EAI Endorsed Transactions
on Internet of Things
Research Article EAI.EU

Detection of Anomalous Bitcoin Transactions in Blockchain Using ML

Soumya Bajpai [1], Kapil Sharma [2] and Brijesh Kumar Chaurasia[3,*]

[1] Department of Computer Science and Engineering, India
[2]Amity University Madhya Pradesh, Maharajpura Dang, Gwalior, India
[3] Pranveer Singh Institute of Technology, Kanpur, India

Abstract

An Internet of Things (IoT)-enabled blockchain helps to ensure quick and efficient immutable transactions. Low-power IoT integration with theBitcoin network has created new opportunities and difficulties for blockchain transactions. Utilising data gathered from IoT-enabled devices, this study investigates the application of ML regression models to analyse and forecast Bitcoin transaction patterns. Several ML regression algorithms, including Lasso Regression, Gradient Boosting, Extreme Boosting, Extra Tree, and Random Forest Regression, are employed to build predictive models. These models are trained using historical Bitcoin transaction data to capture intricate relationships between various transaction parameters. To ensure model robustness and generalisation, cross-validation techniques and hyperparameter tuning are also applied. The empirical results show that the Bitcoin cost prediction of blockchain transactions interms of time series. Additionally, it highlights the possibility of fusing block- chain analytics with IoT data streams, illuminating how new technologies mightwork together to enhance financial institutions.

Keywords: Machine learning (ML), Regressor Model, Blockchain, Bitcoin Prediction, IoT

Received on 10 June 2024, accepted on 26 July 2024, published on 23 August 2024

doi: 10.4108/eetiot.7042

1. Introduction

The digital currency known as Bitcoin was created to function as a private, irreversible payment mechanism utilizing peer-to-peer networks and open-source software. Since Bitcoin has no physical form and is not supported by any organization or government [1]. Bitcoin was inspired by the concept of cryptocurrency or computer-generated money [2], [3]. Cryptocurrencies are only forms of digital, computer-created money. Ituses cryptographic techniques to safeguard transactions. To create tamper-proof records of shared transactions, IoT facilitates data flow from Internet-connected devices to Blockchain networks [3], [4]. Bitcoin is kept in electronic wallets, which can be either hardware or software-based. Private keys kept in wallets provide users access toand control over their Bitcoin assets [4], [5]. Although the cryptographic structure of the blockchain makes Bitcoin transactions usually safe, users must take security measuresto guard their private keys and wallets against theft or hacking. This paper intends to categorise various Bitcoin transactions using machine learning (ML) approaches to better understand organizational and management aspects of regulation and compliance

*Corresponding author. Email: brijeshchaurasia@ieee.org

1.1. Motivation

Tremendous possibilities for a variety of applications and insights can result from the combination of Bitcoin with machine learning. The following is the rationale for com bining Bitcoin and ML:
a) Price Forecast: Regression models and other machine learning algorithms may be taught to analyse previous Bitcoin price data and forecast future price changes. Despite the volatility of cryptocurrency markets, machine learning algorithms may be able to discern patterns and trends that human traders would find challenging to spot.
b) Trading Strategies: Using a variety of indicators, market sentiment research, and historical data, traders and investors may create and improve trading strategies. Algorithms can process and analyse massive volumes of data in real-time to help traders make better-educated decisions.
c) To obtain insights, make wise judgements, and build strategies in the dynamic and ever-evolving world of cryptocurrencies, it is possible to use Bitcoin in combination with machine learning.

1.2.Contributions

The following is a summary of this article's contributions:
a) Develop a classification approach or methodology to forecast the closing price of Bitcoin using ML regression methods. This may require a unique combination of features, a revolutionary model design, or a novel ensemble method on a dataset of Bitcoin transactions.
b) Investigate sophisticated ensemble approaches or meta-learning methods that judiciously combine many regression models, potentially modifying the weights or configurations of the models depending on past performance.
c) Comparison between different types of regression models on the Bitcoin transactions dataset [6–8].

1.3.Structure of the paper

The rest of the article is structured as follows: Section 2 contains related content. In Section 3, a problem formulation is offered. The specifics of the proposed method are covered in Section 4. The outcomes and analysis of the suggested models based on ma- chine learning are presented in Section 5. The Section 6 provides a conclusion and the work's future scope.

2. Related Works

There is a dearth of studies on how to correctly anticipate Bitcoin values using MLalgorithms [9] employed a latent source model created by [10] to forecast Bitcoin's price, resulting in an 89% return in 50 days and a ratio of 4:1.

It has also been investigated [3] how to estimate Bitcoin values using text data from social media and other sources, and this study looked at the frequency, net-work hash rate, and amount of Wikipedia views in addition to sentiment analysis with support vector machines to study the connection between Google Trends views,tweets, and the price of bitcoin. Similar ideas were utilised in [11], but instead of fore- casting trade volume using Bitcoin's price, they used Google Trends views. The often small sample size and propensity for incorrect information to proliferate on message boards like Reddit or through other (social) media channels like Twitter, which artificially inflate or deflate prices, are two drawbacks of such a study, albeit [8]. On Bitcoinexchanges, there is very little liquidity. As a result, there is a greater possibility of market manipulation. Thus, the emotion expressed on social media is not further taken into ac-count. Support vector machines (SVM) and artificial neural networks were used to analyse the Bitcoin blockchain and predict the price of the cryptocurrency [12]. A typical ANN reported 55% accuracy in price direction. They concluded that there was minimalpredictability when utilising solely blockchain data. SVM, Random Forests, and Bino mial generalised linear model (GLM) were used with blockchain data by [13], who alsonoted prediction accuracy above 97% but limited the generalizability of their findings by not cross-validating their models. A disadvantage of both RNN and LSTM trainingis the significant computation required. For example, training 50 distinct MLP models using a network of 50 days Since NVIDIA developed the CUDA framework in 2006, a growing number of applications, including machine learning, have been developed that utilise the GPU's massively parallel capabilities. reported that training and testing of theirANN model was over three times faster when utilising a GPU as compared to a CPU. Similar to [14], classification time increased by an order of 80 times when using an SVM implemented on a GPU as opposed to an alternative SVM method performed on a CPU[15].

3. Problem Formulation

The volume of transactions on IoT-enabled blockchains is enormous, and each transaction is recorded [26], converted into a data block, and then added to a safe, immutable data chain that can only be extended [27]. Identification and classification of transactions out of millions of transactions in terms of close price and other parameters is a big challenge. There is no established method for choosing the data to train any model. Using the ML algorithm won't necessrily guarantee correct results if the data ischosen improperly. For instance, the model can yield unsatisfactory results if we expected a strong fall for

Bitcoin but unintentionally selected a prior data piece from one.Therefore, it is essential to select the appropriate data to train a model to estimate Bitcoin 'close' prices.

4. Proposed Mechanism

In this section, a proposed model, a dataset with a preprocessing approach, and measurement factors are presented. The Bitcoin historical dataset [16] is used to test the machine learning regression algorithm [17]. The major goal of the proposed approach is to use an ensemble strategy to get reliable and accurate forecasts [18]. To successfully predict ensemble machine learning techniques, the following framework is provided in this study [19]. The workflow of the proposed scheme is depicted inFig. 1.

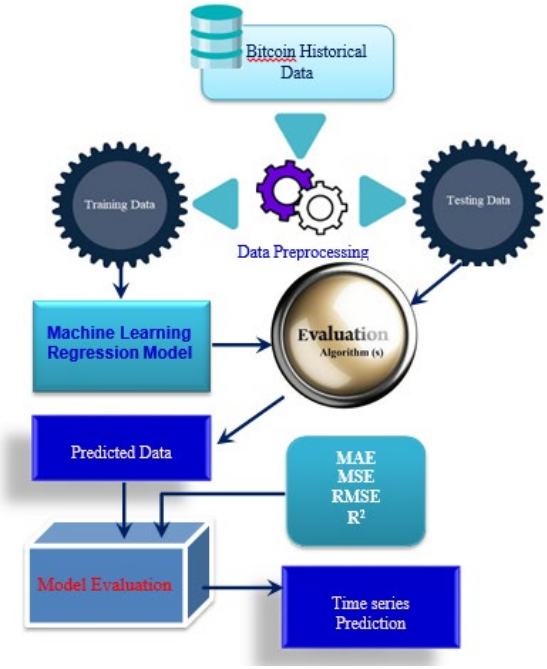

Figure 1. The workflow of the Proposed Model

There are some steps for a Bitcoin regression-based ML algorithm.

- The first phase includes the historical Bitcoin price dataset and relevant data (such as transaction volume, open, close, technical indicators, and sentiment analysis). The dataset is being preprocessed using some preprocessing algorithms to handle null values, data scaling, categorical variables, and unnamed rows and columns to keep the dataset clean.
- Then, using the data, we generate training and test sets. It is a time series, therefore make sure the split maintains the temporal order of the data.
- Choose an approach based on regression that is suitable for time-series pre- diction. The approach could include random forest regression, gradient boosting regression, extreme boosting, extra tree, and lasso regression. To choose the most significant features for the regression model, use methods like feature importance or correlation analysis.
- After that, we apply methods like grid search or randomized search to optimize hyperparameters.
- Apply the appropriate evaluation regression metrics, such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared to assess the model's performance.
- Create a trading strategy based on the model's forecasted Bitcoin price using the trained regression model.
- And finally, in the last step, we simply deploy our model.

4.1.Dataset

The historical price of Bitcoin in US dollars [16], [20] is collected from the Kaggle website. The dataset contains 1460 rows, or instances, representing 1 minute'sworth of data from 2017-09-13 to 2021-09-13, and 8 columns, or characteristics. Timestamp, High, Low, Open, Volume (BTC), Close, Volume (Currency), and Weighted Price are the columns listed here and shown in Fig 2.

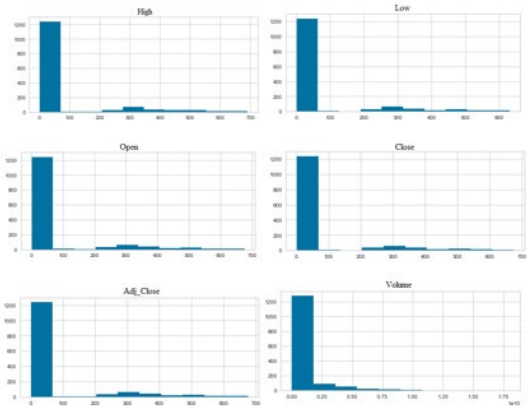

Figure 2. Dataset with features [16]

4.2.Preprocessing

In this phase, for data cleaning, we exclude the other 6 columns from the dataset since our model only considers the closing price of the bitcoin for the relevant days and applies the regression model to forecast the price of the bitcoin. We also eliminate any instances or rows of data that contain missing values.

4.3.Models

We used various types of regression ML approaches to evaluate the historical Bitcoin dataset. Due to the detection's low processing needs, these methods have also been considered. There are five models we used on this dataset: LASSO [21], Gradient Boosting Regression, Extreme Gradient Boosting, Extra Trees Regressor, and RandomForest Regressor [21].

A ML regression model is a particular type of strategy used to predict continuous numerical values based on input data. Regression is a supervised learning method in which the model discovers a link between the input data and the target variable, or thevalue we want to predict.

The regression model seeks to approximate this function by learning from a labelledtraining dataset. The function often takes the form of a mathematical equation.

Regression algorithms come in many forms, but some that are used in this article areas follows:

- **Least Absolute Shrinkage and Selection Operator (LASSO) regression**
  The ordinary least squares (OLS) cost function is modified by a penalty factor in this particular sort of linear regression [21]. The model is encouraged to choose and shrink the coefficients in the direction of zero by this penalty term, which is dependent on the absolute values of the model's coefficients. Since lasso regression has a propensity to make some coefficients absolutely zero, it is frequently used for feature selection since it effectively eliminates some characteristics from the model. The Lasso regression cost function can be expressed as follows:

  $$Cost\ Function = RSS + \lambda * \Sigma|\beta i|$$

  Where RSS (residual sum of squares) is the sum of squared differences between the predicted values and the actual target values, just like in ordinary least squares regression. λ (lambda) is the regularization parameter or penalty term. It controls the strength of the penalty applied to the absolute values of the coefficients. Higher values of λ result in more regularization, leading to more coefficient shrinkage and feature selection. In this context, represents the coefficients of the regression model, which are the parameters to be learned during the training process.

- **Gradient Boosting Regression**
  Progressive Boosting For regression tasks, regression is a potent ensemble machine learning approach. It is a member of the group of boosting algorithms that aggregate the predictions of numerous weak learners, often decision trees, to develop an effective forecasting model. Gradient boosting is renowned for its great accuracy and capacity for handling intricate data interactions.

- **Extreme Gradient Boosting (XGBoost)**
  It is a sophisticated and incredibly effective version of the Gradient Boosting method that was created to handle massive datasets and achieve cutting-edge performance in a variety of ML applications [24], including regression. To enhance the overall effectiveness of the Gradient Boosting method, XGBoost incorporates several optimizations and regularization strategies.

- **Extra Trees Regressor**
  An ML algorithm from the ensemble learning family is the Extra Trees Regressor [22]. It is a variation of the Random Forest technique that is employed for regression jobs that need to predict a continuous numerical output. The acronym "Extremely Randomised Tree" is "Extra Trees".

- **Random Forest Regressor**
  It is applied to problems involving regression in which the objective is to foretell a continuous numerical output. To build a more reliable and precise predictive model, Random Forest [23] combines the capabilities of many decision trees.

4.4.Measurement Factors

It's crucial to assess a regression model using various kinds of evaluation measuresto fully comprehend how well it performs. Different measures offer various perspectives on the model's resilience, accuracy, and bias. Some common metrics forassessing regressors are listed below:

- **Mean Squared Error (MSE)**
  It computes the average squared difference between the actual observed values and the values that were anticipated. Understanding the size of mistakes and determining how well a model matches the data are two areas where MSE is very helpful. Better model performance is indicated by lower MSE values.
- **Root Mean Squared Error (RMSE)**
  This assessment measure is frequently used in regression analysis. The square root of the squared differences between the actual observed values and the anticipated values is calculated. The accuracy and precision of a predictive model's predictions may be evaluated using RMSE.
- **Mean Absolute Error (MAE)**
  This assessment measure is frequently used in regression analysis. It calculates the mean absolute difference between the actual observed values and those that were projected. The model's accuracy is measured simply and easily by MAE.

- **R2(R Square)**
  A key statistic used in regression analysis to assess how well a model fits the observed data is

the coefficient of determination, sometimes known as R- squared. It offers a gauge for how effectively the independent variables (features) in a regression model explain the variability in the dependent variable (target).

5. Results and Analysis

In this section, results and discussions are discussed. The models are implemented in Jupyter. The simulation is carried out on an Intel Core i9-13900K 5.8 GHz processor with 32 GB of RAM and the Windows 10 operating system.

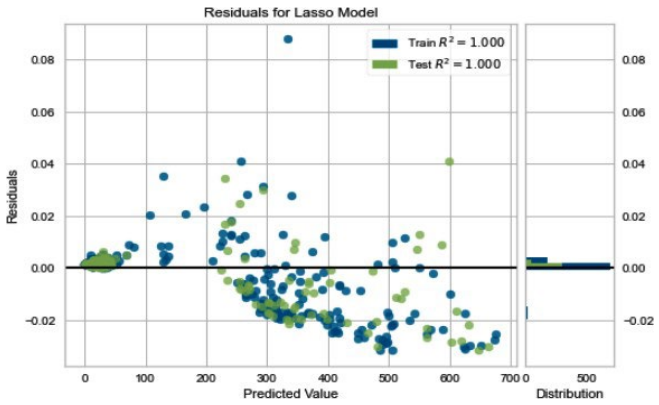

Figure 3. Residuals for Lasso Model

In Fig. 3, the residuals graph for the best model, *i.e.*, Lasso, Model residuals represent the differences between the actual observed values and the predicted values generated by the regression model. Analysing residuals can provide insights into the model's performance, uncover patterns, and identify potential issues. Here the blue dots represent the training data and the green dots represent the testing data, In this graph we find out maximum dots are available between 0.00 to -0.01 which represents the minimum error.

In Fig. 4, the prediction error graph for the best model, *i.e.*, the Lasso model, indicates that this model is a good fit for this dataset, so this concludes that it is an error- free prediction. Here, the best fit and identity are shown in this graph, along with the maximum number of best-fit data points performed across the entire range of predictedvalues

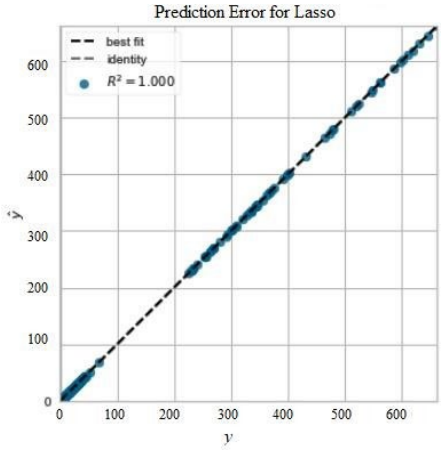

Figure 5. Prediction Error Graph for Lasso Model

The learning curve in Fig. 5 shows a model's performance changes as the volume oftraining data rises. It can help determine whether a model is overfitting, underfitting, or reaching the right level of generalization. In this graph, we identify that learning curvedata is in generalized form so we get accurate results from this model.

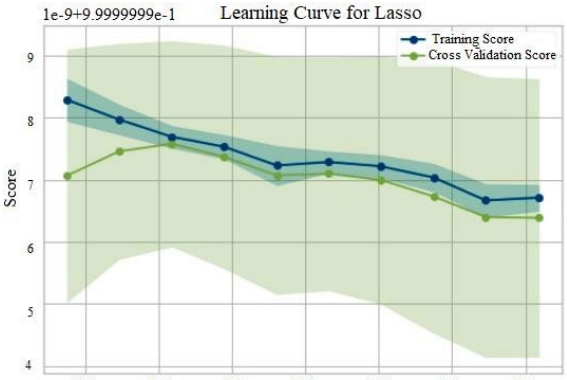

Figure 4. Learning Curve for Lasso Model

In Fig. 6, the prediction of the close value of Bitcoin blockchain transactions ac- cording to the date is accurately predicted from the Historical Bitcoin Dataset.

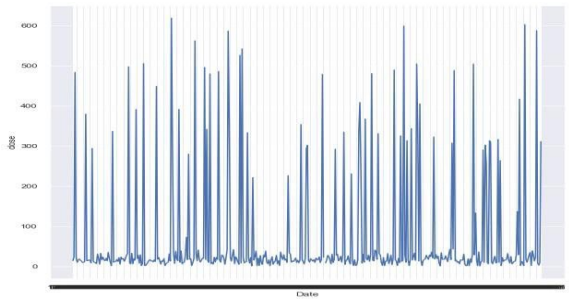

Figure 6. Predicted close value for Lasso Model

In Fig. 7, the performance indicators are used to compare and examine the out- comes. Lasso has a minimum MAE and maximum R^2, and another model has

a higher value of MAE. That's why we conclude that the Lasso machine learning regressor model has a higher prediction rate than other models. We get an MAE value of 0.0038 and an R^2 value is 1.000 in our finding. The result shoes that the peroformance of XGBoost model is better in terms of MAE, MSE, RSME, and R^2.

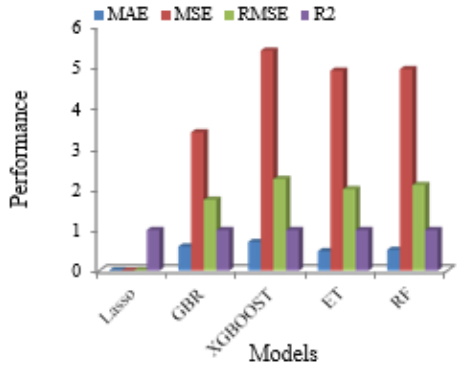

Figure 7. Comparative Analysis of ML Models

6. Conclusion and Future Work

In this paper, we have proposed a technique for selecting the data range from which to train an ML regression prediction model. The work also studied many methods for predicting the price of Bitcoin using various ML algorithms. To train the prediction model is provided a more accurate outcome, which data fraction should be used, there is a gap in techniques. To create our model, we employed evaluations like RMSE, MAE, MSE, and R2. An accurate forecast is obtained after training a LASSO Predicts Model [24]. In future, Predicting liquidity patterns in Bitcoin exchanges can aid in optimizing trading strategies and executing large trades more efficiently [25].

References

[1] Gao M, Lin S, Tian X, He X, He K, Chen S (2024) A bitcoin service community classification method based on Random Forest and improved KNN algorithm. In IET Blockchain- published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology, 1-11. https://doi.org/10.1049/blc2.12064.

[2] Kaastraa I, Boydb M (1996) Designing a neural network for forecasting financial time series. Neurocomputing 10 (3):215–236. https://doi.org/10.1016/0925-2312(95)00039-9

[3] Srivastava S, Chaurasia BK, Singh D (2023) Chapter 18 - Blockchain-based IoT security solutions. In: Pandey R, Goundar S, Fatima, Distributed Computing to Blockchain: Archi- tecture, Technology, and Applications, Elesvier, Ch-18, 327–339. https://doi.org/10.1016/B978-0-323-96146-2.00020-6

[4] Saxena R, Arora D, Nagar V (2023) Classifying Transactional Addresses using Supervised Learning Approaches over Ethereum Blockchain. Procedia Computer Science 218:2018– 2025. https://doi.org/10.1016/j.procs.2023.01.178

[5] Georgoula I, Pournarakis D, Bilanakos C, Sotiropoulos DN (2015) Using Time-Series and Sentiment Analysis to Detect the Determinants of Bitcoin Prices. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2607167

[6] Matta M, Lunesu I, Marchesi M (2015) Bitcoin spread prediction using social and web search media. CEUR Workshop Proceedings 1388. Online available at: https://ceur-ws.org/Vol-1388/DeCat2015-paper3.pdf, Last accessed on 22 Jan 2024.

[7] Gu B, Konana P, Liu A, Rajagopalan B, Ghosh J (2006) Identifying Information in StockMessage Boards and Its Implications for Stock Market Efficiency. Workshop on Infor mation Systems and Economics, 1–6. Online available at: https://www.ideal.ece.utexas.edu/pdfs/151.pdf, Last accessed on 22 Jan 2024.

[8] Greaves, Alex and Au B (2015) Using the bitcoin transaction graph to predict the price ofbitcoin. Quoted, 1-8. Online available at:https://snap.stanford.edu/class/cs224w- 2015/projects_2015 Using_the_Bitcoin_Transaction_Gra ph_to_Predict _ the_Price_of_Bitcoin.pdf, Last accessed on 22 Jan 2024.

[9] Madan I, Saluja S, Zhao A, (2015) Automated Bitcoin trading via machine learning algo- rithms,1– Online available at:https://cs229.stanford.edu/proj2014/ Isaac% 20Madan,%20Shaurya%20Saluja,%20Aojia%20Z hao,Automated% 20Bitcoin%20Trading%20via% 20Machine%20Learning%20Algorithms.pdf, Last accessed on 22 Jan 2024.

[10] Catanzaro B, Sundaram N, Keutzer K (2008) Fast support vector machine training and classification on graphics processors. In the proceedings of the 25th International Conference on Machine Learning, 104–111. https://doi.org/10.1145/1390156.1390170

[11] Adhikari M, Hazra A, Menon VG, Chaurasia BK and Mumtaz S (2021) A Roadmap of Next-Generation Wireless Technology for 6G-enabled Vehicular Networks. In IEEE Inter-net of Things Magazine 4(4): 79–85. https://doi.org/10.1109/IOTM.001.2100075

[12] Karthik MG, Krishnan MBM (2021) Detecting Internet of Things Attacks Using Post Pruning Decision Tree-

Synthetic Minority Over Sampling Technique. International Journal of Intelligent Engineering and Systems 14:105–114. https://doi.org/10.22266/ijies2021.0831.10.

[13] Saxena R, Arora D, Nagar V (2023) Efficient blockchain addresses classification through cascading ensemble learning approach. In International Journal of Electronic Security and Digital Forensics 15(2): 195-210. https://doi.org/10.1504/IJESDF.2023.129278.

[14] Rai S, Chaurasia BK, Gupta R, Verma S (2023) Blockchain-based NFT for Healthcare System. In IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT).700–704. https://doi.org/10.1109/CSNT57126.2023.10134632

[15] Alsaif SA (2023) Machine Learning-Based Ransomware Classification of Bitcoin Trans- actions. Applied Computational Intelligence and Soft Computing 2023:6274260, 1-10. https://doi.org/10.1155/2023/6274260

[16] Bitcoin Historical Data, Online available at: https://www.kaggle.com/datasets/mcziel- inski/bitcoin-historical-data, Last accessed on 29 March 2023.

[17] Jatoth C, Jain R, Fiore U, Chatharasupalli S (2022) Improved Classification of BlockchainTransactions Using Feature Engineering and Ensemble Learning. In Future Internet 14(1):1–12. https://doi.org/10.3390/fi14010016

[18] Wimalagunaratne M, Poravi G (2018) A predictive model for the global cryptocurrency market: A holistic approach to predicting cryptocurrency prices. In 8th International Conference on Intelligent Systems, Modelling and Simulation (ISMS), 78–83. https://doi.org/10.1109/ISMS.2018.00024

[19] Sin E, Wang L (2018) Bitcoin price prediction using ensembles of neural networks. In 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discov-ery, 666–671. https://doi.org/10.1109/FSKD.2017.8393351

[20] Karthik S, Bhadoria RS, Lee JG, Sivaraman AK, Samanta S, Balasundaram A, Chaurasia BK, Ashokkumar, S (2022) Prognostic Kalman Filter Based Bayesian Learning Model forData Accuracy Prediction. In CMC-Computers Materials & Continua 72 (1): 243–259. https://doi.org/10.32604/cmc.2022.023864

[21] Ranstam J, Cook JA (2018) LASSO regression. British Journal of Surgery 105:1348. https://doi.org/10.1002/bjs.10895

[22] Sweta B, Siva RKS, Praveen KRM, Kaluri R, Singh S, Gadekallu TR, Alazab M, Tariq U(2020) A Novel PCA-Firefly Based XGBoost Classification Model for Intrusion Detectionin Networks. Electronics, MDPI 9(219): 1-16. https://doi.org/10.3390/electronics9020219

[23] Pelletier Z, Abualkibash M (2020) Evaluating the CIC IDS-2017 Dataset Using Machine Learning Methods and Creating Multiple Predictive Models in the Statistical Computing Language R. In International Research Journal of Advanced Engineering and Science 5 (2):187–191

[24] Bajpai S, Sharma K, Chaurasia BK (2023) Intrusion Detection Framework in IoT Net- works. In SN Computer Science, Springer, Special Issue on Machine Learning and Smart Systems 4(350): 1-16. https://doi.org/10.1007/s42979-023-01770-9

[25] Saranya T, Sridevi S, Deisy C, Chungd TD, Khane MKAA (2020) Performance Analysis of Machine Learning Algorithms in Intrusion Detection System: A Review. Procedia Computer Science (171): 1251–1260. https://doi.org/ 10.1016/j.procs.2020.04.133

[26] Sharma AK and Chaurasia BK (2023) Blockchain-based NFT for Evidence System. In: Roy, B.K., Chaturvedi, A., Tsaban, B., Hasan, S.U. (eds) Cryptology and Network Securitywith Machine Learning. ICCNSML 2022. Algorithms for Intelligent Systems. Springer, Singapore, 441-451 https://doi.org/10.1007/978-981-99-2229-1_37

[27] Saxena R, Arora D, Nagar V, Chaurasia BK (2023) Privacy Provisioning on Blockchain Transactions of Decentralized Social Media. In Blockchain technology for social media computing, IET, England & Wales, Ch-6, 978-1-83953-543-7, 93 –117, 2023. https://doi.org/10.1049/PBSE019E_ch