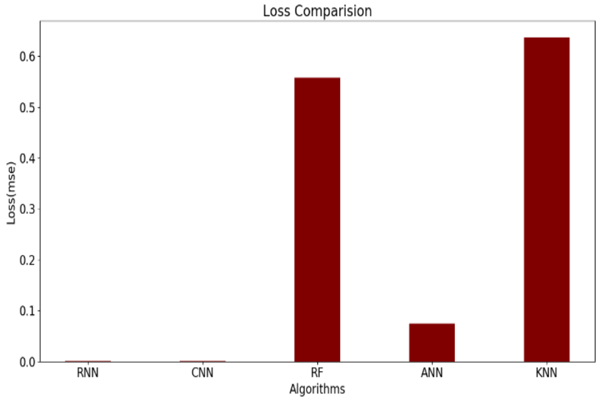

Fig. 9. Loss Comparison Histogram (All Algorithms)

After looking at all the figures and experimental results, we can conclude that CNN has shown the best results, as not only it has shown better accuracy, it has also shown the least loss, unlike other algorithms. We can see in Figure 6 that the accuracy shown by CNN, k-NN and RF is almost similar but as shown in Figure 9, we see that the losses we get shows us that k-NN and Random Forest have the highest and the second highest results, respectively but CNN has shown least loss. From the overall analysis we experimentally conclude that the CNN model has shown the best predictive capability followed by LSTM or RNN. But the non-deep learning model algorithms have shown less accuracy than the deep learning model algorithms.

IV. Conclusion

Nowadays machine learning is being used everywhere. The algorithms are improving from time to time with the advent of new tools and methods. So, in this paper the major contribution is to give a detail discussion about each and every algorithm used in this analysis. The null values in the dataset have not been considered and been left out. The prediction of Bitcoin values as done in this paper is really beneficial. In future the algorithms should be modified to achieve much more accurate results. The scaling of data can be improved and should be subjected for further research. Similarly, the data set size is a problem. Many studies don't have long, and detailed data sets and the values predicted from them become very inaccurate. So, studies have to be done to maximize accuracy in small datasets and the overall accuracy of prediction can be made a great by improving the algorithms.

References

[1] Albawi, Saad, Tareq Abed Mohammed, and Saad Al-Zawi. "Understanding of a convolutional neural network." In 2017 International Conference on Engineering and Technology (ICET), pp. 1-6. IEEE, 2017.

[2] Jang, Huisu, and Jaewook Lee. "An empirical study on modeling and prediction of bitcoin prices with bayesian neural networks based on blockchain information." Ieee Access 6 (2017): 5427-5437.

[3] S. Tiwari, A. Bharadwaj and S. Gupta, "Stock price prediction using data analytics," 2017 International Conference on Advances in Computing, Communication and Control (ICAC3), Mumbai, 2017, pp. 1-5, doi: 10.1109/ICAC3.2017.8318783.

[4] Luan, Yuandong, and Shaofu Lin. "Research on Text Classification Based on CNN and LSTM." In 2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), pp. 352-355. IEEE, 2019.

[5] S. Velankar, S. Valecha and S. Maji, "Bitcoin price prediction using machine learning," 2018 20th International Conference on Advanced Communication Technology (ICACT), Chuncheon-si Gangwon-do, Korea (South), 2018, pp. 144-147, doi: 10.23919/ICACT.2018.8323676.

[6] Song, Xijuan, Jijiang Huang, and Dawei Song. "Air quality prediction based on LSTM-Kalman model." In 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), pp. 695-699. IEEE, 2019.

[7] Yao, Wanye, Pu Huang, and Zhaoxin Jia. "Multidimensional LSTM networks to predict wind speed." In 2018 37th Chinese Control Conference (CCC), pp. 7493-7497. IEEE, 2018.

[8] Xu, Baoxun, Yunming Ye, and Lei Nie. "An improved random forest classifier for image classification." In 2012 IEEE International Conference on Information and Automation, pp. 795-800. IEEE, 2012.

[9] Bernard, Simon, Laurent Heutte, and Sébastien Adam. "On the selection of decision trees in random forests." In 2009 International Joint Conference on Neural Networks, pp. 302-307. IEEE, 2009.

[10] Kouzani, A. Z., Saeid Nahavandi, and K. Khoshmanesh. "Face classification by a random forest." In TENCON 2007-2007 IEEE Region 10 Conference, pp. 1-4. IEEE, 2007.

[11] C. Lopez-Martin, S. Banitaan, A. Garcia-Floriano And C. Yanez-Marquez, "Support Vector Regression For Predicting The Enhancement Duration Of Software Projects," 2017 16th Ieee International Conference On Machine Learning And Applications (Icmla), Cancun, 2017, Pp. 562-567, Doi: 10.1109/Icmla.2017.0-101.

[12] Chen, Ziqi, Bing Li, and Bo Han. "Improve regression accuracy by using an attribute weighted KNN approach." In 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), pp. 1838-1843. IEEE, 2017.

[13] Taunk, Kashvi, Sanjukta De, Srishti Verma, and Aleena Swetapadma. "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification." In 2019 International Conference on Intelligent Computing and Control Systems (ICCS), pp. 1255-1260. IEEE, 2019.

[14] Zhang, Xiang-Sun. "Introduction to artificial neural network." In Neural Networks in Optimization, pp. 83-93. Springer, Boston, MA, 2000.

[15] Singh, Yogesh, Pradeep Kumar Bhatia, and Omprakash Sangwan. "ANN model for predicting software function point metric." ACM SIGSOFT Software Engineering Notes 34, no. 1

(2009): 1-4.

[16] N. Jmour, S. Zayen And A. Abdelkrim, "Convolutional Neural Networks for Image Classification," 2018 International Conference On Advanced Systems And Electric Technologies (Ic_Aset), Hammamet, 2018, Pp. 397-402, Doi: 10.1109/Aset.2018.8379889.

[17] Wei, Qingdong, Fengjing Shao, and Ji Liu. "Research Summary of Convolution Neural Network in Image Recognition." In Proceedings of the International Conference on Data Processing and Applications, pp. 39-44. 2018.

[18] Sun, Yingyi, Wei Zhang, Hao Gu, Chao Liu, Sheng Hong, Wenhua Xu, Jie Yang, and Guan Gui. "Convolutional neural network based models for improving super-resolution imaging." Ieee Access 7 (2019): 43042-43051.

[19] Shi, Zejian, Minyong Shi, and Chunfang Li. "The prediction of character based on recurrent neural network language model." In 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS), pp. 613-616. IEEE, 2017.

[20] Yang, Tzu-Hsuan, Tzu-Hsuan Tseng, and Chia-Ping Chen. "Recurrent neural network-based language models with variation in net topology, language, and granularity." In 2016 International Conference on Asian Language Processing (IALP), pp. 71-74. IEEE, 2016.

[21] Mishra, Manish, and Monika Srivastava. "A view of artificial neural network." In 2014 International Conference on Advances in Engineering & Technology Research (ICAETR-2014), pp. 1-3. IEEE, 2014.