The utilization of four models for the prediction of ETH's realized fluctuations

Meiyu Song, Yueqi Yu, Yile Zhang, Zizhen Wang

Mathematics and Physics, Xi'an Jiaotong-Liverpool University, Suzhou, China

*Corresponding author: Meiyu.Song23@student.xjtlu.edu.cn

Keywords: ETH, GARCH, neural network, forecasting

Abstract: Model predictions are very important for the cryptocurrency market, and accurate model predictions are important for the application of blockchain in the market and to improve the socio-economic benefits. In order to predict the realized fluctuation of cryptocurrency more accurately, based on the four models of BASELINE MODEL, RANDOM WALK, GARCH, NEURAL NETWORK, combining the current situation at home and abroad and the advantages in big data processing, by comparing the metrics of RMSE and RMSPE, we conclude that the optimal model for the realized fluctuation is random walk.[1]

1. Introduction

Our world is evolving due to digitization, which is bringing about cutting-edge financial channels and new technologies like cryptocurrencies—which are just blockchain-based financial applications. Our group studied the development history of cryptocurrencies and used four models to study the volatility of digital currencies. A growing body of research has examined how the volatility characteristics of cryptocurrencies and other financial assets are similar as a result of the growing popularity of digital currencies (see, for example, Baur et al., 2018; Dyhrberg, 2016a; Bouri et al., 2017a; and Klein et al., 2018). In the background of the epidemic, the question of whether cryptocurrency should be viewed as an asset or a currency is hotly debated in academia (Yuneline, 2019; White et al., 2020, among others) because of the significant price increase and volatility that has been seen since 2017. For instance, before February 2017, the price of Bitcoin was less than $1000. It rose over $20,000 in December 2017, but it fell to about $8000 in February 2018. In May 2018, it increased once again, reaching $13,000, then in December 2018, it dropped sharply to about $3,000. The average price of Bitcoin in 2019 was approximately $7000.The exponential increase in speculative activity weakens the effectiveness of pertinent portfolio diversification measures and increases the volatility of cryptocurrency markets (Katsiampa, 2017).We needs to use technological means to achieve predictability in response to the volatility of cryptocurrencies. So far,a few researches have examined predictability in cryptocurrencies. For example, Catania et al. (2018a) investigated cryptocurrency, Hotz-Behofsits et al. (2018) used a time-varying parameter VAR with t-distributed measurement errors and stochastic volatility to describe cryptocurrencies. Prediction utilizes a number of different multivariate and univariate models. The question of whether Tether, another cryptocurrency backed by the USD, is directly influencing the price of Bitcoin to make it more predictable was examined by Griffin and Shams (2020); also see Gandal et al. (2018). In 2019, Trucíos compared the value at risk (VaR) and one-step-ahead volatility predictability of Bitcoin using multiple volatility models. His findings support earlier research by Charles and Darné (2019) and Catania et al. (2018b), which highlighted the importance of treating extreme observations carefully when analyzing bitcoin returns. In particular, he discovered that a score-driven model that incorporates Student's t distributed innovations and time-varying volatility performs better than a lot of GARCH-type models. According to Chu et al. (2015)'s statistical analysis of log returns of the BTC vs USD exchange rate, the generalized hyperbolic distribution appears to be the best option for modeling the unconditional distribution of cryptocurrency time series. According to Núñez et al. (2019), the returns of Bitcoin can be fitted by the normal inverse Gaussian distribution (NIG).[2]

  DOI: 10.25236/iceesr.2024.045

2. Model

2.1 GARCH model

The GARCH (Generalized Autoregressive Conditional Heteroskedasticity) model is a statistical and prediction model. The model volatility or alteration heteroskedasticity in time from some data. It is often employed in the financial sector to model the volatility of financial assets.

The general form of the GARCH (p, q) model is as follows:

Calculation of Conditional Variance: $\sigma^2(t) = \omega + \sum_{i=1}^{p} \alpha_i \cdot \varepsilon^2(t-i) + \sum_{j=1}^{q} \beta_j \cdot \sigma^2(t-j)$

Where,

σ^2(t) is the conditional variance (volatility) at time

ω (omega) is the continual term in the GARCH model, standing for the baseline level of conditional difference.

p is the order of the GARCH effect (ARCH part), standing for making a difference on squared errors from the past p time steps on the current conditional variance.

$\alpha_i$are the coefficients of the ARCH effect, representing the weights of squared errors from the past p time steps on the current conditional variance.

$\varepsilon(t-i)$ is the error term at time $t-i$.

q is the order of the GARCH effect (GARCH part), representing making a difference past q conditional variance on the current conditional variance.

β_j are the coefficients of the GARCH effect, representing the weights of past q conditional variances on the current conditional variance.

Meaning of Conditional Variance: The conditional variance denotes an estimate of the future date and fluctuation of observations given a set of known foregone observations and conditional variances.

The main use of the GARCH model is to model and forecast the volatility of financial markets. This is also the basis model for our currency forecast. By estimating the model's parameters, one can obtain an estimation of future market volatility.

2.2 LSTM model

(Long Short-Term Memory) is irregular of recurrent neural networks (RNNs), especially good fit for processing sequential data, as it can arrest long-term needs. Here are the basic formulas for LSTM:

LSTM calculations can be partitioned into the following steps:

(1) Input Gate:

Switches which date should add to the cell state.

Formula: $i_t = \sigma(W_{ii} \cdot x_t + b_{ii} + W_{hi} \cdot h_{t-1} + b_{hi})$

Where,

i_t is the input gate activation at time step t.

σ is the sigmoid activation function.

W_{ii} and b_{ii} are weights and biases associated with the input gate.

W_{hi} and b_{hi} are the weights and biases for the input gate's hidden state.

(2) Forget Gate:

Controls which state should be redundant from the cell state.

Formula: $f_t = \sigma(W_{if} \cdot x_t + b_{if} + W_{hf} \cdot h_{t-1} + b_{hg}$

Where f_t is the forget gate output, σ is the sigmoidal activation function, $W_{hf}$and $b_{hg}$are weights and biases about the forget gate.

(3) Cell State Update:

Apprises the cell state based on the input and forget gates.

Formula: $g_t = \tanh(W_{ig} \cdot x_t + b_{ig} + W_{hg} \cdot h_{t-1} + b_{hg})$

Where g_t is the candidate cell state at time step t. tanh is the hyperbolic tangent activation function.

(4) Output Gate:

Controls data should be passed to the hidden state and used as the network output.

Formula: $o_t = \sigma(W_{io} \cdot x_t + b_{io} + W_{ho} \cdot h_{t-1} + b_{ho})$

Where o_t is the output gate output, σ is the sigmoid activation function.
(5) Hidden State Update:
Revises the hidden state grounded on the output gate
Formula: $h_t = o_t \cdot \tanh(c_t)$
Where h_t is the new hidden state, and $\tanh(c_t)$ is the hyperbolic tangent activation function.
LSTM can effectively to process the date to make it become a powerful tool for tasks.

3. Experimental analysis

In this section, we first consider using financial data provided by Yahoo Finance. This data records the historical price data of a specified asset with the ticker symbol "ETH USD" (Ethereum cryptocurrency) acquired by the yfinance library between January 1, 2018 and the present, and then plots the closing price, to select the most appropriate model for predicting Ethereum's closing price volatility by comparing the gap between the predicted value and the actual value of the four models.

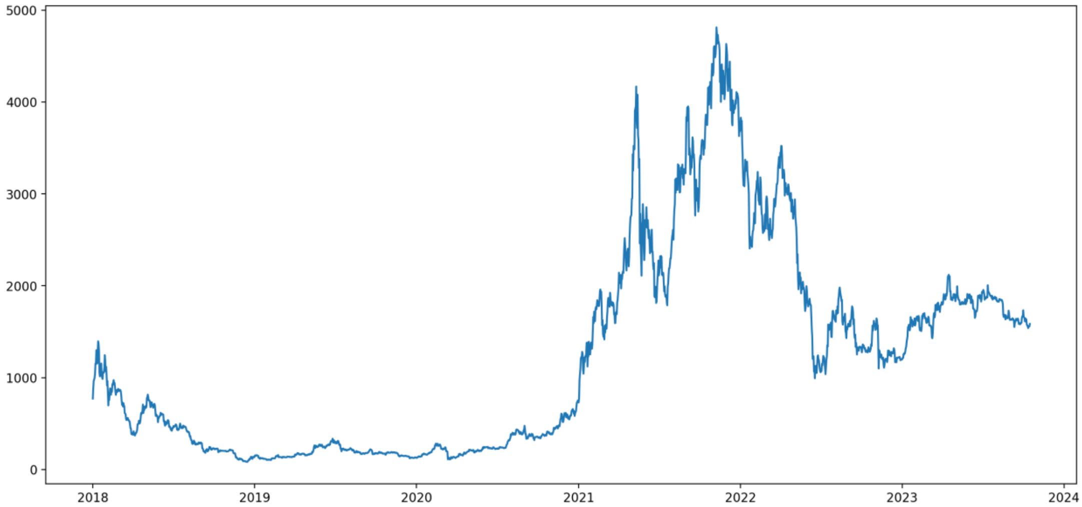
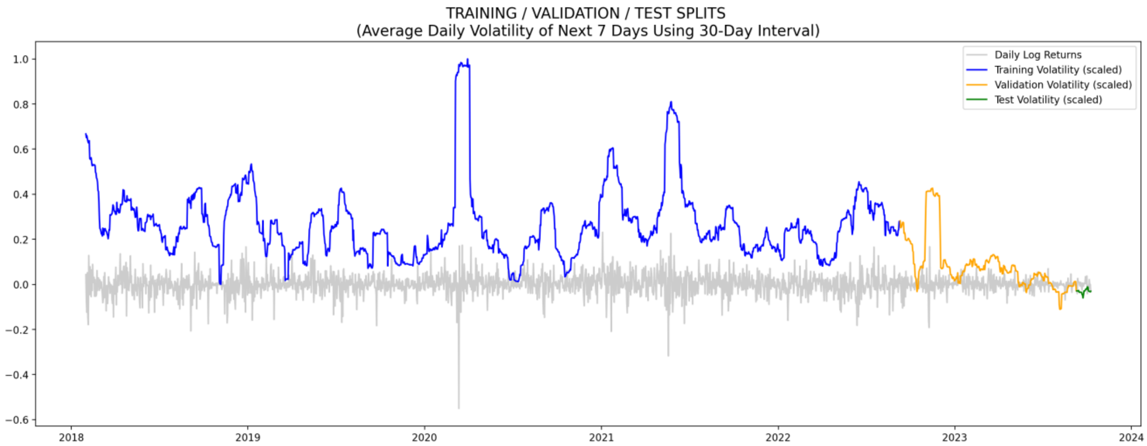


Figure 1 Closing price chart

Figure 1 is a time series graph. We use Log-return for visualization and analysis in this study because log-returns have smaller data intervals, are easy to graph, are easy to observe and compare, and can better handle price fluctuations and percentage changes. The overall volatility of the chart is high, especially in the overall upward trend from 2021 to 2023, indicating that the stock price has increased.



(Average Daily Volatility of Next 7 Days Using 30-Day Interval)

Figure 2 Training/Validation/Test Splits

The code in this section scales the volatility of the training, validation, and test sets so that they fall into the same range, ensuring that the model is not affected by different scales when working with this data. Figure 2 shows the normalized daily implementation volatility for the training set, validation set, and test set.

Study duration:

January 31, 2018, to September 9, 2022, totaling 1683 days as training set;A total of 365 days from September 10, 2022, to September 9, 2023, as validation set;A total of 30 days from September 10, 2023, to October 9, 2023, as a test set.

Daily Log Returns (gray curve): indicates daily log returns. This is the raw data used to calculate the implementation volatility.

Training Volatility (Blue curve): This is the normalized realized volatility of the training set. Your model will use this data for training.

Validation Volatility (Orange curve): This is the normalized realized volatility of the validation set. It is used for validation and adjustment during model training.

Test Volatility (Green curve): This is the normalized realized volatility of the test set. It is used to evaluate the performance of the model on future data.[3]

3.1 Mean Baseline Model

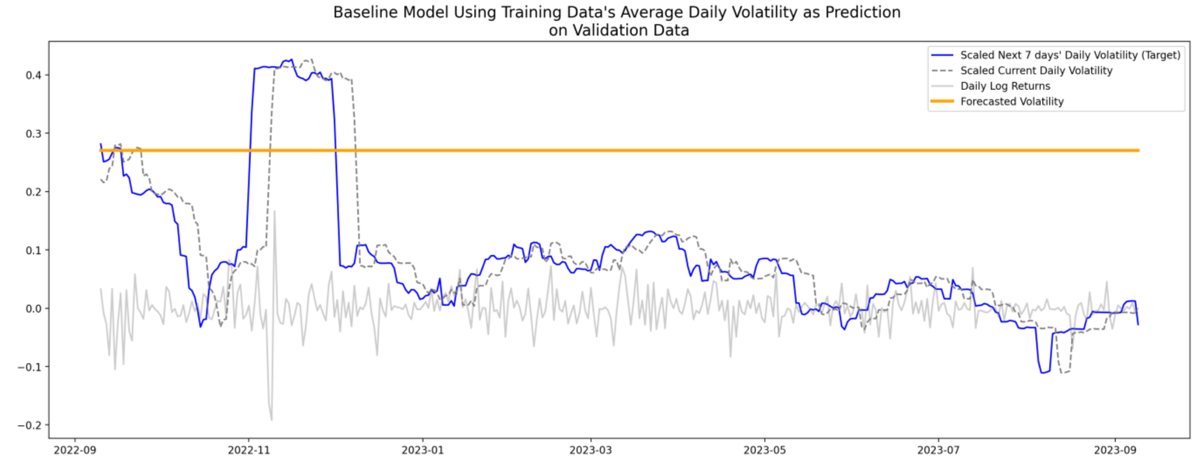


Figure 3 mean baseline Model

The average daily volatility of the training data is used as the baseline model for forecasting in Figure 3 using data from September 2022 through September 2023, compared to the daily volatility. The average daily volatility is a straight yellow line, a far cry from the blue graph of daily volatility.

The RMSPE of the mean baseline model is 89.277882 and the RMSE is 0.217141. Both values are large, so the predictive ability of the mean baseline model is poor.

3.2 Naive Random Walk Forecasting

For volatility Forecasting, Naive Random Walk Forecasting can be described as using the average of the first n realized volatility as a forecast for the next n time steps. In this article, the average realized volatility for the previous 7 days is used as a forecast of volatility for the next 7 days.

Figure 4 uses the daily volatility of the last 7 days as the forecast of the next 7 days. The coincidence in the chart is similar, and the volatility of the forecast model is more accurate.

Through the RMSPE of Random Walk model on the verification set, it can be seen that the smaller the value of RMSPE, the more accurate the prediction of the model, because it means the smaller the error of the prediction of the model relative to the true value.

The RMSE of the Random Walk model on the validation set is 0.06895. This value represents the root-mean-square error between the model's prediction and the actual value. Here, a lower RMSE indicates that the model performs better on the validation set.[4]

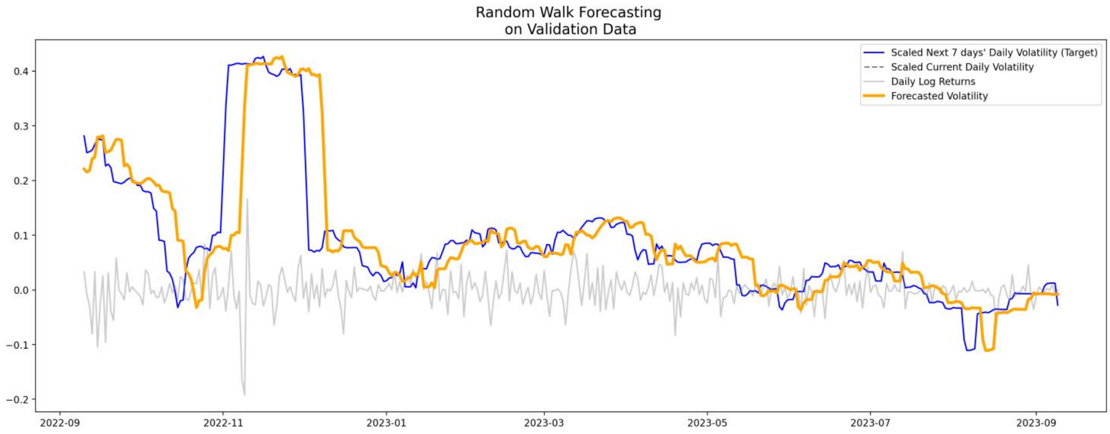


Figure 4 Random Walk Forecasting

3.3 GARCH Model

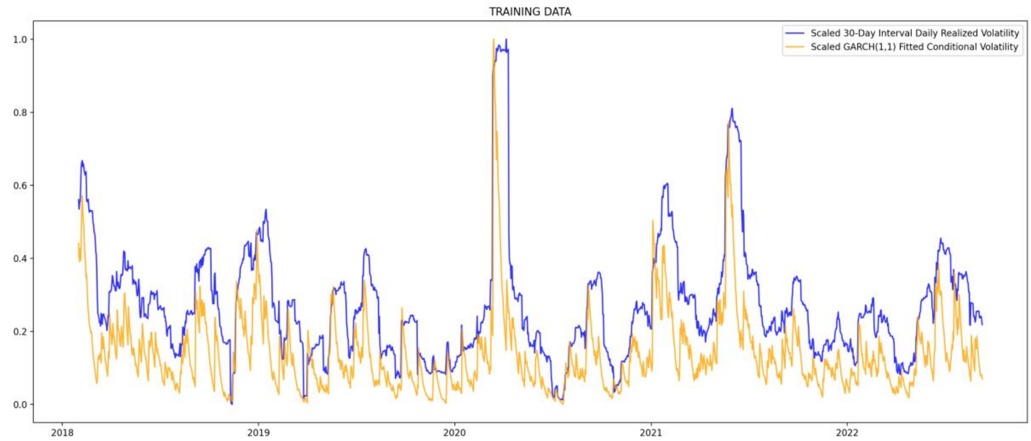


Figure 5 Training data

After harmonizing the training data (percent returns), the GARCH model outputs the conditional volatility of the training portion of the time series. We scale the array of harmonized conditional volatilities, plot them, and compare them to the realized volatilities (also scaled) calculated above in Figure 5.

The blue line is the scaled 30-day interval daily realized volatility, and the yellow line is the scaled GARCH (1, 1) fitted conditional volatility, with essentially the same volatility and a smaller difference.

GARCH models better overlay the test dataset in terms of volatility and size, with better predictive power of the models. [5]

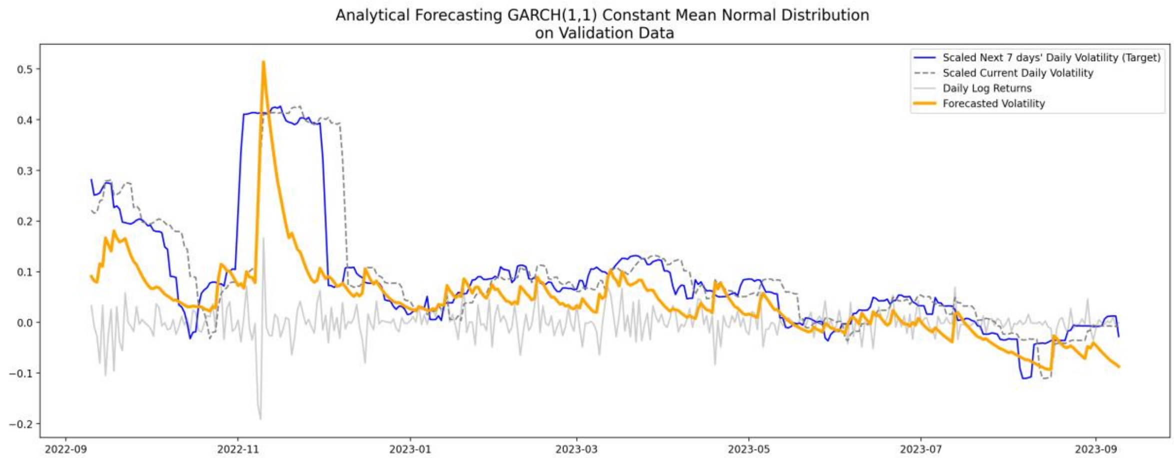


Figure 6: Analytical-based forecasting

The blue line in Figure 6is the future seven-day volatility and the yellow line is the predicted volatility. The predicted value is closer to the volatility trend of the actual target value, which also reflects the better accuracy and performance of the GARCH (1,1) model.

The RMSPE of the GARCH model is 0.745761 and the RMSE is 0.133625, both of which are relatively small values, giving the model a better predictive ability compared to the mean-baseline.

3.4 Neural Networks

The fully connected neural network you use is a simple neural network structure that is suitable for basic time series prediction tasks.

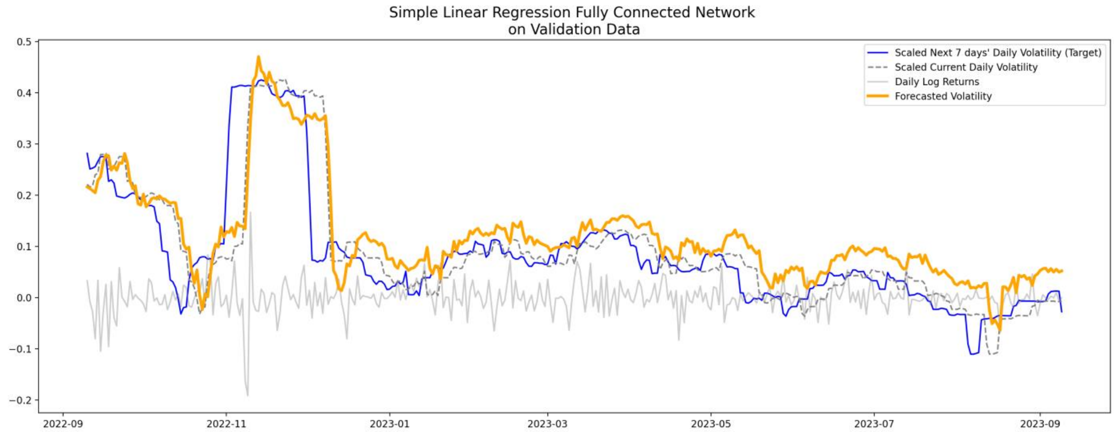


Figure 7 Simple Linear Regression Fully Connected Network

Figure 7 is used to visualize the prediction results of the fully connected neural network model on the validation set compared to the target values.

The graph contains two curves:

The blue curve represents the target value of the validation set, which is the actual normalized (scaled) future volatility.

The orange curve represents the predicted value of the model on the validation set, that is, the model's estimate of future volatility.

By observing the trend and comparison of these two curves, we can initially evaluate the performance of the model. The orange curve can closely follow the blue curve, so the prediction effect of the model is better.

4. Conclusions

Table 1 Comparison of forecasting models

	Model	Validation RMSPE	Validation RMSE
0	Mean Baseline	89.277882	0.217141
1	Random Walk	4.015127	0.068950
2	GARCH(1,1) | Constant Mean | Normal Dist	4.959604	0.084466
3	Simple LR Fully Connected NN | n_past=14	16.148377	0.076383

In Table 1, we observe that through the above several models, we study the volatility of the Ethereum digital coin. RMSPE measures the root mean square of percentage error, so it provides a way to measure prediction error. In RMSPE, the random walk model predicts the most accurate volatility and provides the best performance. RMSE is a measure of a model's prediction error that calculates the square root of the mean squared variance between the model's predicted value and the true value. In this context, the smaller the value of RMSE, the better, indicating the smaller the prediction error of the model. In RMSE, it is still the random walk model that predicts the volatility most accurately. So in summary, random walk models are very good at predicting cryptocurrency volatility. [6]

References

[1] Ftiti, Z., Louhichi, W., & Ben Ameur, H. (2021). Cryptocurrency volatility forecasting: What can we learn from the first wave of the COVID-19 outbreak? Annals of Operations Research, 1-26.

[2] Adcock, R., & Gradojevic, N. (2019). Non-fundamental, non-parametric Bitcoin forecasting. Physica A: Statistical Mechanics and its Applications, 531, 121727.

[3] Catania, L., & Grassi, S. (2022). Forecasting cryptocurrency volatility. International Journal of Forecasting, 38(3), 878-894.

[4] Conrad, C., Custovic, A., & Ghysels, E. (2018). Long-and short-term cryptocurrency volatility components: A GARCH-MIDAS analysis. Journal of Risk and Financial Management, 11(2), 23. Translated with www.DeepL.com/Translator (free version)

[5] Cheikh, N. B., Zaied, Y. B., & Chevallier, J. (2020). Asymmetric volatility in cryptocurrency markets: New evidence from smooth transition GARCH models. Finance Research Letters, 35, 101293.

[6] Palamalai, S., Kumar, K. K., & Maity, B. (2021). Testing the random walk hypothesis for leading cryptocurrencies. Borsa Istanbul Review, 21(3), 256-268.