CRYPTOCURRENCY PRICE PREDICTION BASED ON MACHINE LEARNING

[1]Mr.R.Dhamodharan, [2]Ann Theresa Tom, [3]Hemalatha M, [4]Leelavathi V, [5]Lesa S
[1]AP/CSE, [2,3,4,5]UG Scholar
[1,2,3,4,5]Department of CSE
Vivekanandha College of Engineering for Women, Namakkal, India
[1]dhamodharan@vcew.ac.in, [2]anntheresatom6@gmail.com, [3]lhema6505@gmail.com
[4]leelavathi572002@gmail.com, [5]lesasekar2000@gmail.com

***Abstract*—In this study, aiming at the problem that the price of Bitcoin varies greatly and is difficult to predict, a hybrid neural network model based on convolutional neural network (CNN) and long short-term memory (LSTM) neural network is proposed. The transaction data of Bitcoin itself, as well as external information, such as macroeconomic variables and investor attention, are taken as input. Firstly, CNN is used for feature extraction. Then the feature vectors are input into LSTM for training and forecasting the short-term price of Bitcoin. The result shows that the CNN-LSTM hybrid neural network can effectively improve the accuracy of value prediction and direction prediction compared with the single structure neural network. The finding has important implications for researchers and investors in the digital currencies.**

I. INTRODUCTION

Nowadays, electronic payment is mainly based on traditional electronic payment tools, while digital currency relies on the virtual currency which is on the basis of the blockchain technology. Compared with traditional currency and electronic trading, it has a fast transaction speed but low cost as it uses a decentralised peer-to-peer network and does not require the third-party payment platform. What is more, the transaction is more secure and transparent because it is difficult to crack or forge by using encryption algorithms and automatic authentication mechanisms. Consequently, since the birth of digital currency, it has quickly gained widespread attention.In particular, Bitcoin is the world's first distributed super-sovereign digital currency. It was proposed and built by a Japanese programmer Satoshi Nakamoto in 2009, relying on an electronic payment system based on encryption technology and P2P (Point to Point) technology. In April 2010, Bitcoin was first publicly traded at a price of only $0.03 and in January 2013 the price still did not exceed $15 with an intermittent price peak close to $1000 in December of the same year. Moreover, in December 2017, the price reached a new high of $19,000 but then fell sharply to $6700 in February 2018 and even reached at 3200 in December at the same year. It can be seen that the price of Bitcoin has climbed all the way, butgradually declined, fluctuating greatly. Bitcoin also carries out seven 24-h round-the-clock trading, non-government-led, having no price limit, consequently, the price trend is obviously characterized by skyrocketing and plunging. As a new type of investment tool, its frequent price changes naturally raise a question: whether the price of Bitcoin can be predicted. This is a significant question, especially given Bitcoin's short history and the fact that its price is easily influenced by the attitudes of governments around the world. This paper studies whether the price of Bitcoin can be predicted based on internal information such as the historical price of Bitcoin and external information such as market factors. The data used for prediction is not only limited to the historical transaction information of Bitcoin, but also includes the macroeconomic variables and investor attention towards Bitcoin. At the same time, artificial intelligence technology is introduced into Bitcoin price prediction. In this paper, convolutional neural network (CNN) is used to extract the characteristics that have a great influence on Bitcoin price in the data set, and then long short-term memory (LSTM) is used for price forecasting. A CNN-LSTM hybrid neural network-based price prediction model is proposed. The current research focuses on the accuracy and direction of Bitcoin price forecasting.

II. LITERATURE SURVEY

Early research on bitcoin debated if it was in fact another type of currency or a pure speculative asset, with the majority of the authors supporting this last view on the grounds of its high volatility, extreme short-run returns, and bubble-like price behavior (see e.g., Yermack 2015; Dwyer 2015; Cheung et al. 2015; Cheah and Fry 2015). This claim has been shifted to other well-implemented cryptocurrencies such as Ethereum, Litecoin, and Ripple (see e.g., Gkillas and Katsiampa 2018; Catania et al. 2018; Corbet et al. 2018a; Charfeddine and Mauchi 2019). The opinion that cryptocurrencies are pure speculative assets without any intrinsic value led to an investigation on the possible relationships with macroeconomic and financial variables, and on other price determinants in the investor's behavioral sphere. These determinants have been shown to be highly important even for more traditional markets. For instance, Wen et al. (2019) highlight that Chinese firms with higher retail investor attention tend to have a lower stock price crash risk. That bitcoin prices are mainly driven bypublic recognition, as Li and Wang (2017) call it—measured by social media news, Google searches, Wikipedia views, Tweets, or comments in Facebook or specialized forums—was also investigated in the case of other cryptocurrencies. For instance, Kim et al. (2016) consider user comments and replies in online cryptocurrency communities to predict changes in the daily prices and transactions of Bitcoin, Ethereum, and Ripple, with positive results, especially for bitcoin. Phillips and Gorse (2017) use hidden Markov models based on online social media indicators to devise successful trading strategies on several cryptocurrencies. Corbet et al. (2018b) find that bitcoin, ripple, and Litecoin are unrelated to several economic and financial variables in the time and frequency domains. Sovbetov (2018) shows that factors such as market beta, trading volume, volatility, and attractiveness influence the weekly prices of Bitcoin, Ethereum, Dash, Litecoin, and Monero. Phillips and Gorse (2018) investigate if the relationships between online and social media factors and the prices of Bitcoin, Ethereum, Litecoin, and Monero depend on the market regime; they find that medium-term positive correlations strengthen significantly during bubble-like regimes, while short-term relationships appear to be caused by particular market events, such as hacks or security breaches.

III. PROPOSED METHODOLOGY

Sources of information can be divided into internal information (different parameters of Bitcoin) and external information (macroeconomic factors and investor attention).Among them, the internal information includes the opening price, the highest price, the lowest price, the closing price, the trading volume, and the transaction amount of Bitcoin. They are derived from Kraken Bitcoin exchange trading data provided by Quandl (**https://www.quandl.com/**). Huang *et al*. declare that technical indicators can be used to predict the price of Bitcoin, so this paper chooses three technical indicators: relative strength index (RSI), money flow index (MFI) and on balance volume (OBV). As a new type of investment tool, the price of Bitcoin is considered to be related to macroeconomic variables. Hence, the following indicators are selected in this paper: crude oil futures price, gold price, S&P 500 Index, NYSE Index, NASDAQ Index, Federal Funds Rate, and Yuan–Dollar exchange rate. The data are all from the Wind Database. According to [6], the long prediction period may lead to enormous prediction error. So the forecast period used in this paper is 3 days, namely, the characteristic parameter data of the previous 3 days is used to predict the closing price of Bitcoin on the fourth day. Each set of samples in the constructed data set has a 51-dimensional feature that contains all the features of the previous 3 trading days. And the time interval of all data in this paper is considered, ranging from 30 December 2016 to 31 August 2018. The first 588 samples are used as the training set for model training, and the last 20 samples are applied in the test set to verify the feasibility of the model.

1)Data Preprocessing: Attribute such as: price of open, high, low, close, adjusted close price taken from huge dataset are fed as input to the models for training to pre-process the data techniques like normalization & one hot encoding in applied on dataset.

2)Data Gathering: Daily data for the four channels hasbeen monitored since 2013. First, the Bitcoin price history, from which it is extracted. The coin market is the head of the market with its open API. Second, data from Blockchain included, especially we prefer standard block size, user address number, the amount of production, and the number of miners. We find it objectionable to have Blockchain data, given the endless measurement problem, on the other hand, the number of accounts, by definition related in price movements, as the number of accounts increases, it could mean more transactions that take place (perhaps by exchanging different parties and not just by transferring Bitcoins to another address), or by signalling more users joining the network. Third, in the emotional details, we find that over time the term 'Bitcoin' was used by the Py Trends library. Finally, two indicators are considered, those of S&P 500 and Dow and Jones. Both are refundable Yahoo Finance API. In total, this makes 12 features. Pearson interaction between symptoms is shown in Figure 2. Some attributes do not exist closely related, for example, financial indicators suitable for each other, but not for any of the attributes associated with bitcoin. Also, we see that Google Trends are related to Bitcoin transactions.

*3)Data Cleansing:*From exchange data, we look only at the related Volume, Close, Unlock, higher prices, and market capitalization. In all data sets, if the NaN values are found to be correct there, it is replaced by a description of the appropriate attribute. After this, all data sets are merged into one, according to the magnitude of the time. If we look at the Bitcoin price movement during the period from 2013 to 2014, we have seen fit to remove data points prior to 2014, which is why the details that will be transferred to the neural network are dormant from 2014 to September 2018.

*4)Data Normalization:*Deciding how to get used to the timeline, especially finance is by no means easy. What else, as a sixth rule, the neural network must load data taking large amounts of different data (referring to different time series scales, such as exchange rate, and Google Trends). Doing so can create major gradient updates that will prevent the network from changing. Doing reading easy on the network, data should have the following features:

Min-MaxScaling, where the data inputs are mapped on a number from 0 to 1:x’=x-min(X)/max(X)-min(X)

IV. CNN-LSTM HYBRID NEURAL NETWORK

CNN model is one of the most typical and widely used ANN in recent years. It mimics the perception of local information by biological vision cells. Local connection and layer-by-layer calculation are used to extract the data features, and finally, the global information is synthesised through the full connection. Its basic structure includes convolution layer, pooling layer and full connection layer.The convolution layer uses a convolution operation instead of a matrix multiplication, and each convolution kernel can extract a feature of the input data. The weight sharing method is adopted in the convolution operation, which effectively reduces the number of parameters, decreases the complexity of neural network training, and improves the training speed. The pooling layer can reduce the dimension of input data and the size of data volume. Commonly used pooling methods include mean pooling, maximum pooling etc. The formula is as follows:

$$y_j^l = f\left(\sum_{i \in M_j} x_i^{l-1} \boldsymbol{k}_j + b_j^l\right) \quad (2)$$

$$x_j^l = g(\beta_j^l \text{pooling}(x_j^{l-1}) + b_j^l)$$

LSTM is an improved RNN model, which can effectively solve the problem of gradient disappearance and gradient explosion in the RNN model. It is suitable for processing long-term sequence data and solving long-term dependence. Its basic unit is the memory module, containing the memory unit and three gates controlling the memory unit, namely Input Gate, Output Gate and Forget Gate. The gate is the structure that determines the selective passage of information. If the output value of the sigmoid function is 0, it is discarded completely, while if it is 1, it passes completely. Fig. 1 shows the basic unit of LSTM neural network.

(i) *Forget Gate:* The purpose of Forget Gate is to determine what information will be discarded. Reading in the output of the previous layer h_{t-1} with the current input x_t, the gate outputs f_t and assigns the current cell C_{t-1}, the calculation formula of f_t is as follows:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad (4)$$

where σ presents the sigmoid function.

(ii) *Input Gate:* The role is to update based on existing information. First, run the sigmoid function to get i_t and decide which values to enter. Then, according to the tanh function, a candidate value vector $\widetilde{C}_t$ is obtained, which is multiplied with i_t and added to the state C_t. The formula for this part is as follows:

$$i_t = (W_i x_t + U_i h_{t-1} + b_i) \quad (5)$$

$$\widetilde{C}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \quad (6)$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \widetilde{C}_t \quad (7)$$

(iii) *Output Gate:* Output the information of the current point. After running a sigmoid function to get o_t and determining which parts will be output, C_t is processed by the tanh function to obtain a value between −1 and 1. Finally, the value is multiplied with o_t to decide the ultimate output:

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad (8)$$

$$h_t = o_t \cdot \tanh(C_t)$$

Here the data is the above data collected from(.CSV)is loaded as data set matrices and normalization isperformedandthereafteritgoeswithsplittingofdataandimprovethembytraining.Afterawhileoftrainingit should be reshaped as X and Y where x = t and Y =t+1.Astherearedifferenttypesofmodelswithdifferentfeaturesweneedtoselecta Modelswhichdoes major amount of work and suits our way of solving a problemso we use LSTMmodel.

V. CONCLUSION

In this paper, several approaches for crypto currencies like Bitcoin price prediction were investigated. We compared the results of prediction with Linear Regression, Linear Regression with Features, and Recurrent Neural Networks with LSTM cells. The research contribution of this technique is that we predicted a numerical value of price instead of performing binary classification, as well as used multiple features to train the model. The LSTM method performed notably better than the other two approaches, and we believe that further research on using Neural Networks for time-series prediction is very promising to financial data analytics and other fields. Our work can be extended further using some of the approaches described in the Related Work section. Namely, the LSTM-based model can be used as a part of the autonomous trading agent. It is worth investigating the scalability of our proposed approach. In particular, the important questions for further research are how far into the future should the price be predicted, and how many Bitcoins should be traded at a time by the autonomous agent. The LSTM model, as well as the autonomous agent-based on it, can be further enhanced with sentiment analysis. Historical sentiments from Twitter, the number of search queries from Wikipedia and Google, and other metrics reflecting the public interest in Bitcoin can be used to influence the weights during model training. Moreover, the current sentiments can be combined with the prediction of the LSTM model to influence an autonomous trading agent's decision whether to buy or sell Bitcoins at a given moment of time. Based on our research, it can be concluded that the future direction of work on predicting prices of stocks and crypto currencies has to take multiple metrics and features into account. Econometric data and sentiment analysis were only incorporated into projects that used simpler algorithms, such as Bayesian Regression. However, the projects involving DNNs or RNNs with LSTMs were trained on the datasets containing only the historical prices of stocks. This attests to the novelty of our work: we used multiple features in training the Neural Network, which is something not found in the research literature. Incorporating multiple features into prediction methods with these types of Neural Networks has the potential to produce more accurate results.

VI.RESULTSAND DISCUSSION

To make the model learn we need to focus on LSTM & RNN to allow identification of smaller sequence patterned data and the price of the next day is predicted. A python library known as ”Keras” utilized for the creation of neural network uses TensorFlow backend.

The layers are as follows:

- Input layer (takes data of shape n samples x 50 x 37)
- Bidirectional LSTM layer(returnsasequence,100 cells)
- Dropout layer (20% dropout — reduces overfitting)
- Bidirectional LSTM layer (returns a sequence, 100 cells)
- Dropout layer (20% dropout — reduces overfitting)Bidirectional LSTM layer (doesn’t return a sequence, 50 cells)
- Output layer (returns the predicted next day price of Bitcoin)

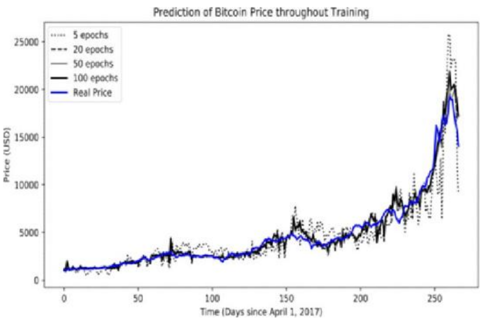

Fig.1. Bitcoin Prediction

1)Training

The training of the data needs a large amount of filtered data sets. When data reaches a stabilization point it doesn’t increases it results efficiency so by using some function we can improve it by loss function, activation function, optimizer.

2)Linear Regression

The main concept in our project is prediction and so to build a relationship between those models we use Linear

Regression to draw a relationships between those data objects. Linear relationship among target as well as predictors is found using Linear regression.

When it is possible to express a variable accurately by the other, then the relationship between them are known as deterministic. For instance, temperature can be predicted in terms of Fahrenheit when it is provided in degree Celsius.

Relationship among variables cannot be determined accurately using statistical relationship. The relationship between weight as well as height can be considered as an example.

The main objective is to attain a line that best fits the data. A line is said to be best fit when it possesses a smaller total prediction error. Distance between points in regression line is called as error..

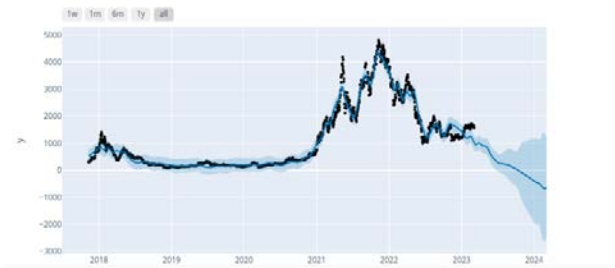

Fig.2. Forecast report of bitcoin

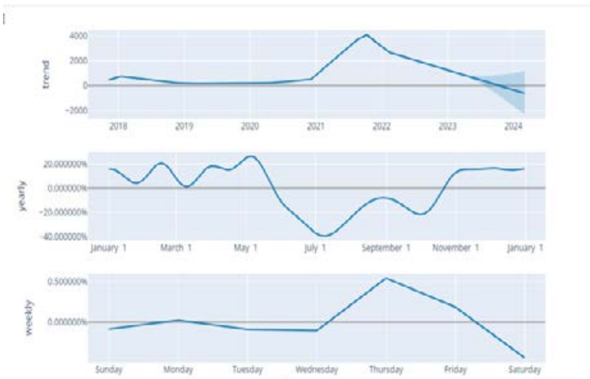

Fig.3. Bitcoin datasets of one year

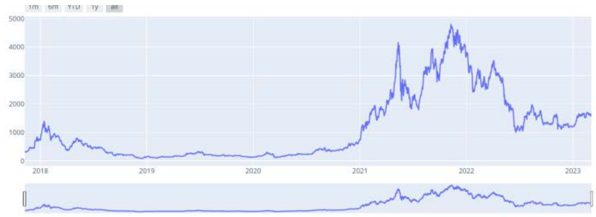

Fig .4. Time series plots of Ethereum open source

VII. FUTURE ENHANCEMENT

Crypto currencies, the newest class of digital assets, is obviously one of the top technological disruptions in recent times. The domain of crypto currencies has been useful in driving plausible improvements, especially in 2021, with many notable improvements. Crypto currencies have been through three essential ingredients of adoption, innovation, and integration for achieving stupendous growth. However, it is also important to think of the future of crypto currency and what it holds for everyone. What are the top trends you can expect in the domain of cryptocurrencies next year? Experts point out the possibility of 2022 being a slow year after the massive growth in 2021. The interest of people in crypto has doubled by huge margins in the last year. Interestingly, crypto is no longer a topic restricted to investors as it has also entered popular culture with many celebrities associating themselves with crypto assets. The following discussion helps you uncover an outline of the predictions for cryptocurrency future and their implications for investors. An understanding of the possible future for cryptocurrencies could help in preparing for changes in the crypto ecosystem over the next 5 or 10 years.

The implementation of the above research is yet to be worked on the given proposed solution. This research paves the way for further work, by analyzing the algorithm leads to the improved results. In future, additional data can be used for the predictive analysis of cryptocurrency for the precise price rate. Using machine learning algorithms as given will improve the prediction itself by learning the data provided.

The future of cryptocurrencies in 2022 and beyond is considerably uncertain. Crypto adoption increased by unreal margins during the pandemic and the crypto market registered humongous levels of trading volume in 2021. The cryptocurrency future predictions for 2022 emphasize regulations and approval for crypto ETFs alongside institutional adoption.

All of these factors would affect not only the general crypto industry but also the experience of users and enterprises. In the long run, crypto has the potential to replace various conventional financial instruments. However, it can also work in tandem with existing financial services and products such as traditional brokerage accounts. At the same time, the perception of real value with cryptocurrencies alongside the emerging utilities of crypto presents conclusive implications for larger crypto adoption rates. Learn more about cryptocurrencies and the crypto ecosystem to figure out more details about its future prospects.

VIII. REFERENCE

1. Augustin, Peter& Salman, Agha. (2019). Predictive Analytics in Cryptocurrency Using Neural networks: A Comparative Study. International Journal of Advanced Technology and Engineering Exploration. 7. 425.

2. A.ElBahrawy, L. Alessandretti, A. Kandler, R. Pastor-Satorras and A. Baronchelli,"Evolutionary dynamics of the cryptocurrency market", Royal Society open science, vol. 4, no. 11, pp. 170623, 2017.

3. Huang J.Z., Huang W, Ni J.: ‘Predicting Bitcoin returns using High Dimensional technical indicators’, *J.FinanceDataSci.*, 2018, **5**, pp. 140 155, http://www.sciencedirect.com/science/article/pii/S2405918818300928.

4. Qiu L., Fanbin W., Longxiu Y. *et al*.: ‘Prediction of Bitcoin price based on wavelet analysis’, *Neijiang sci. Technol.*, 2015, **36**, (5), pp. 129 – 130.

5. Jing L.: ‘Using BP neural network to construct Bitcoin marketforecastingmodel’, *FinanceAccount.Month.*,2016, **536**, (21), pp. 33 – 36.

6. D. Shah and K. Zhang, "Bayesian regression and Bitcoin", Massachusetts Institute of Technology."

7. I.Georgoula, D.Pournarakis,C. Bilanakos,D. Sotiropoulos and G.Giaglis, "UsingTime-Series and Sentiment Analysis to Detect the Determinants of Bitcoin Prices".

8. Babitha, D., Ismail, M., Chowdhury, S., Govindaraj, R., & Prakash, K.B. (2020). Automated road safety surveillance system using hybrid cnn-lstmapproach.International Journal of Advanced Trends in ComputerScience and Engineering,9(2),1767-1773. doi:10.30534/ijatcse/2020/132922020.

9. Babitha, D., Jayasankar, T., Sriram, V. P., Sudhakar, S., &Prakash, K.B. (2020). Speech emotion recognition using state-of-art learning algorithms. International Journal of Advanced Trends in Computer Science and Engineering, 9(2), 1340-1345. doi:10.30534/ijatcse/2020/67922020.

10. Bharadwaj, Y. S. S., Rajaram, P., Sriram, V. P., Sudhakar, S., Prakash, K. B. (2020). Effective handwritten digit recognition using deep convolution neural network. International Journal of Advanced.

11. Enireddy, V., Gunda, K., Kalyani, N. L., & Prakash, K. B. (2020).Nature inspired binary grey wolf optimizer for feature selection in the detection of neurodegenerative (parkinson) disease. International Journal of Advanced Trends in Computer Science and Engineering, 9(3),3977-3987. doi:10.30534/ijatcse/2020/222932020.

12. Ismail, M., Prakash, K. B., & Rao, M. N. (2018). Collaborative filtering-based recommendation of online social voting. International Journal of Engineering and Technology(UAE), 7(3), 1504-1507.doi:10.14419/ijet.v7i3.11630.

13. Kolla, B. P., & Raman, A. R. (2019). Data Engineered content extraction studies for Indian web pages doi:10.1007/978-981-10-8055-5_45 Retrieved from www.scopus.com.

14. Meghana, A. S., Sudhakar, S., Arumugam, G., Srinivasan, P., &Prakash,K. B. (2020). Age andgender prediction using convolution, resnet50 and inception resnetv2. International Journal of Advanced Trends in Computer Science and Engineering,9(2)13281334.doi:10.30534 /ijatcse/2020/65922020.

15. Prakash, K., Lakshmi Kalyani, N., Vadla, P. K., & Naga Pawan, Y. V. R.(2020). Analysis of mammography for identifying cancer cells using convolution neural networks. International Journal of Advanced Trends in Computer Science and Engineering,9(2).11841188. doi:10.30534/ijatcse/2020/44922020.