IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 13 May 2023, accepted 9 June 2023, date of publication 20 June 2023, date of current version 6 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3287888

RESEARCH ARTICLE

Enhancing Cryptocurrency Price Forecasting Accuracy: A Feature Selection and Weighting Approach With Bi-Directional LSTM and Trend-Preserving Model Bias Correction

MUHAMMED RAFI[1], QUBLAI ALI KHAN MIRZA[2], MUHAMMAD IZAAN SOHAIL[1], MARIA ALIASGHAR[1], ARISHA AZIZ[1], AND SUFIAN HAMEED[1]

[1]School of Computing, Department of Computer Science, National University of Computer and Emerging Sciences, Islamabad 44000, Pakistan
[2]Cyber and Technical Computing Department, University of Gloucestershire, GL50 2RH Cheltenham, U.K.

Corresponding author: Muhammed Rafi (muhammad.rafi@nu.edu.pk)

This work was supported in part by the University of Gloucestershire's QR-Fund; and in part by the Faculty Research Support Program (FRSG-Fall-2021) by the National University of Computer and Emerging Sciences, Islamabad, under Grant 11-71-8/NU-R/21.

ABSTRACT A cryptocurrency is a digitized, encrypted, and decentralized virtual currency, which is impossible to counterfeit or double-spend. It is one of the very popular investment instruments and traded in blockchain based crypto exchanges on ever growing volume. It is quite volatile due to imbalance of supply and demand, government regulations, investor sentiment and above all media hype. Cryptocurrency price forecasting is an active area of research and several approaches have been proposed recently. This study proposed a price forecasting model based on three vital characteristics (i) a feature selection and weighting approach based on Mean Decrease Impurity(MDI) features. (ii) Bi-directional LSTM and (iii) with a trend preserving model bias correction (CUSUM control charts for monitoring the model performance over time) to forecast Bitcoin and Ethereum values for long and short term spans. The data for both currencies were analyzed in three different intervals: (i) April 01, 2013 to April 01, 2016 (ii) April 01, 2013 to April 01, 2017 and (iii) April 01, 2013 to December 31, 2019. Extensive series of experiments were performed and evaluated on Root Mean Square Errors (RMSE). For bitcoin forecasting, the model achieved RMSE values 3.499 for interval 1, 5.070 for interval 2 and 6.642 for interval 3. Similarly, for Ethereum RSME of 0.094, 0.332, 3.027 are obtained for the three intervals respectively, On a new test-set collected from January 01, 2020 to January 01, 2022 for the two cryptocurrencies we obtained an average RSME of 9.17, with model bias correction, Comparing with the prevalent forecasting models we report a new state of the art in cryptocurrency forecasting.

INDEX TERMS Blockchain, cryptocurrency, machine learning.

I. INTRODUCTION

Digital technologies are driving transformative change. Everyone, nowadays, appears to be in the midst of digital transformation. The economic paradigm is changing as new technologies continue to redefine product and factor markets, impacting businesses and work in profound ways. Based on a recent report, 24.3% of global GDP will come from digital economy by 2025 [1]. It can be added that the necessity of digital transformation became more obvious after the Corona Virus, since more automated processes are now in practice than they were before. According to a new McKinsey Global Survey of executives, their companies have accelerated the digitization of their customer and supply-chain interactions and of their internal operations by three to four years. And the share of digital or digitally enabled products in their portfolios has accelerated by a shocking seven years [2].

The associate editor coordinating the review of this manuscript and approving it for publication was Jon Atli Benediktsson.

Digitization has led a consensus to emerge among national regulators and global standard setting bodies that blockchain technology offers society and the economy significant new prospects [3]. Blockchain offers to take advantage of the disruptive cryptocurrency application potential to not only minimize the negative aspects of these technologies, but also to leverage their beneficial aspects. The two most well-known cryptocurrencies are Bitcoin and Ethereum their volatility can be seen in Fig. 1 and Fig. 2. Bitcoin being the dominant one in digital currency market due to its high price. The major objective is to provide a global network of transactions and exchanges by allowing two willing parties to negotiate directly with each other without resorting to an expensive intermediary. The advent and eminence of these cryptocurrencies have led various merchants around the world to accept payments in Bitcoin. Further, online buying and selling platforms such as Amazon, eBay, or Craigslist allow this method of payment too [4].

Ethereum aims to provide a blockchain with a built-in Turing-complete programming language. Its ultimate intend is to integrate and improve on the concepts of scripting, altcoins, and on-chain meta-protocols, allowing developers to create arbitrary consensus-based applications with the scalability, standardization, feature-completeness, ease of development, and interoperability offered by these different paradigms all at once [5]. An Ethereum blockchain, however, is similar to the Bitcoin blockchain. The main difference lies that Ethereum blocks contain not only the block number, difficulty, nonce, etc, but also the transaction list and the most recent state [6].

Our main motivation comes from paper [7] and we used it as a base paper. The base paper has used the price indicators of Bitcoin up to December 31, 2019 and provides a near-to-accurate estimate of BTC price forecast using machine learning model on short-term (1, 7, 14 days) and mid-term (30, 60, 90 days) basis. However, the cited work uses wrapper approach to filter out important pricing features of BTC and then manually reduced the 53 indicators extracted from their approach to 12 prime features. This is both, a time-expensive and a complex methodology to draw important attributes of BTC. We believe that with increasing popularity of cryptocurrencies among the individuals, it is necessary to focus on multiple cryptocurrencies rather than a single one. Therefore, we address both the digital currencies, Bitcoin and Ethereum, due to their widespread use and high valuation. In our study, the use of vanilla Bi-LSTM model manifests greater efficiency on the grounds of an improved time-saving and automated feature selection approach, the Mean Decrease Impurity (MDI) approach, which selects optimal indicators completely by itself, resulting in highly-accurate price forecast and better results than the reference paper. Moreover, our research not only highlights the important indicators in the prediction, but also represents the feature relation in a graphical form displayed on an interactive website. Alongside this, it uses model agnostic methods and several corrections are done. Rapidly changing scenarios were adding bias to the previous model; hence, we discount the bias and make use of CUSUM control chart, leading to better performance of the model in long-term.

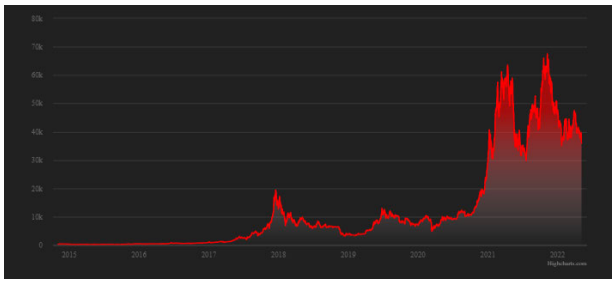

FIGURE 1. Bitcoin prices.

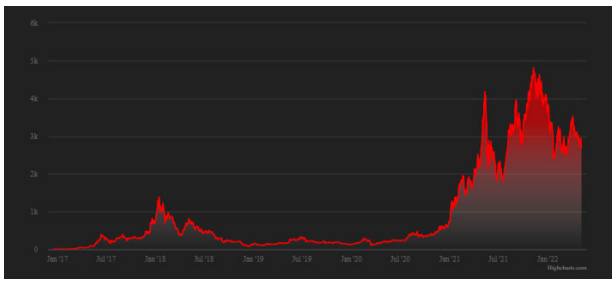

FIGURE 2. Ethereum prices.

In summary following are some of the contributions from this research work:

- - We have collected transaction level market data of two cryptocurrencies for performing forecasting and comparing with [7]
- - We proposed a feature selection approach based on Mean Decrease Impurity(MDI) for the price time series of the two cryptocurrencies.
- - A Bi-directional LSTM based model is proposed for price forecasting using MDI features and weighting schemes.
- - In order to deal with fluctuation in price a trend preserving model bias correction using CUSUM control chart is proposed and for the selected forecasting horizons lowest errors were obtained.

This paper is organized as follows. Section II disucss some relevant work for cryptocurrency price forecasting, in which we disucssed some statistical, machine learning and deep learning approaches. In section III, we presents our methodology starting from data collection, prepossessing and prepare for model implementation and validation. In section IV, we present results from our experiments. Section V and VI discusses insight from the results and conclusion respectively.

II. RELATED WORK

In recent years, both investors and scholars have been paying close attention to the rise in Bitcoin prices [8].Literature has studied various metrics such as BTC-related historical technical indicators [8], [9], [10] and social media interest [11], [12], [13], [14] to investigate the predictability of BTC returns. Reference [9] divided the daily bitcoin

return domain into 21 intervals with the goal of predicting which interval the next day return will be in. They created a return-predictor/model based solely on bitcoin's past prices, employing 124 technical indicators.

Twitter is rapidly being utilised as a news source, educating users about the currency and its rising popularity, impacting buying decisions. As a result, a cryptocurrency user or trader can gain a purchasing or selling advantage by swiftly recognizing the impact of tweets on price direction [12].To see if Bitcoin prices respond to useful signals obtained from Twitter and Google Trends, researchers used a diffusion model. The empirical findings suggest that Bitcoin prices are influenced in part by social media attention, implying a sentimental desire for information demand [11]. Moreover, [15] examined user comments in online cryptocurrency communities in order to forecast volatility in cryptocurrency prices(Bitcoin and Ethereum) and transaction volume. They found that BTC is particularly correlated with the number of positive comments on social media. They reported an accuracy of 79% along with Granger causality test, which implies that user opinions are useful to predict the price fluctuations.

Pure models, when it comes to time-series forecasts, solely use historical data on the variable to be predicted. Autoregressive Integrated Moving Average (ARIMA) [16] is a pure time-series forecast model. Reference [17] shows an ARIMA-based time-series forecast for BTC prices for the next day.For univariate and stationary time-series data, pure time-series models are more suited. For the following reasons, we focus on machine learning with higher level features rather than typical models in this paper. To begin with, Bitcoin and Ethereum values are extremely volatile and non-stationary.The data contains a high number of features, and the suggested machine learning methodology covers autocorrelation, seasonality, and trend effects, whereas pure time-series models require human tuning to address these effects during the training phase.

Explanatory models, on the other hand, use a function of predictor variables to forecast the target variable in the future. Model-based time-series forecasting systems have the drawback of assuming data distributions in advance. For example, [17] and [18] are based on a log-transformation of the BTC prices. Reference [18] revealed the effect of Bayesian neural networks (BNNs) by analyzing the time series of Bitcoin process.By using daily data from September 11, 2011 to August 22, 2017 they compared the Bayesian neural network (BNN) to other linear and nonlinear benchmark models in an empirical investigation on modelling and predicting the BTC price. They discovered that BNN Bitcoin (BTC) prices from April 2013 to April 2020 are good predictors of BTC log-transformed price, which explains the BTC price's extreme volatility.

Considering bitcoin prices are nonlinear and non-stationary, data distribution assumptions may have a negative impact on forecast performance. Statistical techniques to time series modelling, such as ARIMA models, can only be used to model stationary time series (those whose attributes do not change depending on the time the series is viewed), and therefore do not account for seasonality. Machine Learning models can potentially solve this problem by capturing non-linearity in data caused by external causes. The inherent nonlinear and non-stationary features of data are exploited by machine learning-based approaches. By considering the underlying reasons that influence the predicted variable, they can also learn from the explanatory features.Several studies have used machine learning based models to predict the price of cryptocurrencies. Reference [19] divided Bitcoin's price into two categories: daily and high-frequency. For Bitcoin 5-minute interval price prediction, they employed Random Forest, XGBoost, Quadratic Discriminant Analysis, Support Vector Machine, and Long Short-term, with an accuracy of 67.2%. Reference [20] forecasted the Bitcoin exchange rates (maximum, minimum and closing prices), considering daily data.They used ANN,RNN and SVM models with SVM obtaining the best results, a MAPE of 1.28%.

Reference [21] investigated the predictive power of blockchain network-based features on the future price of Bitcoin. They employed a SVM based regression model with 1.98 MSE.A Bayesian optimised recurrent neural network (RNN) and a Long Short Term Memory (LSTM) network were implemented by [22]. The LSTM had the highest classification accuracy of 52% and the lowest RMSE of 8%. Reference [23] suggested (RNN) methods based on GRU,LSTM, and LSTM(bi-LSTM models) to predict the prices of Bitcoin (BTC), Litecoin (LTC), and Ethereum (ETH). The GRU model outperformed the long short-term memory (LSTM) and bidirectional LSTM (bi-LSTM) models in terms of prediction for all forms of bitcoin. With MAPE percentages of 0.2454%, 0.8267%, and 0.2116% for BTC, ETH, and LTC, respectively, GRU gives the most accurate prediction for LTC. References [24], [25], [26] employed LSTM and GRU models for predicting the price of cryptocurrencies.Standard LSTM and GRU models gave an MSE loss of 0.02085 and 0.02113 respectively [25]. Reference [27] employed an RNN-based LSTM model to forecast bitcoin prices. The results were obtained by extrapolating graphs and Root Mean Square Error was 3.38. Lastly [28] investigated the predictability of the bitcoin market across prediction horizons ranging from 1 to 60 min.They tested various machine learning models and found that, while all models outperform a random classifier, recurrent neural networks and gradient boosting classifiers were especially well-suited for the examined prediction tasks.

III. METHODOLOGY

In this research, our primary objective is on the time-series prediction of Bitcoin and Ethereum prices, using a machine learning-based model and monitoring shifts in it. A time series is a sequence of results collected over time. Time-series forecasting is done on the basis of historical time-stamped values, where scientific forecasts are made using this data to

predict the values in the near future. It may not always give accurate predictions, but an estimate of forecasted values, which may vary dramatically- especially when dealing with time series data's varying variables and features not under control. Our intent is to forecast the value of variable x, target variable, in the future, such that

$$\hat{x}[t+s] = f(x[t], x[t-1], \ldots, x[t-n]), s > 0 \quad (1)$$

where s is the forecasting horizon. Here, we take in account the forecasting horizon at the end of 1, 7, 14, 30, 60 and 90 days, scaling the time of the relevant historical data we gathered.

The development of the dataset is the first step in the ML-based time-series approach. Followed by, training the machine learning model and making predictions for one day ahead. Forecasting cryptocurrency prices has foundational interdependencies that are difficult to grasp and model. Variance, standard deviation and moving averages are a few statistical parameters that alter overtime. These interrelationships manifest themselves as technical indicators, which are described in the next section. The Bitcoin and Ethereum price attributes were acquired from available open data sources, which were then preprocessed. In the preprocessing stage, the obtained data is cleaned and scaled using the MinMax scaler. The data gathered for both, BTC and ETH, are then processed and split into three intervals. In order to find the relevant indicators, feature selection is employed. Additionally, we created various datasets for three time intervals with varying forecasting horizons and feature selection is carried out separately for each of these. The feature selection procedure is demonstrated in Fig (no.). It plays a key role in machine learning-based time-series forecasting since it optimizes the process and raises the predictive ability of machine learning algorithms by determining the most significant features and eliminating the inessential and unrelated ones. Feature selection is implemented using Random Forest (RF) and Mean Decrease in Impurity (MDI) as Random Forest tree-based methods are typically ranked by how well it increases the node purity. The average reduction in the impurity throughout all the trees is known as gini impurity. For classification, the gini gain splitting is described by the MDI metric, which integrates the weighted average of each of the individual trees by the number of samples it splits. Gini impurity index is then calculated by

$$G = \sum_{i=1}^{n_c} p_i(1-p_i) = 1 - \sum_{i=1}^{n_c} p_i^2 \quad (2)$$

where, pi is the ratio of classes in target feature and nc is the number of classes in the target feature. It basically assesses the change and irregularity within a group of elements and determines the likelihood of misclassifying a feature on the basis of grouping of classes within a set. For regression, gini impurity is the Residual Sum of Squares (RSS), defined by

$$RSS = \sum_{i=1}^{n} (y_i - f(x_i)^2) \quad (3)$$

where, yi is ith value of variable to be predicted, f(xi) is the predicted value of yi and n is the upper limit of summation. Machine learning models for both, regression and classification, are trained on training split and validated using the hold-out approach.

ML has been shown to improve the processing of both structured and unstructured data flows, thereby recognizing precise patterns within large sets of data quickly. It outperforms the traditional approaches to time-series forecasting. Bi-directional Long Short Term Memory (Bi-LSTM) is used for the classification, as well as regression. Here, every input sequence contains data from both past and the present. As a result, by combining LSTM layers from both directions, it yields more meaningful results. This, alongside the model summary, is shown in Fig (no.) (add bi-lstm basic fig and model summary). The model consists of 128 units with ReLU activation function, y = max(0, x), outputting the input directly if it is a positive value, else gives zero. A dropout layer is added which drops 20% of neurons in order to avoid overfitting, followed by a dense layer. The loss function used is Mean Square Error (MSE), defined as

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \quad (4)$$

where, n is the number of data points, yi is the observed values and Yî is the predicted values. This depicts the closeness of data points to the regression line. Adam optimizer is used at compilation of the model as it requires fewer tuning parameters and takes lesser time to compute.

Non-stationary time-series data indicates a range of statistical characteristics. Therefore, highly dynamic scenarios result in addition of bias to the model. The cumulative sum (CUSUM) control chart tracks tiny changes in the process by computing the average of each sample and calculates the total number of deviations from a target. Either of the following quantities is plotted:

$$S_m = \sum_{i=1}^{m} (\bar{x}_i - \hat{\mu}_0) \; or \; S'_m = \frac{1}{\sigma_{\bar{x}}}\sum_{i=1}^{m} (\bar{x}_i - \hat{\mu}_0) \quad (5)$$

Against the total sample m, where m0 is the in-control mean estimate and sigma-x-bar is the known standard deviation of the mean of the sample. CUSUM charts are made use of after modeling to discount the bias and monitor the small shifts for improved model performance.

A. DATA

A web scraper written in Python 3.6 was used to obtain data about Bitcoin and Ethereum from https://bitinfocharts.com. The data for both currencies was analyzed in three different intervals: i) 2013 April 1 - 2016 April 1, ii) 2013 April 1 - 2017 April 1, and iii) 2013 April 1 - 2019 December 31st. These 3 intervals were used to compare with the models given by [7]. More than 700 technical indicators were included in the data. Using the feature selection methods outlined later

TABLE 1. Intervals and timeline for forecasting.

S-no	Interval Batch	Interval Start	Interval End
1	Interval Batch-1	April 2013	April 2016
2	Interval Batch-2	April 2013	April 2017
3	Interval Batch-3	April 2016	April 2019
4	Interval for Testset	Jan 2020	Jan 2022

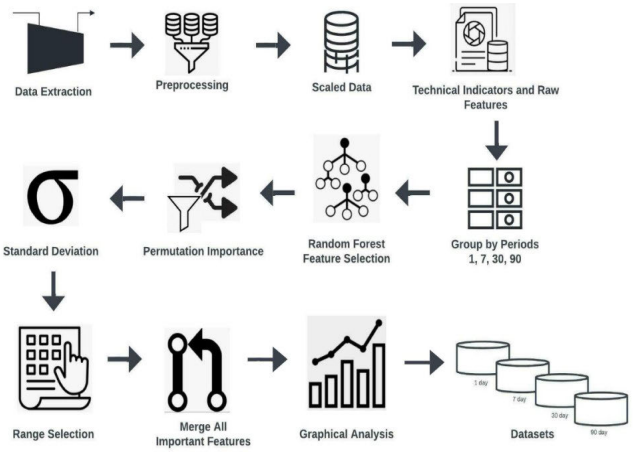

FIGURE 3. DL-based time-series forecast using technical indicators.

in the document, a set of features were chosen. Simple Moving Average (SMA), Exponential Moving Average (EMA), Relative Strength Index (RSI), Weighted Moving Average (WMA), Standard Deviation (STD), Variance (VAR), Triple Moving Exponential (TRIX), and Rate of Change(ROC) are some of the technical indicators.

These technical indicators are calculated over three, seven, fourteen, thirty, and ninety days.The end-of-day closing prices are regarded as raw values. Table 1 lists the raw attributes that these technical indicators are based on. Technical indicators display properties not immediately apparent in raw data, such as variances and standard deviations as a function of time. These technical indicators were created to display these aspects in the BTC and ETH price time-series. For example, instead of merely showing the raw transactions and hashrates, they illustrate how the ETH or BTC price is related to the standard deviation of the transactions or hashrate across 14 day intervals. The process of data extraction and processing is shown in Fig-3.

We also collected a new test-set for the two currencies from January 01, 2020 to January 01, 2022 and applied our proposed forecasting methodology.

B. PREPROCESSING

In Data preprocessing a number of steps were done to clean and scale the data. Sklearn Simple imputer was used to fill the missing data in the dataframe using the most frequently used strategy and fit transform together to increase model accuracy.The cleaned data was then turned into a data frame, resulting in a two-dimensional data set with rows and columns.Using the sklearn library function train test split, the data was split into test and train groups, with the test size set to 33% of the total data and random state of 42 to control the shuffling before splitting data. The features were scaled using the minmax scaler for training the LSTM model. The features are shifted between 0 and 1 with the minmax scaling, while the relative magnitude of the outliers is preserved. The shape of the original distribution is preserved by MinMaxScaler. It has no effect on the information included in the original data.

C. FEATURE SELECTION

Feature selection improves the machine learning process and increases the predictive power of machine learning algorithms by selecting the most important variables and eliminating redundant and irrelevant features.It is necessary to improve model performance. Several alternative approaches were used to extract and prune features iteratively.Firstly, RandomForestRegressor was initialized to fit a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.The bootstrap was set to False thus the whole dataset was used to build each tree and the n_estimator was set to default.

Next Impurity-based feature importance of RandomForestClassifier was implemented followed by Sklearn's permutation_importance feature selection to improve accuracy and select best features to be fed to the model.Permutation_Importance returns us importances_mean which is stored in a sorted manner using the formula given below:

$$importances\_mean[i] - 2 * importances\_std[i] > 0, \quad (6)$$

where std stands for standard deviation. If the condition is met then the feature is appended into the selected list of features.

Random forests is an ensemble learning method for regression and other problems that works by training a large number of decision trees. The mean or average prediction of the individual trees is returned for regression tasks.A random forest, unlike a single decision tree, may forecast using hundreds of trees, yielding better results. It does not necessitate substantial training and is suitable for small datasets and quick evaluation. One of the most extensively used measures of feature relevance, Mean Decrease Impurity (MDI), wrongly allocates high importance to noisy features, resulting in systematic bias in feature selection [29].The improvement in the "Gini gain" splitting criterion, which contains a weighted mean of the individual trees' improvement in the splitting criterion created by each variable, is described by the mean decrease in impurity (Gini) significance metric. The Gini impurity index is calculated as follows:

$$G = \sum_{i=1}^{n_c} p_i(1 - p_i) = 1 - \sum_{i=1}^{n_c} p_i^2 \quad (7)$$

where nc represents the number of classes in the target variable and pi represents the class ratio. In other words,

TABLE 2. Raw Indicators.

Features	Description
Transactions	The number of sent and received Bitcoin payments
Block size	Transactional information cryptographically linked in the blockchain. The maximum block size is currently set at 1 megabyte
Sent from addresses	These are distinct Bitcoin addresses from which payments are made everyday
Difficulty	The daily average mining difficulty. The difficulty is computed by the network after a specified number of blocks have been created so that the time it requires to mine a block remains around 10 min
Mining profitability	The profitabilityi in USD/day for 1 terahash per second (THash/s)
Sent BTC	The total Bitcoins sent daily
Fee-to-reward ratio	The ratio of the fee sent in a transaction to the reward for verifying that transaction by the other users
Median transaction fee	The median of transaction fees in Bitcoin
Average transaction fee	Each transaction can have an associated transaction fee determined by the sender. The transaction fee is received by the miners who verify the transaction. Transactions with higher fees incentivize the Bitcoin miners to process them sooner than transactions with lower fees
Block time	The time required to mine one block. Usually, it is around 10 min but can fluctuate depending on the hashrate of the network
Hashrate	The daily total computational capacity of the Bitcoin network. Hashrate indicates the speed of a computer in completing an operation
Median transaction value	The median value of the transactions in Bitcoin
Active addresses	The number of unique addresses participating in a transaction by either sending or receiving Bitcoins
Top 100 to total	The ratio of Bitcoins stored in the top 100 accounts to all the other accounts of Bitcoin
Average transaction value	The average value of the transactions in Bitcoin

it measures the disorder of a set of elements. It is calculated as the probability of mislabeling an element assuming that the element is randomly labeled according to the distribution of all the classes in the set.

D. BiDirectional LSTM, MACHINE LEARNING MODEL

We predicted the bitcoin and Ethereum prices using the BiDirectional LSTM model. An LSTM network is a type of recurrent neural network (RNN) that can learn long-term dependencies between time steps of sequence data. This deep learning model is very beneficial for time-series data modeling and forecasting. Because the daily Bitcoin and Ethereum price and its characteristics are time series data, LSTM may be used to make price forecasts. The diagram below illustrates the architecture of a simple LSTM network for regression. The network starts with a sequence input layer followed by an LSTM layer. The network ends with a fully connected layer and a regression output layer.A common LSTM unit is composed of a cell, an input gate (i), an output gate (o) and a forget gate (f). The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell. Fig. 5 demonstrates How Bidirectional LSTM work. The equations for the gates, cell state, candidate cell state and the final output in LSTM are given by:

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i) \tag{8}$$

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o) \tag{9}$$

$$tildec_t = tanh(w_c[h_{t-1}, x_t] + b_c) \tag{10}$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \tag{11}$$

$$h_t = o_t * tanh(c^t) \tag{12}$$

where:
$i_t \rightarrow$ represents input gate.
$f_t \rightarrow$ represents forget gate.
$o_t \rightarrow$ represents output gate.
$\sigma \rightarrow$ represents sigmoid function.

FIGURE 4. LSTM network.

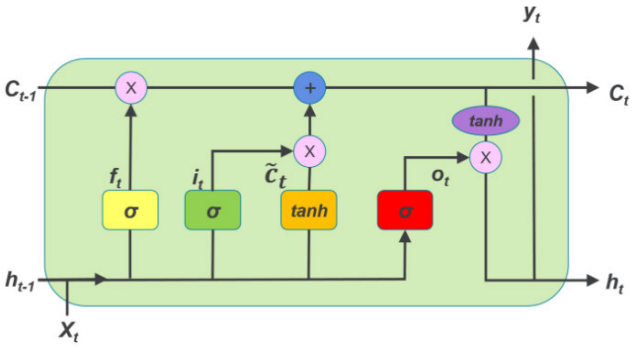

FIGURE 5. LSTM Architecture.

$h_{t-1} \rightarrow$ output of the previous lstm block (at timestamp (t-1)).
$x_t \rightarrow$ input at current timestamp.
$c_t \rightarrow$ cell state(memory) at timestamp(t).
$\tilde{c} \rightarrow$ represents candidate for cell state at timestamp(t).

E. CUSUM CONTROL CHART

A successful forecasting system design includes developing and executing processes to monitor the performance of the forecasting model. Regardless of how well the forecasting model initially performs, it is expected that its performance may decline over time. The time series pattern may shift due to internal inertial forces. The forces that drive the process may change over time or as a result of external factors and additional clients entering the market. A change in the level or slope of the variable being projected, for example, could occur. It's also feasible that the data's intrinsic

unpredictability will rise. As a result, performance monitoring is critical.

A cumulative sum (CUSUM) chart is a type of control chart used to monitor small shifts in the process mean. It uses the cumulative sum of deviations from a target. The CUSUM chart plots the cumulative sum of deviations from the target for individual measurements or subgroup means [30]. The CUSUM is highly useful in detecting changes in the monitored variable's level. It operates by accumulating deviations above the desired goal value T0 (typically zero or the average forecast error) with one statistic C+ and deviations below the target with another statistic C. The higher and lower CUSUMs are the statistics C+ and C, respectively [31].

For implementing CUSUM on our models firstly variance of the entire list of predicted values was calculated along with standard-deviation(std). After which the std was divided by the number(n) of values predicted and boundary limits were defined by multiplying the standard-deviation with 3. A graph represented in figure 6 was plotted representing all predicted values along with upper and lower limit to determine how many values fell in range.Depending on the number of values in range it is decided where to train the model again with recent data to get improved accuracy and precision in predicted values.

IV. RESULTS

In this section, the results of prediction models of Ethereum and Bitcoin are summarized.

The following measures are used to assess the performance of regression models: mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE). It is indeed ideal to have a model with low MAE, MAPE, and RMSE. These measurements show us how accurate our forecasts are and how far off they are from the actual data. A model that predicts unpredictable values on occasion will have a larger RMSE value, even if it has a lower MAE or MAPE. As a result, all three measures are considered when evaluating the models.The Root of the Mean of the Square of Errors is RMSE, while the Mean of Absolute Value of Errors is MAE. The disparities between the predicted values (values predicted by our regression model) and the actual values of a variable are called errors in this case.While MAPE is the sum of the individual absolute errors divided by the demand (each period separately). It is the average of the percentage errors. They're calculated as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{13}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{14}$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \tag{15}$$

TABLE 3. Bitcoin results.

Interval	MAE	RMSE	MAPE	R^2
Interval 1	2.40124	3.49908	0.786439	0.999652
Interval 2	3.05531	5.07083	0.723439	0.999621
Interval 3	3.70745	6.64741	0.66458	0.999846

TABLE 4. Ethereum results.

Interval	MAE	RMSE	MAPE	R^2
Interval 1	0.0626881	0.0945622	4.9981	0.999244
Interval 2	0.201017	0.332451	3.82481	0.998378
Interval 3	1.60717	3.02708	2.98182	0.999817

TABLE 5. Test-set results.

Interval	RMSE
Test-set(Average without Model Bias Correction)	209.1921
Test-set(Average with Model Bias Correction)	9.1873

where,

$$y_i = actual\ value$$
$$y_p = predicted\ value$$
$$n = number\ of\ observations\ rows$$

Table 2 and Table 3 show the results of the regression models for the three intervals for Bitcoin and Ethereum. BTC prices did not encounter considerable volatility in Interval I, from April 2013 to April 2016, as seen in Fig. 1. MAPE of 0.78, ii) RMSE of 3.49, and iii) MAE of 2.40 were recorded by the LSTM model during this interval. In Interval I, [7] reported a MAPE of 0.93, RMSE of 3.01, and MAE of 2.20 for their LSTM model.

Interval II, which runs from April 2013 to April 2017, has substantially higher BTC prices near the end, although it is reasonably stable similar to Interval I. It reported a 0.72 MAPE, RMSE of 5.07, and MAE of 3.05.In comparison, MAPE of 1.98, RMSE of 10.55 and MAE of 6.55 were reported by [7].

After April 2017, BTC prices had the most volatility, which is detailed in Interval III (April 2013 to December 2019).In this interval LSTM model reported i)MAPE of 0.66 ii)RMSE of 6.64 and iii)MAE of 3.70.In this interval [7] using LSTM model reported a i)MAPE of 3.61 ii)RMSE of 135.76 and MAE of 62.90. The results for Ethereum are displayed in Table 3.

V. DISCUSSION

The actual price, as well as variation in the prices are the two key factors of Bitcoin and Ethereum price modeling. As demonstrated in the study, the former can be accomplished with reduced error percentages; however, the latter remains an open challenge. The cryptocurrency prices are highly

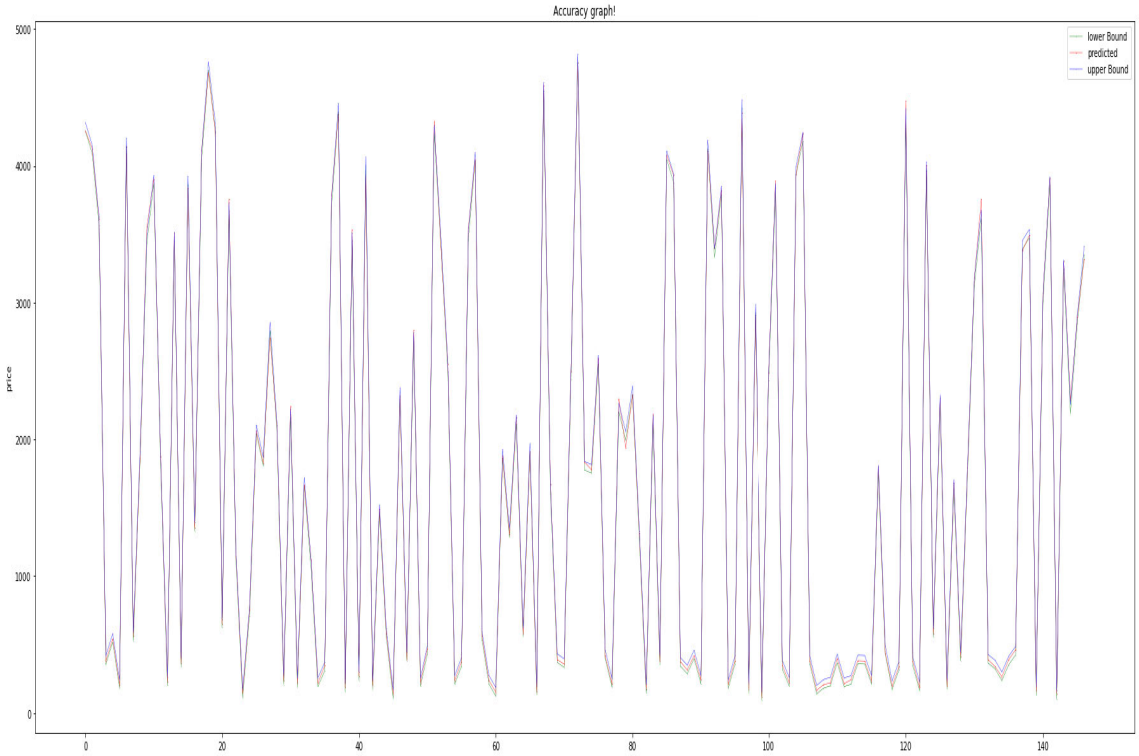

FIGURE 6. Cusum Graph.

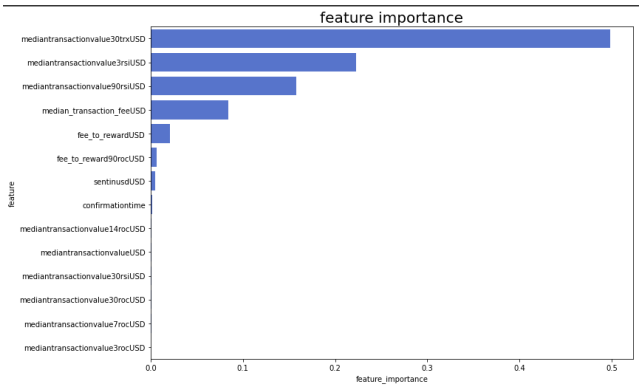

FIGURE 7. Importance features of Ethereum Jan2021-Jan2022.

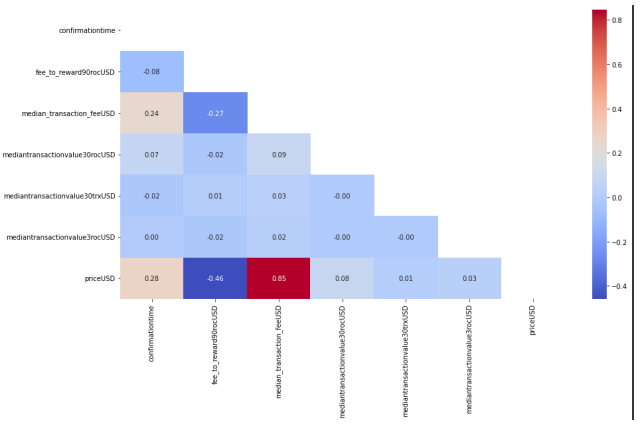

FIGURE 8. Feature importance HeatMap Ethereum Jan2021-Jan2022.

volatile, hence the fluctuations. Among these, Bitcoin and Ethereum prices vary more rapidly due to the sentiments of stakeholders and users, accompanied by their growing supply and demand in the cryptomarket. Therefore, their prices are purely random in nature and no individual set of characteristics can be used to predict their future prices. Nonetheless, researchers have made significant progress in estimating the Bitcoin and Ethereum values on the basis of assorted feature sets. This paper includes indicators that are directly related to blockchain. For example, from an investor's perspective, the open, high, low and close values of the two cryptocurrencies will be the most effecting attributes. For mining purposes, difficulty and hashrate are the key features. A trader; however, will be more interested in the moving averages such as SMA, EMA and WMA. All these can be classified as time-series characteristics. Technical indicators, on the other hand, helps in mapping these swift changing raw attributes into simple time-series attributes to estimate baselines. When technical markers spanning multiple time periods are merged, a huge feature set suitable for machine learning is created.

In order to find useful features, it is important to shortlist indicators whilst preserving the shape of original distribution. This is achieved by the MinMax scaler that scales the values in the range of 0 to 1, thereby optimizing the process. The

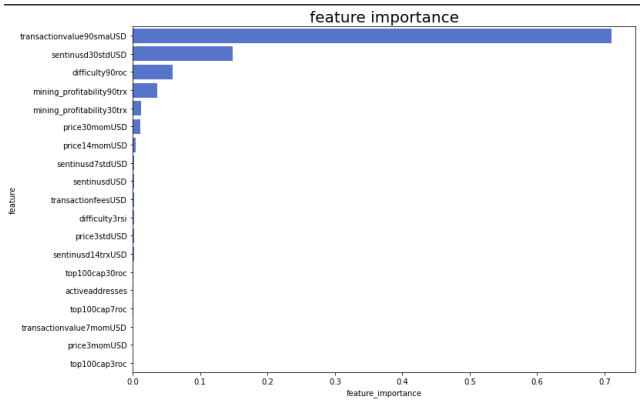

FIGURE 9. Importance features of Bitcoin Jan2021-Jan2022.

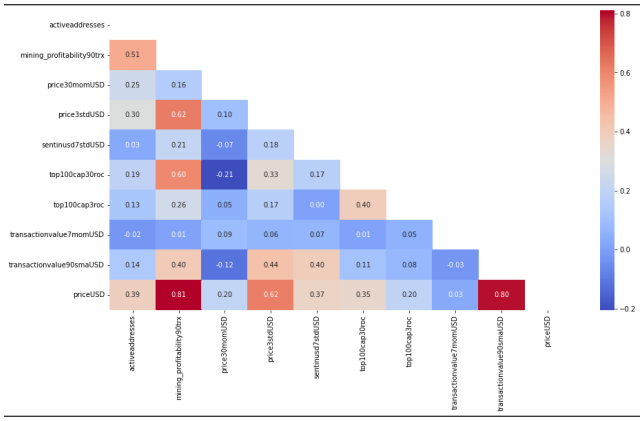

FIGURE 10. Feature importance HeatMap Bitcoin Jan2021-Jan2022.

feature selection method discussed is systematic and eliminates the highest ranked elements since those attributes are known to be adding noise to data. This helps in boosting up the performance of the model as the optimal indicators are automatically selected thus, computationally faster. MDI approach helped in achieving this benchmark by ranking attributes on the basis of impurity and leaving out relevant features only.

LSTM networks are a kind of Recurrent Neural Networks (RNN) that have the ability to work with sequence dependent problems. Bi-directional LSTMs are an improvement on traditional LSTM models, which allows it to have comprehensive and complete information at any point in the given sequence, whether it is prior or later to it. This enables it to access long-range context in either direction and exhibits better predictions; therefore, used as the only modeling technique followed by CUSUM control charts. This control chart helps in monitoring even the slightest shift in the value, consequently allowing us to determine a thresholding value in the estimated price ranges of Bitcoin and Ethereum, directing to better performance of the model in future. On a out-of-sample dataset, that we collected for the two cryptocurriences from January 01, 2020 to January 01, 2022. Our model without bias corrections using produced RMSE of 209.1921, through the effective use of CUSUM model bias correction it reduced to average RMSE of 9.1873 which is a clear indication that our model is able to adjust model bias very accurately.

VI. CONCLUSION

In this paper, we leverage the LSTM model to forecast Bitcoin and Ethereum prices for short-term and mid-term. It implements a deep learning model to anticipate remarkably precise prices of Bitcoin and Ethereum at the end of the day, short-term (1, 7 and 14 days) and mid-term (30, 60 and 90 days). LSTM outperforms other machine learning and deep learning algorithms, hence used in this study. Two approaches for feature selection are implemented- wrapper and MDI approach. The indicators shortlisted by MDI approach results in improved accuracy with the same model. For further improvisation, we use the CUSUM control chart for monitoring the shift in the prices of both cryptocurrencies mentioned above. The overall accuracy for forecasting Bitcoin prices using these improvement measures was up to 73.3%. The RMSE and MAPE are as low as 6.67% and 0.66%, respectively. Ethereum price forecasting resulted in accuracy up to 54.%. The RMSE and MAPE are as low as 3.03% and 2.98%, respectively. The results of performance evaluation metrics show significant improvement in the recent literature in predicting both, fluctuation in prices and daily closing price. The results are promising and can be made use of in a multitude of sectors, including blockchain and AI research.

Our findings reveal that while it is possible to make a satisfactory estimate of Bitcoin and Ethereum prices, forecasting the swing in movement, whether increase or decrease, is far more challenging. This research represents finest performance scores with the MDI approach using the LSTM model. However, both feature selection approaches and classification models need to be investigated deeper. Instead of using daily prices for historic data, prices and other technical attributes on hourly basis can be extracted and used. Moreover, ensemble learning techniques can be used to improve the forecasting results in the future. The use of CUSUM control charts in this research is limited to monitoring the changes affecting our model. Additionally, determining the exact features and making changes taking them into account can be implemented in the future too. Apart from this, historic data for relevant technical indicators of other cryptocurrencies such as Tether, Ripple and XRP can also be used to forecast their prices.

On the basis of this study, subsequent work that can be done is using artificial intelligence and machine learning for cryptocurrency price prediction for gauging potential risk in financial technologies. This can help in anomaly and fraud detection. Money laundering in cryptocurrencies can be seen to have currently spiked. Thus, our future work aims to focus on the mentioned domain by considering financial threats that may be faced in the crypto world in order to control the rise in such unethical activities.

REFERENCES

[1] I. Peña-López. (2017). *Digital Spillover. Measuring The True Impact Of The Digital Economy*. [Online]. Available: http://www.huawei.com/minisite/gci/en/digitalspillover/files/gcidigitalspillover.pdf

[2] I. Bashir, *Mastering Blockchain: Distributed Ledgers, Decentralization and Smart Contracts Explained*. Birmingham, U.K.: Packt, 2017.

[3] D. Unal, M. Hammoudeh, and M. S. Kiraz, "Policy specification and verification for blockchain and smart contracts in 5G networks," *ICT Exp.*, vol. 6, no. 1, pp. 43–47, Mar. 2020.

[4] M. Ertz and É. Boily, "The rise of the digital economy: Thoughts on blockchain technology and cryptocurrencies for the collaborative economy," *Int. J. Innov. Stud.*, vol. 3, no. 4, pp. 84–93, Dec. 2019.

[5] V. Buterin. (2013). *Ethereum White Paper: A Next Generation Smart Contract & Decentralized Application Platform*. [Online]. Available: https://github.com/ethereum/wiki/wiki/White-Paper

[6] D. Vujicic, D. Jagodic, and S. Randic, "Blockchain technology, bitcoin, and ethereum: A brief overview," in *Proc. 17th Int. Symp. INFOTEH-JAHORINA (INFOTEH)*, Mar. 2018, pp. 1–6.

[7] M. Mudassir et al., "Time-series forecasting of Bitcoin prices using high-dimensional features: A machine learning approach," *Neural Comput. Appl.*, 2020, doi: 10.1007/s00521-020-05129-6.

[8] S. Asante Gyamerah, "Are bitcoins price predictable? Evidence from machine learning techniques using technical indicators," 2019, *arXiv:1909.01268*.

[9] J.-Z. Huang, W. Huang, and J. Ni, "Predicting bitcoin returns using high-dimensional technical indicators," *J. Finance Data Sci.*, vol. 5, no. 3, pp. 140–155, Sep. 2019, doi: 10.1016/j.jfds.2018.10.001.

[10] R. Adcock and N. Gradojevic, "Non-fundamental, non-parametric bitcoin forecasting," *Phys. A, Stat. Mech. Appl.*, vol. 531, Oct. 2019, Art. no. 121727, doi: 10.1016/j.physa.2019.121727.

[11] D. Philippas, H. Rjiba, K. Guesmi, and S. Goutte, "Media attention and bitcoin prices," *Finance Res. Lett.*, vol. 30, pp. 37–43, Sep. 2019, doi: 10.1016/j.frl.2019.03.031.

[12] J. Abraham, D. Higdon, J. Nelson, and J. Ibarra, "Cryptocurrency price prediction using tweet volumes and sentiment analysis," *SMU Data Sci. Rev.*, vol. 1, no. 3, p. 1, 2018.

[13] D. Shen, A. Urquhart, and P. Wang, "Does Twitter predict bitcoin?" *Econ. Lett.*, vol. 174, pp. 118–122, Jan. 2019, doi: 10.1016/j.econlet.2018.11.007.

[14] N. Aslanidis, A. F. Bariviera, and Ó. G. López, "The link between cryptocurrencies and Google trends attention," *Finance Res. Lett.*, vol. 47, Jun. 2022, Art. no. 102654, doi: 10.1016/j.frl.2021.102654.

[15] Y. B. Kim, J. G. Kim, W. Kim, J. H. Im, T. H. Kim, S. J. Kang, and C. H. Kim, "Predicting fluctuations in cryptocurrency transactions based on user comments and replies," *PLoS ONE*, vol. 11, no. 8, Aug. 2016, Art. no. e0161197, doi: 10.1371/journal.pone.0161197.

[16] N. A. Bakar and 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.*, vol. 4, no. 11, pp. 130–137, 2017, doi: 10.22161/ijaers.4.11.20.

[17] Z. H. Munim, M. H. Shakil, and I. Alon, "Next-day bitcoin price forecast," *J. Risk Financial Manage.*, vol. 12, no. 2, p. 103, Jun. 2019, doi: 10.3390/jrfm12020103.

[18] H. Jang and J. Lee, "An empirical study on modeling and prediction of bitcoin prices with Bayesian neural networks based on blockchain information," *IEEE Access*, vol. 6, pp. 5427–5437, 2018, doi: 10.1109/access.2017.2779181.

[19] Z. Chen, C. Li, and W. Sun, "Bitcoin price prediction using machine learning: An approach to sample dimension engineering," *J. Comput. Appl. Math.*, vol. 365, Feb. 2020, Art. no. 112395, doi: 10.1016/j.cam.2019.112395.

[20] D. C. A. Mallqui and R. A. S. Fernandes, "Predicting the direction, maximum, minimum and closing prices of daily bitcoin exchange rate using machine learning techniques," *Appl. Soft Comput.*, vol. 75, pp. 596–606, Feb. 2019, doi: 10.1016/j.asoc.2018.11.038.

[21] A. Greaves and B. Au, "Using the bitcoin transaction graph to predict the price of bitcoin," Stanford Univ., Stanford, CA, USA, Tech. Rep., 2015.

[22] S. McNally, J. Roche, and S. Caton, "Predicting the price of bitcoin using machine learning," in *Proc. 26th Euromicro Int. Conf. Parallel, Distrib. Netw.-Based Process. (PDP)*, Mar. 2018, pp. 339–343, doi: 10.1109/pdp2018.2018.00060.

[23] M. J. Hamayel and A. Y. Owda, "A novel cryptocurrency price prediction model using GRU, LSTM and bi-LSTM machine learning algorithms," *AI*, vol. 2, no. 4, pp. 477–496, Oct. 2021, doi: 10.3390/ai2040030.

[24] S. Biswas, M. Pawar, S. Badole, N. Galande, and S. Rathod, "Cryptocurrency price prediction using neural networks and deep learning," in *Proc. 7th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, vol. 1, Mar. 2021, pp. 408–413, doi: 10.1109/icaccs51430.2021.9441872.

[25] S. Tanwar, N. P. Patel, S. N. Patel, J. R. Patel, G. Sharma, and I. E. Davidson, "Deep learning-based cryptocurrency price prediction scheme with inter-dependent relations," *IEEE Access*, vol. 9, pp. 138633–138646, 2021, doi: 10.1109/access.2021.3117848.

[26] S. E. Charandabi and K. Kamyar, "Prediction of cryptocurrency price index using artificial neural networks: A survey of the literature," *Eur. J. Bus. Manage. Res.*, vol. 6, no. 6, pp. 17–20, Nov. 2021.

[27] S. Marne, S. Churi, D. Correia, and J. Gomes, "Predicting price of cryptocurrency—A deep learning approach," in *Proc. NTASU Conf.*, vol. 9, no. 3, 2020, pp. 1–7.

[28] P. Jaquart, D. Dann, and C. Weinhardt, "Short-term bitcoin market prediction via machine learning," *J. Finance Data Sci.*, vol. 7, pp. 45–66, Nov. 2021, doi: 10.1016/j.jfds.2021.03.001.

[29] X. Li, Y. Wang, S. Basu, K. Kumbier, and B. Yu, "A debiased mdi feature importance measure for random forests," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.

[30] V. Koshti, "Cumulative sum control chart," *Int. J. Phys. Math. Sci.*, vol. 1, no. 1, pp. 28–32, 2011.

[31] D. C. Montgomery, C. L. Jennings, and M. Kulahci, *Introduction to Time Series Analysis and Forecasting*. Hoboken, NJ, USA: Wiley, 2015.

MUHAMMED RAFI was born in Karachi, Pakistan. He received the B.S. and M.S. degrees in computer science from the FAST Institute of Computer Science, University of Karachi, Karachi, in 1996 and 2000, respectively, and the Ph.D. degree in computer science, in 2017. He has more than ten years of experience in software development and also working as a consultant for the local software industry. His current research interests include algorithm development, machine learning, information retrieval, text/data mining, time series analysis, and natural language processing. He has received several travel grant awards for presenting his work at the top conferences. He has served as a judge and technical quality review team at many versions of IEEEXtreme Programming competitions. He has served as a reviewer in various international journals of high impact.

QUBLAI ALI KHAN MIRZA received the Ph.D. degree in network security and machine learning from the University of Bradford. He is currently a Lecturer of cyber security with the School of Business and Technology, University of Gloucestershire, U.K. He has more than ten years of experience in software development, network security, and cloud migration. He is also an experienced Academician/Researcher. His research interests include network security, malware analysis, network analytics, cryptography, cloud computing, and the IoT security. He is a fellow of HEA. He has reviewed several international journals and published in several IEEE conferences and journals.

MUHAMMAD IZAAN SOHAIL was born in Karachi, Pakistan. He received the B.S. degree in CS from FAST-NUCES. He holds vast experience working with blockchain technology and machine and deep learning technology. He holds more than two years of experience as a Software Developer. He is currently working as a Software Engineer for multiple Javascript and PHP-based frameworks. His current research interests include using blockchain as a security protocol and as part of the Information of Things networks.

MARIA ALIASGHAR was born in Karachi, Pakistan. She received the B.S. degree in CS from the FAST-Institute of Computer Science, in 2022. She is interested in exploring the spectrum of artificial intelligence, robotics, machine learning, and natural language processing. She has extensive experience working with machine learning models and Javascript-based frameworks. She is currently working as an Associate Software Engineer at 10Pearls working on various development projects with the goal of becoming an expert in the field. She received the Bronze Medalist.

ARISHA AZIZ was born in Karachi, Pakistan. She received the B.S. degree in computer science from FAST-NUCES, in 2022. She is very motivated and enthusiastic about working with data and has extensive experience in AWS. She is currently employed as a Software Engineer in big data and analytics platforms, where she works closely with data scientists and engineers from other teams. She holds past practical experiences working with artificial intelligence, machine learning, deep learning, and computer vision technologies, and is keen on working with cutting-edge technology and systems. She is an active member of the Medium blog and promotes open-source contributions for the benefit of the tech community.

SUFIAN HAMEED received the Ph.D. degree in networks and information security from the University of Göttingen, Germany. He works as an Associate Professor at the Department of Computer Science, National University of Computer and Emerging Sciences, Pakistan. He also leads the IT Security Laboratories, NUCES. The research laboratory studies and teaches security problems and solutions for different types of information and communication paradigms. His research interests include network security, web security, mobile security, and secure architectures and protocols for Cloud and the IoTs.