STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **12**, Month 2024, pp 0–18.
Published online in International Academic Press (www.IAPress.org)

Predicting the closing price of cryptocurrency Ethereum

Vhukhudo Ronny Rambevha, Caston Sigauke, Thakhani Ravele*

Department of Mathematical and Computational Sciences, University of Venda, South Africa

Abstract Given that cryptocurrencies are now involved in nearly every financial transaction due to their widespread acceptance as an alternative method of payment and currency exchange, researchers and economists have increased opportunities to analyze cryptocurrency prices. Over time, predicting the daily closing price of Ethereum has been challenging for investors, traders, and investment banks because of its significant price volatility. The daily closing price of cryptocurrency is crucial for trading or investing in Ethereum. This report aims to conduct a comparative analysis of the predictive performance of deep machine learning algorithms within a stacking ensemble modeling framework, utilizing daily historical price data of Ethereum from Coindesk, tweets from Twitter spanning from August 1, 2022, to August 8, 2022, and five additional covariates (closing price lag1, closing price lag2, noltrend, daytype, and month) derived from Ethereum's closing price. Seven models are employed to forecast the daily closing price of Ethereum: recurrent neural network, ensemble stacked recurrent neural network, gradient boosting machine, generalized linear model, distributed random forest, deep neural networks, and a stacked ensemble of gradient boosting machine, generalized linear model, distributed random forest, and deep neural networks. The primary evaluation metric is the mean absolute error (MAE). Based on MAE, the RNN forecasts outperform the other models in this study, achieving an MAE of 0.0309.

Keywords Cryptocurrency; Ethereum; Machine learning models; Natural language processing; Recurrent neural network.

AMS 2010 subject classifications: 62M20, 91G70, 91B82

DOI: 10.19139/soic-2310-5070-2076

1. Introduction

1.1. Overview

The word cryptocurrency has become the topic of the 21st century. Many investors trade cryptocurrency, which has become a new financial product. The cryptocurrency was first adopted as a gaming transaction in 2014. Middlebrook [1] suggested that large corporations are indicating that they may legally accept it as an exchange of their goods and services.

Mukhopadhyay et al. [2], cryptocurrency is a digital currency (digital money) in which transactions are verified, and records use a decentralized system. This decentralized system is called a peer-to-peer network and operates as cryptography, ensuring that no one controls the system, neither the government nor rich individuals. The decentralized system differs from the traditional monetary system, whereby a central bank controls a currency. To ensure security and fairness amongst the users, a well-structured complex encryption hashing algorithm constructed from the basis of blockchain technology is put in place.

Blockchain technology was not widely used until a mysterious individual who used Satoshi Nakamoto's pseudonym created the first cryptocurrency, famously known as Bitcoin, in 2009. Due to the hidden identity of the author of Bitcoin, the cryptocurrency environment has been deemed by many as an illegal way of transacting,

*Correspondence to: Thakhani Ravele (Email: ravelethakhani@gmail.com). Department of Mathematical and Computational Sciences, University of Venda, Private Bag X5050, Thohoyandou, 0950, South Africa.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

resulting in high volatility and the cryptocurrency crashing several times. Bitcoin may be the largest cryptocurrency by market value, but it is not the only cryptocurrency in the market.

Ethereum is the software platform enabling cryptocurrency transactions ether'; Ethereum is the second-largest cryptocurrency launched by programmer Vitalik Buterin on the 30^{th} of July 2015 (Warner [3]). On the 7^{th} of May 2021, Ethereum had a market capitalisation of $\$410\ Billion$ according to Times [4]. Two weeks later, the new market capitalization is $\$282\ Billion$. There was a dramatic change in price ether between the 7^{th} of May 2021 and 21^{st} of May 2021. This raised concerns for investors about whether it is a reliable asset due to its high fluctuation.

A much more informal name for machine learning is 'the prophet method', as it is known for its good learning algorithms for making future predictions. The development of machine learning algorithms dates back to the 1970s by experimenting with different architectures of neurons (Shavlik et al. [5]). Machine learning is defined as the ability of a computer to learn a task without explicitly following instructions, guided by algorithms, and then be able to extract meaningful conclusive information (Anderson [6]). Machine learning may either be supervised, semi-supervised or unsupervised.

This incredible learning ability demonstrated by computers leads to the term artificial intelligence. Today, several methods have been developed that one can use to apply artificial intelligence. These include recurrent neuron networks (RNN), gradient boosting machines (GBM), generalized linear models (GLM), distributed random forests (DRF) and deep neural networks (DNN). This paper will use RNN, GBM, GLM, DRF and DNN to predict the daily closing price of Ethereum.

Natural language processing (NLP), which was formally known as natural language understanding (NLU), is a subset of artificial intelligence (AI) used to analyze text through a set of computerized technologies and theories that are put together to imitate human-like language processing (Liddy [7]). NLP methods can process oral or written texts. For this research, only written texts will be processed using NLP methods.

When utilizing social media data, it is important to consider the ethical ramifications of social media analytics for businesses as well as the intended users of the data. Liddy [8] studied predictive risk intelligence using social media (YouTube, Twitter, and local social networks and news media) as an input. This study noted that after reviewing the inputs from social media, some ethical concerns with the use of AI algorithms in predictive intelligence, such as integrity, security and privacy, transparency and algorithmic bias, were established.

We are living in an age where digital communication is thriving. Twitter has been one of the most successful and popular social media of the 21st Century. As of 2021, according to Twitter company metrics, Twitter had 221 million total monetizable daily active users, with more than 500 million tweets sent daily in the fourth quarter of 2021. These users are worldwide, tweeting about different subjects, including many cryptocurrencies traded daily on crypto-exchange platforms.

Ethereum is the world's second-largest cryptocurrency after Bitcoin, which has a much lower value, indicating that it can be used for day-to-day transactions; thus, it is important to predict its closing price due to its rapid price fluctuation. However, previous studies fail to predict Ethereum's closing price accurately. To date, it has attracted investors and risk portfolio managers, predicting its closing price will be a good measure of risk and give investors confidence to invest in the cryptocurrency.

When trading securities such as Ethereum, investors, traders, and investment banks need to know how profitable it is to buy, hold or sell Ethereum. This study intends to forecast the daily closing price of Ethereum using machine learning algorithms RNN, GBM, GLM, DRF and DNN with the help of Ethereum tweets from Twitter.

1.2. Literature review

Predicting the stock market volatility can be dated back to 1995 as a crucial decision-making tool. Fleming et al. [9] aims to predict the stock market volatility. The reason is that stock market volatility, such as portfolio insurance, is essential for investment decision-makers. The study is structured by executing three objectives. Firstly, they examined the properties of the Chicago Board Options Exchange Volatility Index (VIX) and evaluated its predictive power. Secondly, they studied the correlation between volatility and stock market returns and how well VIX predicts stock market volatility. The VIX is modelled using the Black-Scholes framework using the implied volatility of eight Chicago Board Options Exchange options. The results show that VIX strongly correlates with the expected stock market returns, indicating that it is a good measure.

Other variables other than price returns have been used to forecast price volatility. Jain and Jiang [10] aim to predict future price volatility using the limit order book (LOB) from the Shanghai Stock Exchange (SHSE). The data is obtained from SHSE from January 2009 to December 2009. A LOB slope was constructed to be used in the prediction process. It is concluded that LOB efficiently predicts price volatility with the limitations of poor performance during major market-wide movements. The underlying factor of efficiently predicting price volatility is the high correlation between buy orders and future price volatility.

If the security is big enough for market capitalization, its price might impact the stock market volatility. Tang et al. [11] investigated the oil future price predictability power towards the United States (US) market volatility. The autoregressive conditional heteroskedasticity methodology is applied to construct models. Data is obtained from the Thomson Reuters Tick History Database from January 2007 to April 2017. It is concluded that oil future price predictability significantly predicts the US stock market.

Generalized autoregressive conditional heteroskedasticity (GARCH) models are less computationally expensive than ANN (artificial neural network) models; however, the rapid development of computers allows us to compare the best model between the two in predicting historical volatility. Lahmiri [12] investigated the best approach for forecasting currency exchange rate volatility. The data, which consists of daily exchange rates from January 2010 to December 2013, were collected from the federal economic database. The dataset is split into 80% for the training set and 20% for the testing set. The three models compared to each other are the GARCH, exponential-GRACH, and ANN. The ANN model outperforms the other models in predicting historical volatility because it produces the least mean square error.

When assessing a risk portfolio, one needs accurate technical measures. Combining models known to be great volatility predictors may be very fruitful. [13] investigated the effect of combining RNN with LSTM and multiple GARCH models to forecast volatility. The data is collected from the KOSPI 200 index returns between January 2001 and September 2011. The models used in the combination technique are exponential weighted moving average (EWMA), EGARCH and RNN with LSTM. The combined RNN model with LSTM and three GARCH-type models produced the best predictions with a mean absolute error of 0.0107.

The use of RNN predictive power works well with time-series data. Anbazhagan and Kumarappan [14] proposed predicting deregulated electricity market price for the next day in Spain. The architecture of RNN is Elamn Network, which proves to be very robust. The Elman network is compared to other models such as the autoregressive integrated moving average (ARIMA), weighted nearest neighbours, wavelet ARIMA, neural networks with wavelet transform and wavelet-ARIMA radial basis function neural networks. The results conclude that the proposed RNN with Elam Network is the most efficient model.

Derbentsev et al. [15] focused on predicting three cryptocurrencies using Random Forests (RF) and Stochastic Gradient Boosting Machine (SGBM). The three cryptocurrencies are Bitcoin (BTC), Ethereum (ETH) and Ripple (XRP). The dataset is their historical daily close prices. To check the effectiveness of these models, an out-of-sample forecast was made for the selected time series using the one-step ahead technique. For the three cryptocurrencies (BTC, ETH, and XRP), the out-of-sample accuracy of the short-term prediction daily close prices derived by the SGBM and RF fell between 0.92 and 2.61 in terms of Mean Absolute Percentage Error (MAPE). The outcomes confirm that the ML ensembles approach may be applied to forecast cryptocurrency prices.

Poongodi et al. [16] proposed using a time series made up of the closing values of the cryptocurrency Ether every day, two machine learning techniques—linear regression (LR) and support vector machine (SVM)— to predict daily closing prices. Filters with varying weight coefficients are employed to anticipate the price of ether cryptocurrency over a range of window lengths. Cross-validation is a technique used in the training phase to build a high-performance model that is not dependent on the dataset. The results showed that the SVM method has a higher accuracy (96.06%) than the LR method (85.46%).

Deep learning models are known to be good predicting models; recently, researchers have been investigating if they are better when stacked together. Livieris et al. [17] proposed ensemble models evaluated as combinations of long short-term memory (LSTM), Bi-directional LSTM and convolutional layers. The ensemble models were tested for regression (predicting the next hour's cryptocurrency price) and classification (predicting whether the price of a cryptocurrency will rise or fall in relation to the current hour). Empirical results from the study showed

that deep and ensemble learning can effectively support one another in creating robust, steady, and dependable forecasting models.

Henrique et al. [18] investigated the relationship between social media posts and the volatility price movement of cryptocurrency. This was achieved by analyzing the social media posts of a Chinese platform, Sina-Weibo, Sina-Weibo, WeChat, and QQ groups. Sina-Weibo can produce approximately 24000 accompanied by 70000 comments and tweets about cryptocurrency in just eight days. A sentiment dictionary is constructed to categorize three distinct moods from the tweet. These are bag holders, new highs, and abandoned ships. Combining these social media sentiments with an RNN with LSTM fueled by historical cryptocurrency price is more efficient in predicting the volatility price movement than the based ARIMA model by 18.5% accuracy.

Vadivukarassi et al. [19] investigated the polarity of tweets from Twitter as either positive or negative. Luo et al. [20] extracted the tweets from Twitter using Twitter API. A Chi-Square test and a naïve Bayes classifier are used for training and testing the model for selecting the best features to evaluate sentimental polarity using Python. Different features of s 10,100,1000,10000 were applied respectively. It was concluded that as the number of features increases, the accuracy of the selected feature also increases.

Giudici et al. [21] investigated the dynamics of cryptocurrency asset values, particularly how price data is shared between various bitcoin market exchanges and between bitcoin markets and conventional ones. This was accomplished using a preliminary filtering technique based on the random matrix approach to enhance the correlation-based minimum spanning tree method, which hierarchically clusters bitcoin prices from various exchanges and traditional assets. The primary empirical conclusions are as follows: (i) the prices of bitcoin exchanges are positively correlated with one another, with the biggest exchanges, like Bitstamp, driving the prices; (ii) the prices of bitcoin exchanges are unaffected by the prices of traditional assets, but their volatilities are, with a negative and lagging effect.

Khedr et al. [22] researched cryptocurrency price prediction, and during their explanatory analysis, [22] found no seasonal influences in cryptocurrency; statistical prediction methods are difficult to apply. Therefore, machine learning is advantageous in this industry since it can anticipate prices based on experience. At the same time, traditional statistical approaches involve many assumptions that may be impractical. [22] used data from 2010 to 2020 to study various papers using multiple methods, including traditional and machine learning methods. Although there are some challenges, there is room to improve. Machine learning methods perform better than traditional statistical methods.

Luo et al. [20] presented an NLP framework that uses sentiment analysis to analyze the opinions of Twitter users on Human papillomavirus (HPV) vaccination over ten years from 2008 to 2017. Luo et al. [20] used sentiment analysis and AI, amongst other methods, on the phrase $'associationmining'$ search through Twitter. The results showed that from 2008 to 2011 and 2015 to 2016. The top negative words were safety concerns, deaths, adverse/side effects, injuries and scandal, while the top positive words were cervical screens, prevents, vaccination campaigns and cervical cancers. The result from the sentiment analysis helped public health researchers gain a better understanding of the influence of social media on HPV vaccination attitudes and also develop strategies that will deal with misinformation.

The use of machine learning models and the application of NLP with naive Bayes classifying tweets into positive or negative news has been discussed in detail in the above literature review. This research seeks to study the effect of adding tweets as an additional variable in predicting the closing price of cryptocurrency.

1.3. Research highlights and contributions

The main contribution of this study is to carry out a comparative study of the predictive capabilities of some machine and deep learning algorithms with a stacking ensemble modelling framework of ETH closing prices for the next two days using historical Ethereum prices and Ethereum tweets. A summary of the research highlights and contributions are:

- Given the growing interest in cryptocurrency, the study addresses a relevant and timely topic in finance and technology.
- The comparative approach of evaluating multiple models comprehensively analyses their effectiveness in predicting cryptocurrency prices.

- Including natural language processing to analyse social media sentiment adds a unique dimension to the predictive model.

The rest of the paper is organized as follows. A discussion of the modelling closing price of Ethereum using the RNN, Stacked RNN, GBM, GLM, DRF and DNN, Stacked GBM, GLM, DRF and DNN (GGDD) is given in Section 2. Empirical results are presented in Section 3 while Section 4 presents a discussion of the performance of the models. Section 5 concludes.

2. Methods

2.1. Natural language processing Method

Naive Bayes is the NLP method for text (tweet) classification into good or bad news. The classified tweet will be treated as an additional explanatory variable coded 0 and 1 for good and bad news, respectively. The naive Bayes model is part of a group of generative models in NLP. The Naive Bayes model is given in equation (1) (Rish et al. [23]).

$$n(x) = argmax P(n(x) = K|X) = argmax P(X|n(x) = K) * P(n(x) = K), \tag{1}$$

where k is the class containing 1 for good news and 0 for bad news, and $X = (x_1, x_2, ..., x_n)$ is the feature vector. As input is embedded into the NLP model, it will go through the function TextBlob, a classifying technique for tweets into either good/positive news or bad/negative news. The TextBlob will compute the subjectivity and polarity of the tweets. If the polarity of the tweet is positive, it will be considered good news, and if it is negative, it will be considered bad news.

2.2. Recurrent Neural Network with LSTM

RNN is despised for having a gradient vanishing problem, which results in poor results. Through this challenge, a modified version of RNN with long short-term memory (LSTM) was developed by (Williams and Zipser [24]). RNN with LSTM is an advanced version of RNN that remembers past data in memory using three gates in each neuron in the hidden layer. The three gates are the input gate, forget gate and output gate. The input gate modifies the memory; the forget gate decides what details to discard from the block, and the output gate combines the values from the input gate and the forget gate.

RNN algorithm is implemented by initializing weights, forward propagation and backward propagations. After the output has been calculated from the forward propagation during training, the error between the predicted and actual values is used to adjust weights using mini-batch gradient descent. Adjusting these weights from the output layer to the first layer is called Backward propagation. Most RNNs primarily use the following activation functions: sigmoid, hyperbolic tangent (tanh) and Rectified linear unit (ReLu). These activation functions help prevent the gradient from exploding and vanishing.

Equation (2) defines the sigmoid activation.

$$S(x) = \frac{1}{1 + e^{-x}}. \tag{2}$$

The tanh function is defined by equation (3).

$$T(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}. \tag{3}$$

The ReLu function is given in equation (4).

$$R(x) = max(0, x). \tag{4}$$

2.3. Gradient Boosting Machine

The second methodology used is the Gradient Boosting Machine (GBM), a recommended algorithm, which Friedman [25] proposed. The GBM algorithm is a forward learning ensemble method driven by the principle that more accurate approximations can yield better prediction outcomes. Regression trees are progressively constructed on all of the dataset's characteristics by H2O's GBM in a completely distributed manner; each tree is constructed in parallel. Gradient boosting machines have demonstrated notable effectiveness in a variety of real-world applications. They can be easily tailored to meet specific application requirements and establish a connection with the statistical framework, such as learning various loss functions (Natekin and Knoll [26]). In order to produce a more precise response variable estimate, the GBM learning process fits new models one after the other. This technique's main concept is to build new base-learners with the highest possible correlation with the loss function's negative gradient (Natekin and Knoll [26]).

Because of their great adaptability, the GBMs may be tailored to almost any specific data-driven task. It adds great flexibility to the model design, so selecting the best loss function becomes a trial-and-error process. Given a dataset that has a response variable y and a set of explanatory variables $x = \{x_1, ..., x_n\}$ with a training sample $\{y_i, x_i\}_1^N$ of known $(y, x) - values$ (Friedman [25]). The GBM general formula to estimate the response variable is given in equation (5).

$$F(x; \{\beta_m, a_m\}_1^M) = \sum_{m=1}^{M} \beta_(m) h(x; a_m), \tag{5}$$

where function $h(x; a_m)$,is a simple parameterized function of the input variables x, characterized by parameters $a = \{a_1, a_2, ...\}$.

2.4. Generalized Linear Model

The third methodology that will be used is the Generalized Linear Model (GLM). Lee and Nelder [27] describe GLM as a family of models consisting of Gaussian regression, Poisson regression, Binomial regression (classification), Fractional binomial regression, Quasibinomial regression, Multinomial classification, Gamma regression, Ordinal regression, Negative Binomial regression and Tweedie distribution. Given the nature of our data, which is in an integer form, the evaluation model used is Gaussian regression.

According to Lee and Nelder [27], the dependence between a response vector (y) and a covariates vector (x) is modelled by Gaussian as a linear function given by the following function:

$$\hat{y} = x^T \beta + \beta_0, \tag{6}$$

where $x = \{x_1, ..., x_n\}$, $\beta = (x^T x)^{-1} y$ and β_0 is the error coefficient.

2.5. Distributed Random Forest

The Distributed Random Forest (DRF) will be the fourth methodology. Instead of producing a single classification or regression tree, Geurts [28] states that DRF creates a forest of them. Weak learners are based on a subset of rows and columns comprised of each tree. The variance will decrease with more trees. Whether predicting for a class or a numerical value, classification and regression both use the average prediction across all of their trees to arrive at a final prediction.

Cevid et al. [29] suggested a forest design for multivariate responses based on their joint conditional distribution, independent of the estimation target and the data model. If we let $Y = (Y_1, Y_2, ..., Y_d) \in \mathbb{R}^d$ be a multivariate random variable representing the data of interest, but whose joint distribution is heterogeneous and depends on some subset of a potentially large number of covariates $X = (X_1, X_2, ..., Xp) \in \mathbb{R}^p$. An estimate of the conditional distribution will give a certain target object $T(x)$:

$$P(Y|X = x) = P(Y|X_1 = x_1, ..., X_p = x_p), \tag{7}$$

where $x = (x_1, ..., x_p)$ is an arbitrary point in $\mathbb{R}^p$.

2.6. Deep Neutral Networks

The fifth methodology that will be used is the Deep Neutral Networks (DNN). Candel et al. [35] illustrate how a multi-layer feedforward artificial neural network trained by back-propagation stochastic gradient descent is the foundation for deep learning. Numerous hidden layers of neurons with maxout, rectifier, and tanh activation functions may be present in the network. High prediction accuracy is made possible by sophisticated features like adaptive learning rate, rate annealing, momentum training, dropout, L1 or L2 regularisation, checkpointing, and grid search.

Since the deep learning model of the neural network was compiled using h_20. The methodology below primarily focuses on the feedforward architecture used by h_20. This model uses a similar architecture to the RNN, which has a weighted combination of aggregated input signals given by Cevid et al. [29]:

$$\alpha = \sum_{i=1}^{n} w_i x_i + b, \tag{8}$$

having an output signal $f(\alpha)$ transmitted by the connected neuron. Unlike the RNN described in subsection 2.2, the Deep Neutral Network uses the LSTM extension. However, it uses the same activation functions illustrated in equations (2) to (4).

2.7. Stacked Ensemble

The seventh methodology used is the Stacked Ensemble for GBM, GLM, DRF and DNN (ESGGDD). According to LeDell [30], the Stacked Ensemble method is a supervised ensemble machine learning algorithm that uses a technique known as stacking to determine the best configuration of a set of prediction algorithms. This paper stacks GBM, GLM, DRF and DNN to produce a stacked GMM GLM DRF DNN (GGDD model). Another stacked model will combine three RNN models trained at different train test split ratios.

The stack ensemble method used here is from the $h_20 stackedensemblelearningmodel$ to find the optimal combination from several predictions. h_20 stacked ensemble algorithm is included in the h_20 system algorithms here (Stacked Ensembles [31]).

2.8. Model Forecast Accuracy

The best-performing models will be selected under the following criteria:

2.8.1. Root Mean Square Error Root mean squared error (RMSE) is the square root of the second sample moment of the differences between predicted and observed values or the quadratic mean of these differences, Letcher [32]. Its formula is given by:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_{th} - y_{th})^2}{n}}, \tag{9}$$

where $\hat{y}_{th}$ is the predicted value, y_{th} is the actual value n is the total number of observations.

2.8.2. Mean Square Error Mean squared error (MSE) measures residuals squared between the expected and actual observations, Douaik et al. [33]. Its formula is given by:

$$\text{MSE} = \frac{\sum_{i=1}^{n}(\hat{y}_{th} - y_{th})^2}{n}, \tag{10}$$

where $\hat{y}_{th}$ is the predicted value, y_{th} is the actual value and n is the total number of observations.

2.8.3. Mean Absolute Error Mean absolute error (MAE) measures absolute errors between the expected and actual observations, Schneider and Xhafa [34]. Its formula is given by:

$$\text{MAE} = \frac{\sum_{i=1}^{n} |\hat{y}_{th} - y_{th}|}{n}, \tag{11}$$

where $\hat{y}_{th}$ is the predicted value, y_{th} is the actual value and n is the total number of observations.

3. Empirical results and discussion

3.1. Exploratory data analysis

The dataset of daily observations of the closing price in Ethereum's United States Dollar (USD) currency has been obtained from a reliable open source called Coindesk. It is dated from the 1st of January 2022 to the 6th of December 2022.

The primary data used is quantitative. The closing price of Ethereum will be used to calculate lag1 and lag2. The tweets will be extracted using the search of #Ethereum on Twitter. Preprocessing the data includes:

1. Evaluating the lag1 and lag2 from the daily closing price will be used as two covariates.
2. Evaluating the no trend from the daily closing price will be used as one of the covariates.
3. Extracting a monthtype variable from each observation date will be used as one of the covariates.
4. Extracting a type variable from the date of each observation, which is to be used as one of the covariates.
5. Computing positive and negative sentiments from tweets using a naive Bayes classifier.
6. Hot-coding positive news as one and bad news as 0.

The response variable is the daily closing price of Ethereum, which undergoes scaling transformation to reduce the range between maximum and minimum closing price. The data will be split into three train and test splits: 80:20, 90:10 and 95:5. The variables in the dataset are described as follows:

- Ethereum Closing Price (ETH_CP): this is the daily closing price of Ethereum.
- Lag 1 (lag 1): this is the computed first lag of each closing price of Ethereum.
- Lag 2 (lag 2): this is the computed second lag of each closing price of Ethereum.
- Daytype (lag 2): this is the computed day from each date of each closing price of Ethereum.
- Monthtype (lag 2): this is the computed month from each date of each closing price of Ethereum.
- Noltrend 2 (lag 2): this is Ethereum's computed smooth spine of the closing price.
- Tbc (lag 2): this is the computed tweet polarity of each closing price of Ethereum.

Table 1 shows summary statistics of the overall data. It displays the minimum value (min), first quartile (Q1), mean, median, third quartile (Q3), maximum value (Max), Skewness and Kurtosis. There are 294 observations of each variable. The ETH_CP has a minimum of \$996.280 and a maximum of \$3786.640. The range between the minimum and maximum ETH_CP is more than \$2000. Hence, the scaling transformation has reduced the range, making it easier to compute the predictions. The ETH_CP has a skewness of 0.244, and the kurtosis of -1.388 reflects a platykurtic curve that has lighter tails than its normal distribution.

Figure 1 top panel displays ETH_CP, showing a better visual display of closing prices of Ethereum over the months. The quantile-quantile plot illustrated in Figure 1, bottom panel indicates that ETH_CP is not a typical normal distribution since it has heavy outliers in the left and right tails. Furthermore, an evaluation of the chi-square normality test at 5% alpha level of significance with the null hypothesis of ETH_CP is a normal distribution and alternative hypothesis that ETH_CP is not a normal distribution. The test produced a p-value less than α of 5%, indicating that ETH_CP does not display a normal distribution after rejecting the null hypothesis.

The box and whiskers plot in Figure 2 top panel represents the density distribution and the outliers. Figure 2 top panel displays a non-asymmetric distribution, which does not represent a normal distribution with many outliers on its tails. Figure 2 bottom panel evaluates the correlation of the elements in the dataset. Noltrend and monthtype

Table 1. Summary Statistics.

Variables	Min	Q1	Mean	Median	Q3	Max
ETH_CP	996.280	1436.165	2136.670	1924.010	2894.205	3786.640
$lag1$	-430.950	-55.713	-8.390	-2.535	46.880	307.550
$lag2$	-587.490	-78.993	-16.738	-4.535	63.288	358.350
$daytype$	1.000	7.000	13.660	13.500	20.000	30.000
$month$	1.000	3.000	6.143	6.000	9.000	12.000
$noltrend$	1061.278	1459.842	2136.698	1907.680	2887.622	3813.644
$tweets$	0.000	1.000	0.810	1.000	1.000	1.000

are the only covariates with a good correlation with ETH_CP, meaning the variables are highly related to the response variable; the rest have little correlation with ETH_CP.

A total of 788 tweets were extracted from Twitter from the 1st of January 2022 to the 06th of December 2022 using the following query on Twitter APIs: "$-nft - \#NFT - nfts - giveaways - \#giveaway - \#btc - \#bnb - \#bitcoin - followme(\#EthereumOR\#ETHUSD)min_replies : 1min_faves : 1lang : enuntil : 2022 - 06 - 16since : 2022 - 03 - 10 - filter : links$" This query assists in reducing Ethereum tweets noise as it makes sure that the tweets do not contain certain keywords that are not inline with Ethereum. The tweets were then subjected to data cleaning, which includes removing mentions, unwanted symbols, retweets and hyperlinks, as shown in Table 2. Twitter was only used as a sole source of social media because multiple scholars are beginning to see that Twitter can be used to anticipate a wide range of events, particularly financial markets. Another significant reason is that previous studies have demonstrated the ability to forecast market movement for securities and other financial instruments using real-time Twitter data.

Table 2. Overview of tweets.

Date	(1) 2022-01-01 05:55:44	(2) 2022-01-01 17:12:01	(3) 2022-01-01 19:18:45	(4) 2022-01-02 01:28:55	(5) 2022-01-02 13:19:52
$User$	LucidAxies	IMineBlock_com	realsheepship	itsmebutterz	CardanoHumpback
$Tweet$	Holy fucking shit xPolygon is such trash. I d...	Crypto mining has provided for me consistently...	Exchanges on Ethereum are decent with competit...	efiDrew dude's such a fucking idiot to not thi...	think it would be a dream because I am doing i...
$Subjectivity$	0.603571	0.284799	0.388333	0.511111	0.512500
$Polarity$	-0.085714	-0.014881	-0.029167	-0.136111	0.325000
$Analysis$	Negative	Negative	Negative	Negative	Positive

The following steps were taken to clean the tweets that were extracted from Twitter:

1. A tailored algorithm was used to remove and replace symbols and characters that might confuse the sentimental analysis.
2. This tailored algorithm removed '$@mentions'$, '$hashtagsymbols'$, '$\symbols$, '$@^{tm}symbols'$, '$eurosymbols'$, '$retweets'and'hyperlinks'$

Table 2 further shows that by using the Textblob function, subjectivity and polarity were computed to indicate whether the tweet is good or bad news. If the polarity of the tweet is less than one, then it is bad news. If it is greater than one, it is considered good news. Since more than one tweet can be good or bad, the tweets are further aggregated into one category daily. For example, if on the 1st of January, there are three tweets of bad news and one tweet of good news, the four tweets will be aggregated, and the 1st of January will be assigned 0 as $3 > 1$, meaning that the outcome of the 1st of January is negative.

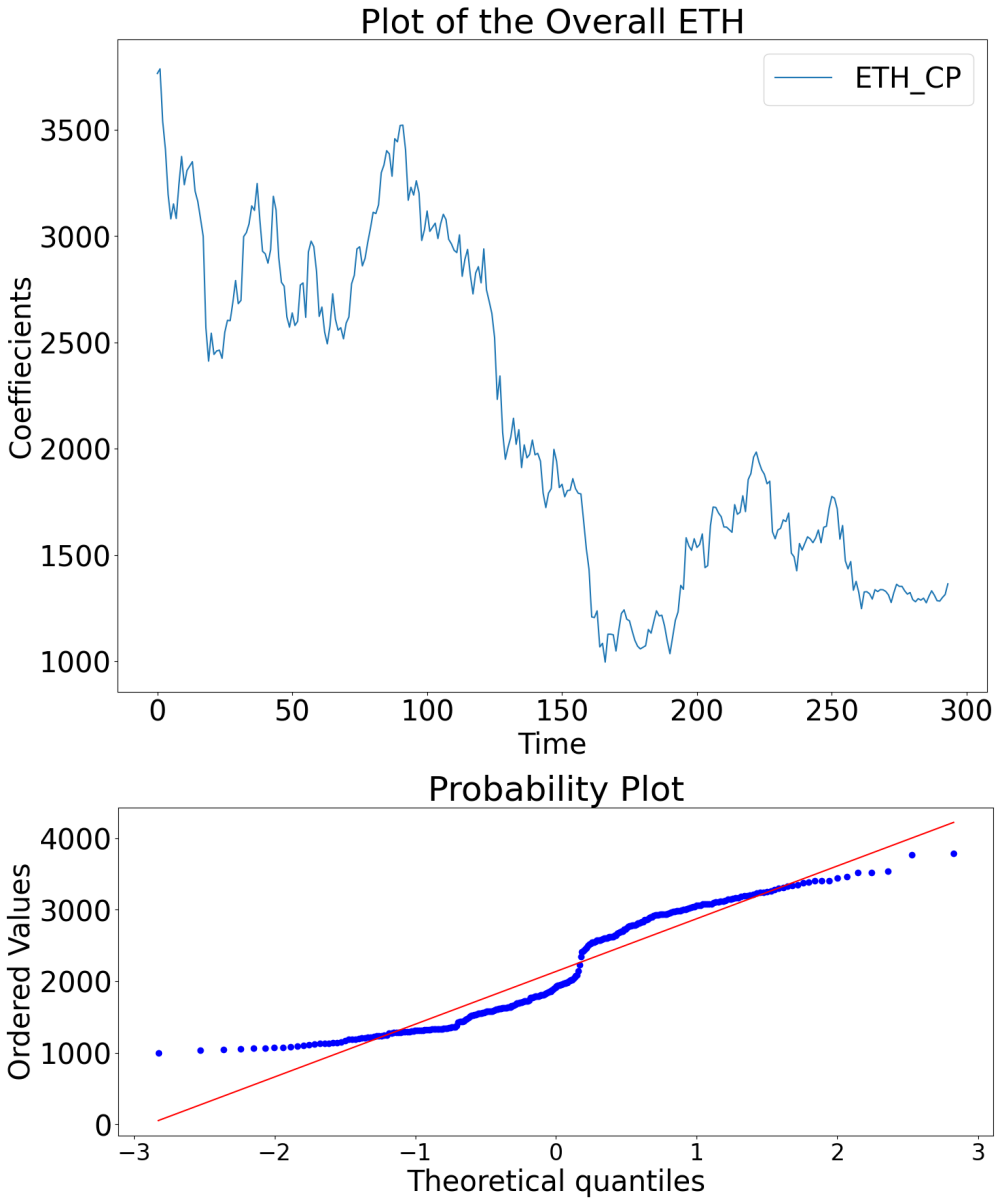

Figure 1. **Top panel:** Graphical representation of ETH_CP.**Bottom panel:** Quantile-quantile plot of ETH_CP.

3.2. Results

Table 3 summarises the accuracy measures of the seven comparative machine learning models evaluated at three train test ratios, i.e. 80:20, 90:10 and 95:5. All models had the same subset of training and testing data. The models were evaluated using three accuracy measures, i.e. MAE, MSE and RMSE. GLM is the model that had the lowest MAE of 21.31 when evaluated at a 95:5 train test ratio. An MAE of 21.31 means that on average days, the GLM predictions will be off the actual value by an error of $21.31. This means that after creating a lower and upper bound of $21.31 when forecasting the closing price of Ethereum, and it happens that it is sufficient when compared to the actual value, the investor can estimate their profitability when trading.

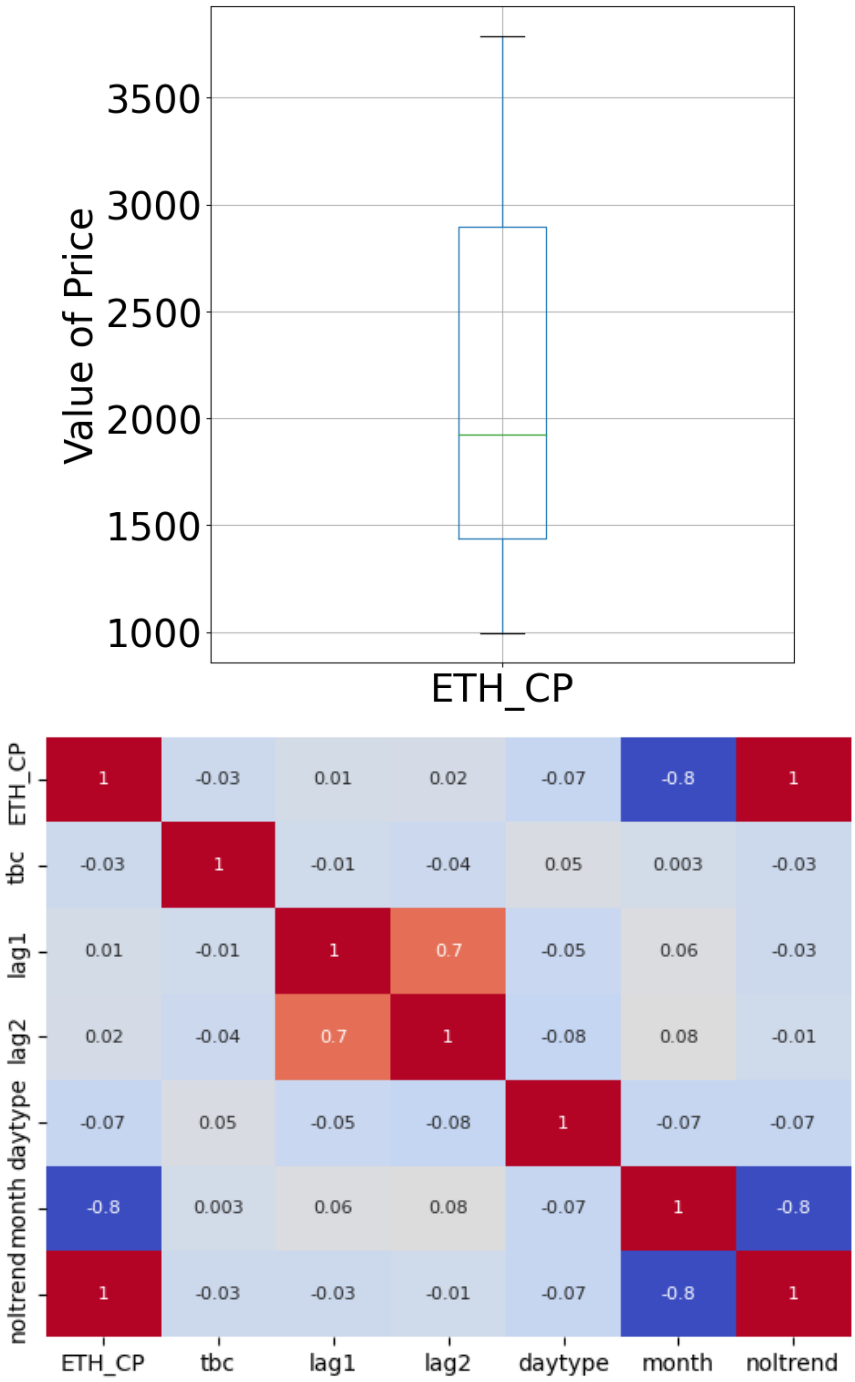

Figure 2. **Top panel:** Box and whisker of ETH_CP. **Bottom panel:** Correlation table.

Furthermore, GLM also had the lowest MSE and RMSE when evaluated at a 95:5 train test ratio. According to MAE, the second-best model that performs well after GLM is the RNN, which produced an MAE of 29.81. This hierarchy of performance was followed by Stacked GGDD, GBM, Stacked RNN and DRF, respectively, up

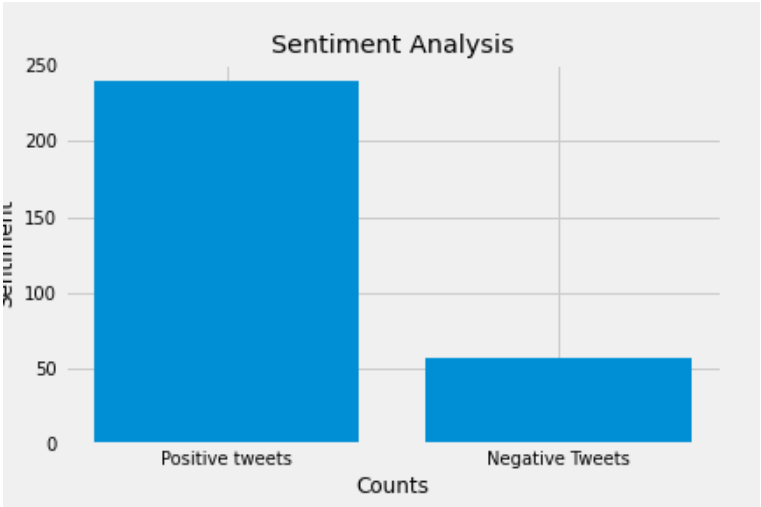

Figure 3. Partial Extracted tweets of #Ethereum.

Table 3. MAE, MSE and RMSE for all models.

Model	80:20 train test ratio			90:10 train test ratio			95:5 train test ratio		
	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
RNN	31.05	2286.86	47.82	42.52	3311.19	57.54	29.81	1914.67	43.76
Stacked RNN	29.04	1941.75	44.07	44.63	3918.78	62.61	56.36	6387.20	79.92
GBM	42.29	3708.81	60.90	40.66	3149.73	56.12	36.92	2039.63	45.16
GLM	29.80	1984.04	44.54	27.07	2369.08	48.67	21.31	1359.95	36.88
DRF	58.24	5508.92	74.22	67.75	8183.31	90.46	81.06	9439.04	97.15
DNN	64.23	6820.55	82.59	42.65	3515.87	59.29	81.28	8644.11	92.97
Stacked GGDD	44.38	4231.78	65.05	48.75	4765.92	69.04	34.67	1964.73	44.33

until we got to the least performing model according to MAE the DNN, which produced an MAE of 81.26 when evaluated at 95:5 train test ratio.

Figures 4 to 7 display the forecasted closing price of Ethereum against its actual closing price of all the models when evaluated at a 95:5 train test ratio. The red lines show the predicted values, whereas the black lines show the observed values. It is evident from 5 (b), which displays the GLM model, that the predicted and observed values are very close to each other, whereas, in 6 (b), which displays the DNN model, it is evident that the observed and actual values are not so close to each other as compared to 5 (b).

4. Discussion

The study focused on the complexities of predicting Ethereum's daily closing price, a cryptocurrency known for its volatile nature. With the widespread acceptance of cryptocurrencies as an alternative payment and currency exchange mode, the financial world has seen a surge in interest from academics, economists, investors, traders, and investment banks. However, accurately forecasting cryptocurrency prices remains a formidable challenge due to their rapid fluctuations.

The length of the training period for the machine learning algorithms has been kept constant to avoid bias in performance from the models. Each model has been trained for 2400 epochs. This training period was selected after multiple training periods were tried. During the selection of the best training length, the presence of outliers and external factors, such as the influence of specific market events that may skew the analysis of the accounted covariates, was considered and considered.

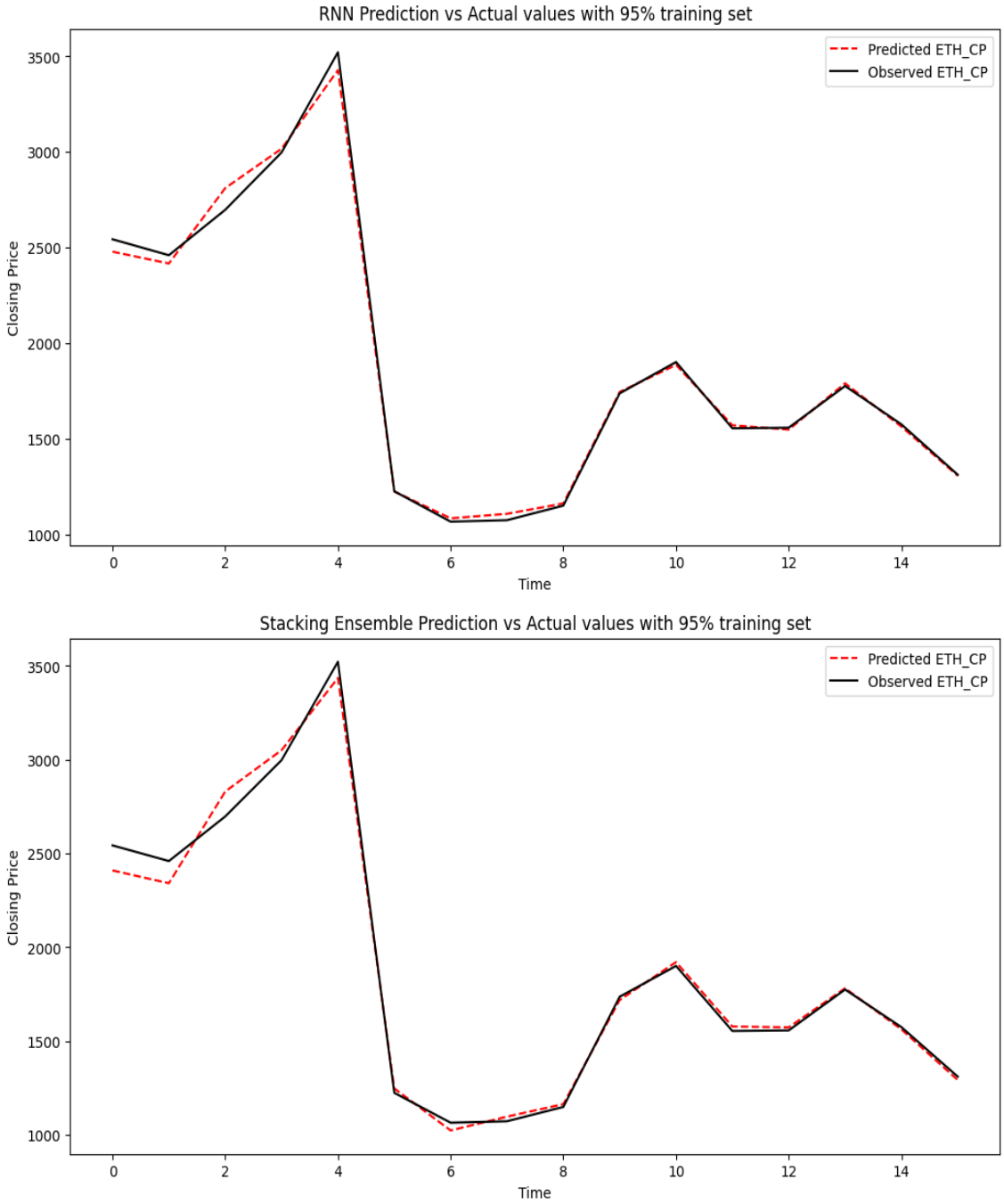

Figure 4. **Top panel:** RNN forecasts. **Bottom panel:** Stacked RNN forecasts.

The RNN was trained on different optimizers, and it was found that optimizer 'MSE' provided the best results; Different dropout layers of different sizes were tuned to provide the best improved model performance. The RNN had three layers; the best performance was obtained from the first layer, the second layer being a 'tanh', the third layer being a sigmoid, and the fourth layer being dense. A smooth and optimal fit spine was used to train GBM, GLM, DRF and DNN to ensure fair comparison and improved model performance.

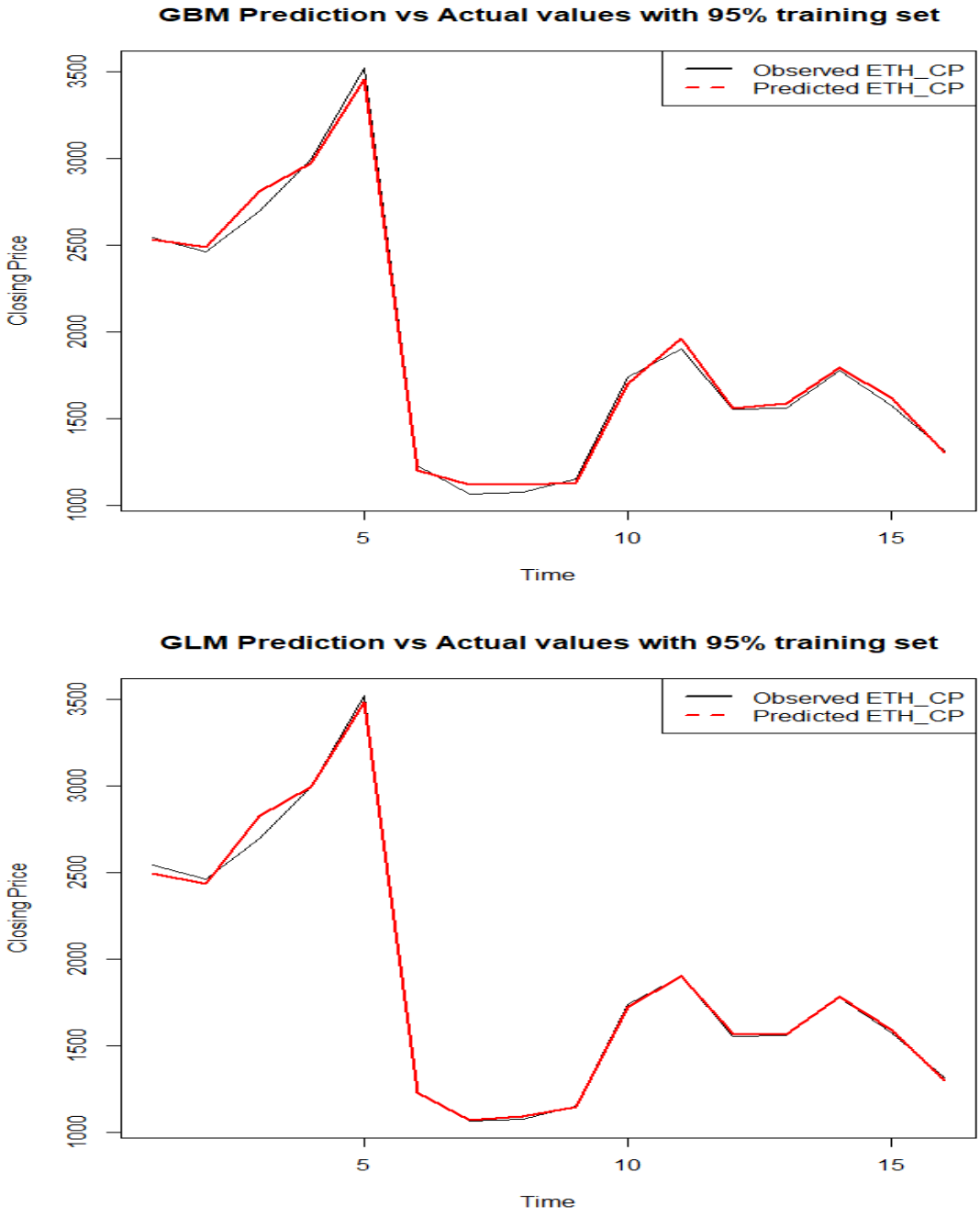

Figure 5. **Top panel:** GBM forecasts. **Bottom panel:** GLM forecasts.

The study uses deep learning algorithms within a stacking ensemble modeling framework to tackle this prediction task. This approach integrates various models to harness their collective predictive power, thus improving the accuracy of the forecasts. The dataset comprises daily historical observations of Ethereum prices sourced from Coindesk, tweets extracted from Twitter spanning from August 1, 2022, to August 8, 2022, and

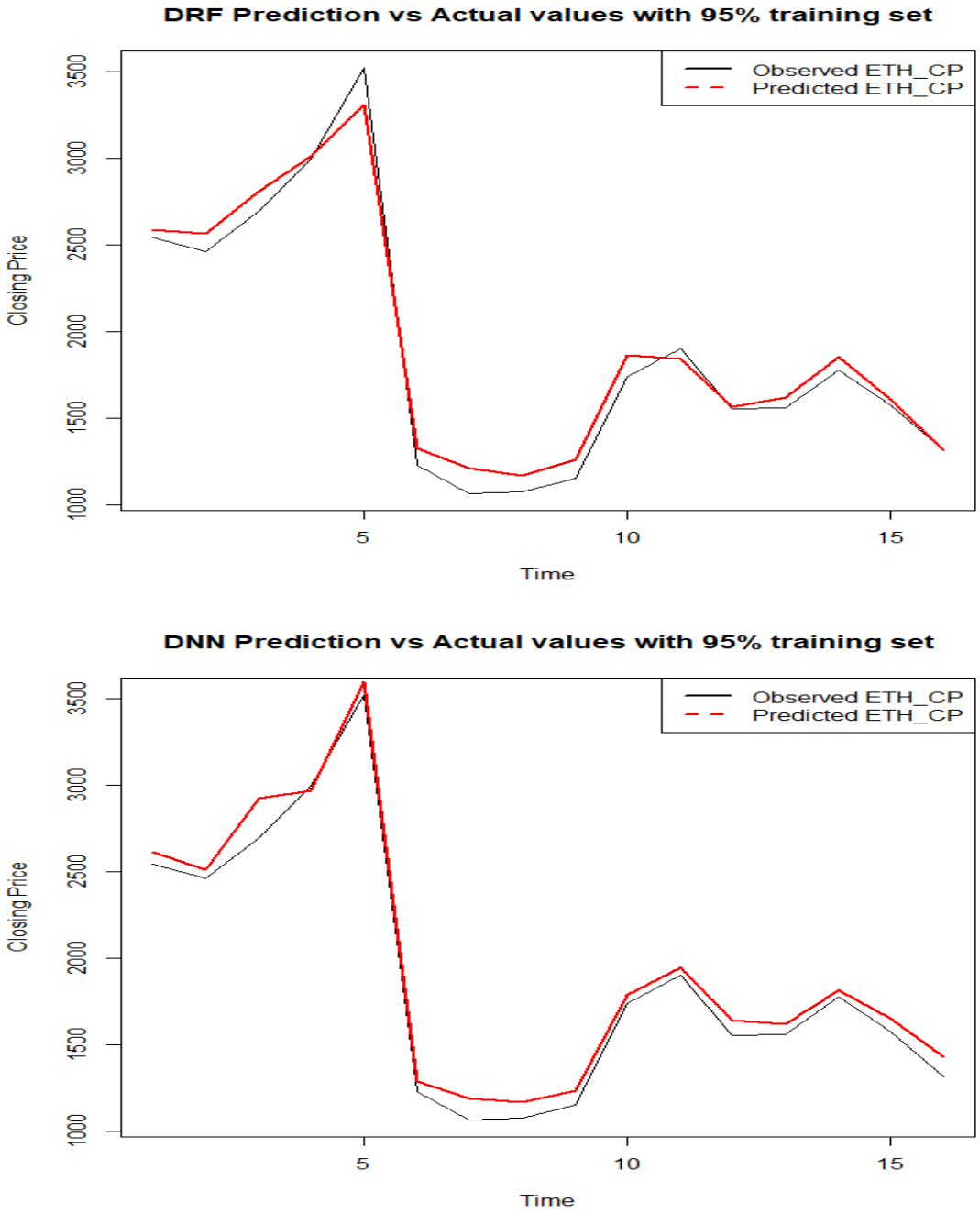

Figure 6. **Top panel:** DRF forecasts. **Botton panel:** DNN forecasts.

five additional covariates derived from Ethereum's closing price (closing price lag1, closing price lag2, noltrend, daytype, and month).

Seven models are employed to forecast Ethereum's daily closing price: Recurrent neural network, ensemble stacked recurrent neural network, Gradient boosting machine, Generalized linear model, Distributed random forest, Deep neural networks and Stacked ensemble (combining Gradient Boosting Machine, Generalized Linear

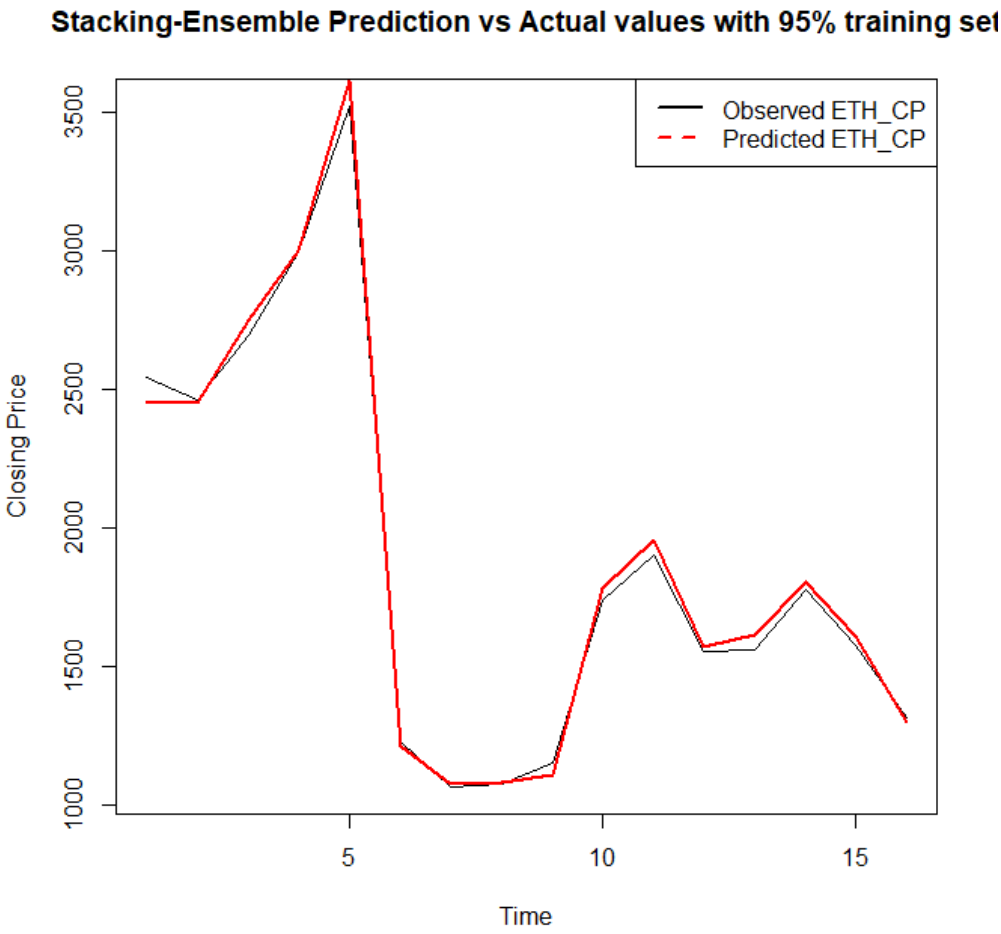

Figure 7. Stacked GGDD.

Model, Distributed Random Forest, and Deep Neural Networks). The main evaluation metric used to assess the performance of these models is MAE, which quantifies the average magnitude of errors in the predictions. The lower the MAE, the better the model's forecasting accuracy.

Empirical results from this study suggest that the RNN model is the best-performing model with an MAE of 21.31. This indicates that, on average, the RNN model's predictions deviate from the actual Ethereum closing prices by approximately 21.31 units. This result suggests that RNNs are particularly good at capturing Ethereum's price movements' complex patterns and dynamics, outperforming other models considered in the study. The best-performing model has been further supported by the lowest RMSE of 36.88 and lowest MSE of 1359.95. In contrast, the model was typically less robust when the RMSE values were larger.

5. Conclusion

The predictive capabilities of various algorithms, including RNN, Stacked RNN, GBM, GLM, DRF, DNN, and Stacked GGDD models, were examined using a stacking ensemble modelling framework to forecast the closing price of Ethereum. Among these, the GLM model, trained with a 95:5 train-test ratio, gave the most accurate

forecasts based on MAE. The MAE for GLM was 21.31, outperforming the RNN (29.81), Stacked RNN (56.36), GBM (36.92), DRF (81.06), DNN (81.28), and Stacked GGDD (34.67) models under the same 95:5 train-test ratio conditions. Therefore, according to MAE, GLM with a 95:5 train-test ratio yielded better results than the other models evaluated in this study. This suggests that when forecasting Ethereum's closing price, GLM is preferable to other models. Precise forecasts of the Ethereum closing price using covariates can help to assess profitability in Ethereum trading.

The study stresses the potential of deep learning techniques, particularly RNNs, in forecasting cryptocurrency prices. However, it is important to note that the cryptocurrency market is highly unpredictable and influenced by numerous factors beyond traditional financial data, including sentiment analysis from social media platforms such as Twitter. Thus, while these findings offer valuable insight, continued research and refinement of forecasting models are essential to effectively navigate the complexities of cryptocurrency trading and investment.

As useful as the tweets were significant when predicting the closing price of Ethereum, as noted from the high correlation between tweets and the closing price of Ethereum, for further improvement, when extracting tweets from X, formerly known as Twitter, one can use different windows/ lag on how impactful a tweet could be after a certain number of days. In addition to this, we also recommend making use of other use from another platform such as Google and crypto blogs. Given the range between the minimum and maximum closing price, normalizing the data before training helped models like GLM and RNN perform better. Other scaling techniques may produce better MAE results.

5.1. Limitations and Recommendations

In the cryptocurrency world, prices are very volatile. Hence, it is essential to encourage research that incorporates new techniques, strategies, and alternative approaches, such as more sophisticated prediction algorithms, feature engineering techniques, additional covariates, sentiment lexicons, topic modeling, or deep learning models specifically designed for NLP and other validation metrics for gaining accurate cryptocurrency price prediction. Exploring the content and patterns of the tweets could provide a better understanding of their impact on Ethereum prices.

This can assist cryptocurrency investors toward potentially increased profits and support policy market behavior. Market-specific events or news on cryptocurrency prices could affect generalized findings; researchers should avoid this limitation in the future. The discussion here could be beneficial for exploring some promising opportunities that remain open in cryptocurrency price prediction research. The consideration of multiple social media platforms could provide a more robust analysis and bring forth new findings.

Acknowledgements:

This research was funded by the DST-CSIR National e-Science Postgraduate Teaching and Training Platform (NEPTTP) http://www.escience.ac.za/.

Data Availability Statement:

Daily historical observations of the price of Ethereum were obtained from Coindesk, a free access website https://www.coindesk.com/price/ethereum/ and tweets extracted from Twitter.

Acknowledgements:

The support of the DST-CSIR National e-Science Postgraduate Teaching and Training Platform (NEPTTP) towards this research is hereby acknowledged. Opinions expressed, and conclusions arrived at are those of the authors and

are not necessarily to be attributed to the NEPTTP. In addition, the authors thank the anonymous reviewers for their helpful comments on this paper.

Conflicts of Interest:

The authors declare no conflict of interest. The funders had no role in the study's design, in the collection, analyses, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

ANN	Artificial neural network
BTC	Bitcoin
DNN	Deep neural network
DRF	Distributed random forest
ETH	Ethereum
GBM	Gradient boosting machines
GLM	General linear model
LSTM	Long short term memory
MAE	Mean absolute error
MAPE	Mean absolute percentage error
MSE	Mean square error
NLP	Natural language processing
NLU	Natural language understanding
RF	Random forest
RMSE	Root mean square error
RNN	Recurrent neural network
SGBM	Stochastic gradient boosting machines
SVM	Support vector machines

REFERENCES

1. Middlebrook, S. T. (2014), *Bitcoin for merchants: Legal considerations for businesses wishing to accept bitcoin as a form of payment*. Business Law Today, p. 1. http://dx.doi.org/
2. Mukhopadhyay, U., Skjellum, A., Hambolu, O., Oakley, J., Yu, L. and Brooks, R. (2016). *A brief survey of cryptocurrency systems, in 2016 14th annual conference on privacy, security and trust (PST), Auckland, New Zealand*. IEEE, pp. 745–752. https://doi.org/10.1109/PST.2016.7906988
3. Warner, J. (2018), *The founder of ethereum: Vitalik buterin*. IG. https://www.ig.com/en/news-and-trade-ideas/forex-news/the-founderof-ethereumvitalik-buterin-41892-180131
4. Times, T. E. (2021). *Cryptocurrency Ethereum is flourishing, but risks linger*. https://economictimes.indiatimes.com/markets/forex/cryptocurrencyethereum-is-flourishing-but-risks-linger/articleshow/82475252.cms
5. Shavlik, J.W., Dietterich, T. and Dietterich, T. G. (1990). *Readings in machine learning*. Morgan Kaufmann, California, United States of America.
6. Anderson, J. R. (1990). *Machine learning: An artificial intelligence approach*. Vol. 3, Morgan Kaufmann.
7. Liddy, E.D. (2001). *Natural Language Processing*. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc.
8. Tilimbe, J. (2019). *Ethical implications of predictive risk intelligence*. The ORBIT Journal, 2(2), 1–28. https://doi.org/10.29297/orbit.v2i2.112
9. Fleming, J., Ostdiek, B. and Whaley, R. E. (1995). *Predicting stock market volatility: A new measure*. Journal of Futures Markets, 15(3), 265–302. https://doi.org/10.1002/fut.3990150303
10. Jain, P. and Jiang, C. (2014). *Predicting future price volatility: empirical evidence from an emerging limit order market*. Pacific-Basin Finance Journal, 27, 72–93. https://doi.org/10.1016/j.pacfin.2014.01.006

11. Tang, Y., Xiao, X., Wahab, M. andMa, F. (2021). *The role of oil futures intraday information on predicting us stock market volatility*. Journal of Management Science and Engineering, 6(1), 64–74. https://doi.org/10.1016/j.jmse.2020.10.004
12. Lahmiri, S. (2017). *Modeling and predicting historical volatility in exchange rate markets*. Physica A: statistical mechanics and its Applications, 471, 387–395. https://doi.org/10.1016/j.physa.2016.12.061
13. Kim, H. Y. and Won, C. H. (2018). *Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple garch-type models*. Expert Systems with Applications, 103, 25–37. https://doi.org/10.1016/j.eswa.2018.03.002
14. Anbazhagan, S. and Kumarappan, N. (2012). *Day-ahead deregulated electricity market price forecasting using recurrent neural network*. IEEE Systems Journal, 7(4), 866–872. https://doi.org/10.1109/JSYST.2012.2225733
15. Derbentsev, V., Babenko, V., Khrustalev, K.I.R.I.L.L., Obruch, H. and Khrustalova, S.O.F.I.I.A. (2021). *Comparative performance of machine learning ensemble algorithms for forecasting cryptocurrency prices*. International Journal of Engineering, 34(1), 140–148. https://doi.org/10.5829/ije.2021.34.01a.16
16. Poongodi, M., Sharma, A., Vijayakumar, V., Bhardwaj, V., Sharma, A. P., Iqbal, R., and Kumar, R. (2020). *Prediction of the price of Ethereum blockchain cryptocurrency in an industrial finance system*. Computers and Electrical Engineering, 81, 106527. https://doi.org/10.1016/j.compeleceng.2019.106527
17. Livieris, I. E., Pintelas, E., Stavroyiannis, S., and Pintelas, P. (2020). *Ensemble deep learning models for forecasting cryptocurrency time series*. Algorithms, 13(5), 121. https://doi.org/10.3390/a13050121
18. Henrique, B. M., Sobreiro, V. A. and Kimura, H. (2018). *Stock price prediction using support vector regression on daily and up to the minute prices*. The Journal of Finance and Data Science, 4(3), 183–201. https://doi.org/10.1016/j.jfds.2018.04.003
19. Vadivukarassi, M., Puviarasan, N. and Aruna, P. (2017). *Sentimental analysis of tweets using naive Bayes algorithm*. World Applied Sciences Journal, 35(1), 54–59. https://idosi.org/wasj/wasj35(1)17/7.pdf
20. Luo, X., Zimet, G. and Shah, S. (2019). *A natural language processing framework to analyze the opinions on HPV vaccination resected in Twitter over 10 years (2008-2017)*. Human Vaccines and Immunotherapeutics, 15(7-8), 1496. https://doi.org/10.1080/21645515.2019.1627821
21. Giudici, P., Polinesi, G. (2021). *Crypto price discovery through correlation networks*. Annals of Operations Research, 299(1), 443–457. https://doi.org/10.1007/s10479-019-03282-3
22. Khedr, A. M., Arif, I., El-Bannany, M., Alhashmi, S. M., and Sreedharan, M. (2021). *Cryptocurrency price prediction using traditional statistical and machine-learning techniques: A survey*. Intelligent Systems in Accounting, Finance and Management, 28(1), 3–34. https://doi.org/10.1002/isaf.1488
23. Rish, I. et al. (2001). *An empirical study of the naive Bayes classifier, in 'IJCAI 2001 workshop on empirical methods in artificial intelligence*, 3, 41–46.
24. Williams, R. J. and Zipser, D. (1989). *A learning algorithm for continually running fully recurrent neural networks*. Neural Computation, 1(2), 270–280. https://doi.org/10.1162/neco.1989.1.2.270
25. Friedman, J. H. (2001). *Greedy function approximation: a gradient boosting machine*. Annals of Statistics, 1189–1232. https://doi.org/10.1214/aos/1013203451
26. Natekin, A., and Knoll, A. (2013). *Gradient boosting machines, a tutorial*. Frontiers in neurorobotics, 7, 21. https://doi.org/10.3389/fnbot.2013.00021
27. Lee, Y., and Nelder, J. A. (1996). *Hierarchical generalized linear models*. Journal of the Royal Statistical Society Series B: Statistical Methodology, 58(4), 619-656. https://doi.org/10.1111/j.2517-6161.1996.tb02105.x
28. Geurts, P., Ernst, D., and Wehenkel, L. (2006). *Extremely randomized trees*. Machine learning, 63, 3–42. https://doi.org/10.1007/s10994-006-6226-1
29. Cevid, D., Michel, L., Näf, J., Bühlmann, P., and Meinshausen, N. (2022). *Distributional random forests: Heterogeneity adjustment and multivariate distributional regression*. Journal of Machine Learning Research, 23(333), 1-79. https://jmlr.org/papers/volume23/21-0585/21-0585.pdf
30. LeDell, E. E. (2015). *Scalable ensemble learning and computationally efficient variance estimation*. University of California, Berkeley.
31. Stacked Ensembles. Available online: https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/stacked-ensembles.html?highlight=stacked(accessedon11February2024).
32. Letcher, T. (2022), *Comprehensive renewable energy*. Second Edition, Elsevier.
33. Douaik, A., VanMeirvenne, M. and Tóth, T. (2005). *Soil salinity mapping using spatio-temporal kriging and Bayesian maximum entropy with interval soft data*. Geoderma, 128(3-4), 234–248. https://doi.org/10.1016/j.geoderma.2005.04.006
34. Schneider, P. and Xhafa, F. (2022). *Anomaly Detection and Complex Event Processing Over IoT Data Streams: With Application to EHealth and Patient Data Monitoring*, Elsevier Science, https://books.google.co.za/books?id=gLICzgEACAAJ
35. Candel, A., Parmar, V., LeDell, E., and Arora, A. (2016). *Deep learning with H2O*. H2O. ai Inc, 1-21. https://h2o.ai/content/dam/h2o/images/marketing/unassigned/DeepLearningBooklet.pdf