Cryptocurrency Price Prediction Using Various Machine Learning models including Ensemble Method and Sentiment Analysis

Sujay S Chakravarthy
Department of Computer Science
PES University, RR Campus
Bengaluru, India
sujas0610@gmail.com

Suhas V
Department of Computer Science
PES University, RR Campus
Bengaluru, India
suhasv2912@gmail.com

Suraj R
Department of Computer Science
PES University, RR Campus
Bengaluru, India
surajx070@gmail.com

***Abstract*— The volatility of cryptocurrency markets poses significant challenges for investors and traders, necessitating robust predictive models to enhance decision-making. This research explores the efficacy of several machine learning models—Long Short-Term Memory (LSTM), Gradient Boosting Machine (GBM) using XGBoost, Random Forest, Artificial Neural Networks (ANN), and an ensemble LSTM-GRU model—in predicting Bitcoin prices. Additionally, the study integrates sentiment analysis of cryptocurrency-related tweets using the VADER sentiment analysis tool to augment predictive accuracy. Historical price data and Twitter sentiment scores underwent rigorous preprocessing before training and evaluation.**

The ensemble LSTM-GRU model, leveraging the combined strengths of LSTM and GRU architectures and enriched with sentiment scores, demonstrated superior performance compared to individual models. Our results indicate that integrating sentiment analysis enhances prediction accuracy, offering valuable insights for market participants. This study contributes to the field of financial forecasting by presenting a novel approach that combines advanced machine learning techniques with social media sentiment analysis, thereby providing practical tools for navigating cryptocurrency market volatility.

Keywords—Cryptocurrency, Price Prediction, Machine Learning, Sentiment Analysis, Financial Forecasting, Ensemble Models, Neural Networks

I. Introduction

Cryptocurrencies have emerged as a transformative force in global financial markets, characterized by their decentralized nature and significant volatility. Bitcoin, the pioneering cryptocurrency, exemplifies this volatility, with price fluctuations often defying traditional market analysis and prediction methods. For investors and traders in these markets, accurate forecasting of cryptocurrency prices is not merely advantageous but imperative for informed decision-making and risk management.

The traditional methods of financial analysis and prediction struggle to capture the complex dynamics inherent in cryptocurrency markets. Factors such as global regulatory shifts, technological advancements, and widespread media coverage contribute to rapid price fluctuations, challenging conventional models. Consequently, there is a growing interest in leveraging advanced computational techniques, particularly machine learning, to enhance the predictive capabilities for cryptocurrencies.

This study focuses on exploring the efficacy of various machine learning models in predicting Bitcoin prices, augmented by sentiment analysis of cryptocurrency-related tweets. Machine learning models such as Long Short-Term Memory (LSTM), Gradient Boosting Machine (GBM) using XGBoost, Random Forest, and Artificial Neural Networks (ANN) are examined alongside an ensemble LSTM-GRU model. These models are chosen for their ability to capture non-linear relationships and patterns in historical price data, which is essential given the volatile nature of cryptocurrency markets.

In addition to historical price data, sentiment analysis from social media platforms, particularly Twitter, is integrated using the VADER sentiment analysis tool. This approach aims to capture the collective sentiment of market participants reflected in social media discussions, which may influence cryptocurrency prices. By combining these two streams of data—historical price trends and sentiment analysis—this study seeks to enhance the accuracy and reliability of cryptocurrency price predictions.

The significance of this research lies in its potential to offer practical insights and tools for investors and traders navigating the unpredictable terrain of cryptocurrency markets. By evaluating and comparing the performance of various machine learning models and assessing the impact of sentiment analysis, this study contributes to advancing the field of financial forecasting in the context of digital assets.

II. Literature Review

Several studies have investigated the use of machine learning and deep learning models for cryptocurrency price prediction. Hamayel and Owda (2021) proposed a model using GRU, LSTM, and bi-LSTM algorithms to predict the prices of Bitcoin (BTC), Ethereum (ETH), and Litecoin (LTC). Their results indicated that the GRU model outperformed both LSTM and bi-LSTM in terms of prediction accuracy, with the lowest Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE) values.

Similarly, a study by Bonde et al. (2024) utilized LSTM neural networks to predict Bitcoin prices and highlighted the effectiveness of LSTM models in capturing long-term dependencies in time-series data. They also incorporated sentiment analysis from social media platforms, demonstrating that public sentiment can significantly impact price movements.

In another study, Patel et al. (2020) proposed a deep learning-based cryptocurrency price prediction scheme for financial institutions. Their model used LSTM and GRU networks to analyse historical price data and sentiment from social media, achieving notable prediction accuracy.

A. *Sentiment Analysis and Finanacial Forecasting*

Sentiment analysis has emerged as a critical component in financial forecasting, particularly in the context of cryptocurrencies. By analysing public sentiment expressed through social media and news articles, researchers have aimed to capture the psychological and behavioural aspects influencing market trends.

Bonde et al. (2024) emphasized the role of sentiment analysis in their cryptocurrency price prediction model. They used the VADER sentiment analysis tool to classify sentiments in tweets as positive, negative, or neutral, which were then integrated into their predictive models. Their findings underscored the importance of considering market sentiment to enhance the accuracy of price forecasts.

B. *Comparative studies and Model Evaluation*

Comparative studies have been conducted to evaluate the performance of different predictive models. Hamayel and Owda (2021) compared the performance of GRU, LSTM, and bi-LSTM models, demonstrating that GRU achieved the best results across multiple cryptocurrencies. Their evaluation metrics included MAPE and RMSE, which provided a comprehensive assessment of model accuracy.

Rebane et al. (2018) conducted a comparative study using Seq2Seq RNNs and ARIMA models for cryptocurrency prediction. They concluded that deep learning models generally outperformed traditional statistical methods like ARIMA in handling the complexities of cryptocurrency price data.

C. *Gaps and Future Directions*

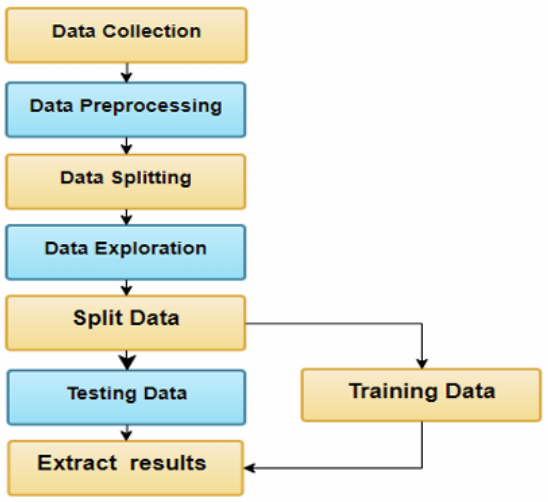
Despite the advancements in cryptocurrency price prediction, several research gaps remain. One significant gap is the limited comparison of multiple machine learning models within a single study. Most research focuses on a few selected models, but there is a need for comprehensive studies that compare a broader range of models, such as LSTM, GBM/XGBoost, Random Forest, ANN, and hybrid models like LSTM-GRU, as done in this project.

Additionally, the integration of extensive Twitter sentiment analysis into predictive models is relatively underexplored. While sentiment analysis has been incorporated in some studies, a more in-depth and extensive analysis of Twitter data, including the use of more sophisticated natural language processing techniques, could provide richer insights and improve prediction accuracy. Future research should aim to utilize larger datasets of tweets and apply advanced sentiment analysis methods to capture the nuanced sentiments of market participants.

Future research should explore the integration of more diverse data sources, including trading volumes and macroeconomic variables, to improve prediction models. Furthermore, the development of hybrid models that combine the strengths of various machine learning and deep learning techniques could provide more robust predictions.

III. METHODOLOGY

The figure below illustrates the sequence of methodology used to obtain results.



A. *Data Collection*

Historical Price Data: The historical price data for Bitcoin was collected from Yahoo Finance. This dataset includes daily closing prices, opening prices, high and low prices, trading volumes, and adjusted closing prices. The data spans several years, providing a comprehensive view of Bitcoin's price fluctuations over time. (July 2021 to July 2024)

Sentiment Data: Sentiment data was collected using the Tweepy Twitter API, which allows for the extraction of tweets related to Bitcoin. Additionally, a larger dataset of tweets was sourced from Mendeley (https://data.mendeley.com/datasets/8fbdhh72gs/5), providing a robust dataset for sentiment analysis. The tweets in these datasets were analysed to gauge public sentiment regarding Bitcoin. (Used a dataset of 39,000 bitcoin tweets that range in dates 2021 to end of 2023)

Date	Open	High	Low	Close	Adj Close	Volume
7/1/2021	35035.98	35035.98	32883.78	33572.12	33572.12	3.78E+10
7/2/2021	33549.6	33939.59	32770.68	33897.05	33897.05	3.87E+10
7/3/2021	33854.42	34909.26	33402.7	34668.55	34668.55	2.44E+10
7/4/2021	34665.57	35937.57	34396.48	35287.78	35287.78	2.49E+10
7/5/2021	35284.34	35284.34	33213.66	33746	33746	2.67E+10
7/6/2021	33723.51	35038.54	33599.92	34235.2	34235.2	2.65E+10
7/7/2021	34225.68	34997.66	33839.29	33855.33	33855.33	2.48E+10
7/8/2021	33889.61	33907.91	32133.18	32877.37	32877.37	2.99E+10
7/9/2021	32861.67	34042.29	32318.88	33798.01	33798.01	2.74E+10

Variable Name	Variable Description	Data Type
Date	Date of Observation	Date
Open	Opening price on the given day	Number
High	High price on the given day	Number
Low	Low price on the given day	Number
Close	Close price on the given day	Number

Fig. 1. Bitcoin Historical Prices Dataset

created_a	favorite_c	full_text	hashtags/(	hashtags/1	hashtags/2
7/1/2021	7	last friday, #eip1559 was	EIP1559		
7/1/2021	9	as part of taproot,	Bitcoin		
7/1/2021	2456	the			
7/1/2021	61	all coindesk 20 assets, which constitute roughly 99% of th			
7/2/2021	643	havenÃĆÂ€Â™t sold any coins from my xcad presale. curr			
7/2/2021	1474	wtf are you doing? @jmxfifa			
7/2/2021	7223	faze kay in the mud. pass it on			
7/2/2021	2937	according to wyckoff #b	BITCOIN		
7/2/2021	0	rt @jmxfifa: addressing			
7/2/2021	2508	cool, but youÃĆÂ€Â™re the owner. you shouldnÃĆÂ€Â™t			
7/2/2021	1494	glad you cleared it up. sad youÃĆÂ€Â™re stepping down b			
7/2/2021	5	as it stands now,	Bitcoin		
7/2/2021	3099	address this whole situation please @jmxfifa			
7/2/2021	12	what is #eip1559 and wh	EIP1559		
7/4/2021	1088	very bullish			

Fig. 2. Bitcoin Tweets dataset for Sentiment Analysis

B. *Data Preprocessing*

Price Data Preprocessing: The historical price data was preprocessed as follows:

Date Formatting: The 'Date' column was converted to a datetime format.

Indexing: The 'Date' column was set as the index.

Missing Values: Missing values in the price data were filled using forward fill (ffill) method to ensure continuity in the dataset.

Sentiment Data Preprocessing: The sentiment data was preprocessed to align with the price data:

Date Formatting: The 'created_at' column in the sentiment data was converted to a datetime format.

Indexing: The 'Date' column was set as the index.

Sentiment Analysis: The VADER sentiment analysis tool was used to calculate sentiment scores for each tweet. The compound sentiment score, which ranges from -1 (negative) to +1 (positive), was used as an indicator of the overall sentiment.

BERT Sentiment Analysis: A BERT-based sentiment analysis model was utilized to classify the sentiment of tweets as positive or negative. The model's scores were adjusted to align with the VADER compound score scale.

Combining Sentiment Scores: The sentiment scores from VADER and BERT were combined to produce a final sentiment compound score. This was done by averaging the scores from both models to leverage the strengths of each.

$$s(t) = \frac{s_{vader}(t) + s_{bert}(t)}{2}$$

Combining Price and Sentiment Data: The price and sentiment data were merged to create a unified dataset:

Joining Data: The price data and sentiment scores were combined using an outer join on the 'Date' index.

Handling Missing Values: Missing sentiment scores were filled with a default value of 0. Any remaining NaN values in the combined dataset were filled using both forward fill (ffill) and backward fill (bfill) methods.

Normalization: Normalization of the input data is essential to scale the features within a specific range, usually [0, 1]. This can be done using MinMaxScaler

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

C. *Model Development*

ANN Model: An Artificial Neural Network (ANN) is a computing system inspired by the biological neural networks of animal brains. ANNs consist of layers of nodes, or neurons, where each node is connected to others through weighted edges. The ANN learns patterns by adjusting these weights based on the error of predictions compared to actual outcomes, through a process known as backpropagation. In the context of Bitcoin price prediction, the ANN captures complex patterns in the historical data but may struggle with the non-linear and highly volatile nature of cryptocurrency prices, as reflected in its moderate performance.

It was developed with the following architecture:

Input Layer: Dense layer with 50 neurons and ReLU activation.

Hidden Layer: Dense layer with 25 neurons and ReLU activation.

Output Layer: Dense layer with 1 neuron. The model was compiled using the Adam optimizer and mean squared error (MSE) as the loss function.

GBM/XGBoost Model: Gradient Boosting Machine (GBM) and its optimized version, XGBoost, are ensemble learning methods that build a strong model by combining the predictions of several weaker models, typically decision trees. GBM works by sequentially adding trees to minimize a loss function, focusing on the errors made by previous models. XGBoost enhances this process with additional regularization and computational optimizations. Despite their success in various domains, their performance in Bitcoin price prediction may be less effective due to the highly non-linear and dynamic nature of the data, as indicated by the high RMSE and MAPE in your results.

The Gradient Boosting Machine (GBM) model was implemented using the XGBoost library. The model parameters included 100 estimators and a learning rate of 0.1. The model was trained using the training dataset, and predictions were made on the test dataset.

LSTM Model: Long Short-Term Memory (LSTM) networks are a type of recurrent neural network (RNN) designed to model sequences and time series data with long-term dependencies. LSTMs overcome the vanishing gradient

problem common in traditional RNNs by using memory cells to store and retrieve information over long periods. This capability makes LSTMs well-suited for Bitcoin price prediction, as they can effectively capture temporal dependencies and trends in the historical data, leading to improved performance compared to traditional models. The LSTM model is defined by the following set of equations:

$$
\begin{aligned}
f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\
i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\
\tilde{C}_t &= \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \\
C_t &= f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \\
o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\
h_t &= o_t \odot \tanh(C_t)
\end{aligned}
$$

where:

- f_t is the forget gate
- i_t is the input gate
- $\tilde{C}_t$ is the candidate cell state
- C_t is the cell state
- o_t is the output gate
- h_t is the hidden state
- σ is the sigmoid activation function
- $\odot$ denotes element-wise multiplication

It was developed with the following architecture:

Input Layer: LSTM layer with 50 units, returning sequences.

Hidden Layer: LSTM layer with 50 units, not returning sequences.

Dense Layer: Dense layer with 25 neurons.

Output Layer: Dense layer with 1 neuron. The model was compiled using the Adam optimizer and mean squared error (MSE) as the loss function.

Random Forest Model: Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the average prediction of individual trees. This technique helps in reducing overfitting and improving generalization. Each tree in the Random Forest considers a random subset of features, which adds to the model's robustness. However, while Random Forests are generally effective for a range of tasks, their performance in your study shows moderate accuracy, possibly due to the inherent complexity and variability in Bitcoin price movements. The Random Forest Regressor predicts the output by averaging the predictions from multiple decision trees:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} \hat{y}_i$$

where:

- $\hat{y}_i$ is the prediction from the i-th decision tree
- N is the number of trees in the forest

It was implemented with 100 estimators. The model was trained using the training dataset, and predictions were made on the test dataset.

LSTM-GRU Ensemble Model: The LSTM-GRU ensemble model combines Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) to leverage the strengths of both architectures. GRUs are a simplified variant of LSTMs, with fewer gates and parameters, which can lead to faster training and inference while still capturing temporal dependencies effectively. By combining LSTM and GRU predictions, the ensemble model can balance the strengths of both, leading to significantly better performance metrics (RMSE and MAPE) than individual models. The addition of Twitter sentiment data further enhances the model's accuracy, demonstrating the impact of social media sentiment on Bitcoin price fluctuations. The GRU model is defined by the following set of equations:

$$
\begin{aligned}
z_t &= \sigma(W_z \cdot [h_{t-1}, x_t] + b_z) \\
r_t &= \sigma(W_r \cdot [h_{t-1}, x_t] + b_r) \\
\tilde{h}_t &= \tanh(W_h \cdot [r_t \odot h_{t-1}, x_t] + b_h) \\
h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t
\end{aligned}
$$

where:

- z_t is the update gate
- r_t is the reset gate
- $\tilde{h}_t$ is the candidate hidden state
- h_t is the hidden state

The ensemble model combined predictions from both LSTM and GRU models. The architecture and process were as follows:

LSTM Model: As described above.

GRU Model: Similar to the LSTM model but using GRU layers instead of LSTM layers.

Ensemble Predictions: The ensemble predictions were computed using an exponential formula where the final prediction is a weighted sum of the LSTM and GRU predictions.

$$\alpha * LSTM + (1 - \alpha) * GRU$$

D. Evaluation Metrics

Metrics Used: The performance of each model was evaluated using two primary metrics:

Root Mean Squared Error (RMSE): This metric measures the standard deviation of the residuals (prediction errors). It provides a measure of how well the model's predictions match the actual values.

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

Mean Absolute Percentage Error (MAPE): This metric measures the accuracy of the model's predictions as a

percentage. It provides a normalized measure of prediction accuracy, allowing for comparison across different scales.

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

Comparison and Visualization: The performance of the models was compared using RMSE and MAPE. Visualizations, including plots of the actual vs. predicted prices, were used to illustrate the models' performance.

IV. Results

Model Performance:

TABLE I.

Model	Root Mean Squared Error (RMSE)	Mean Absolute Percentage Error (MAPE)
ANN	2620.93	3.38%
GBM/XgBoost	5,136.36	6.42%
Random Forest	3258.47	3.82%
LSTM	2560.15	3.42%
LSTM-GRU Ensemble (without twitter sentiments)	1182.86	1.64%
LSTM-GRU Ensemble (with twitter sentiments)	311.67	0.158%

*Evaluation metrics for predicted bitcoin prices

Comparison and Visualization:

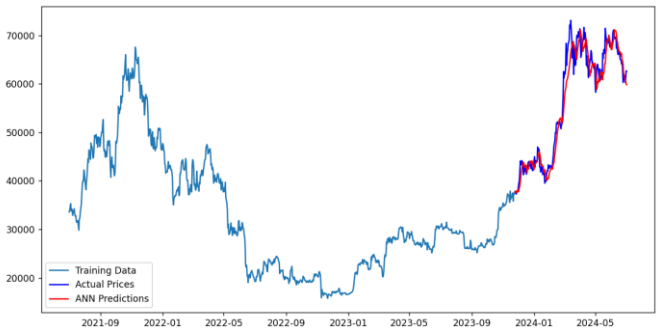


Fig. 3. Bicoin Price Prediction using ANN Model

Fig. 4. Bicoin Price Prediction using GBM Model

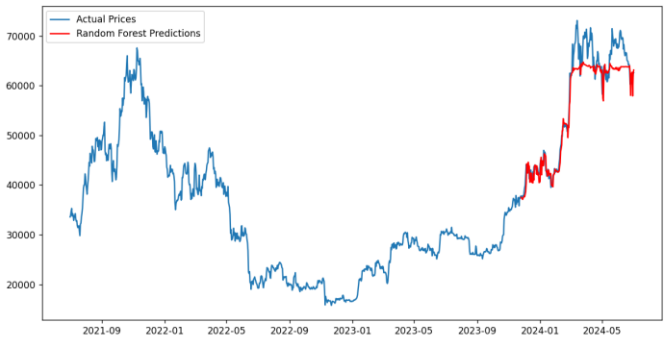


Fig. 5. Bicoin Price Prediction using Random Forest Model

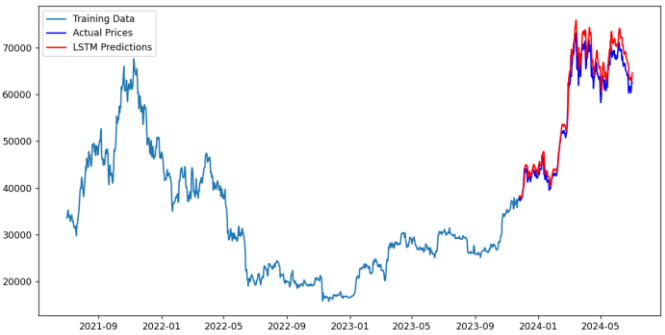


Fig. 6. Bicoin Price Prediction using LSTM Model

Fig. 7. Bicoin Price Prediction using LSMT-GRU ensemble Model (without twitter sentiment compound)

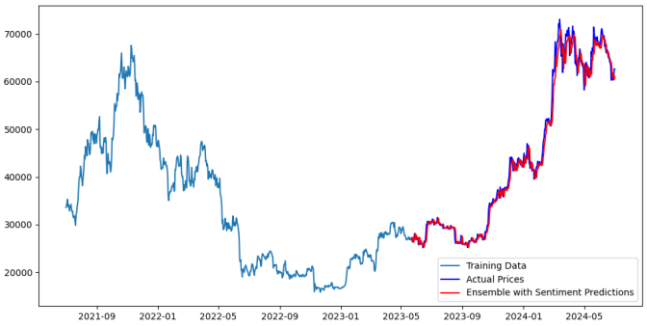


Fig. 8. Bicoin Price Prediction using LSMT-GRU ensemble Model (with twitter sentiment compound)

The results of your study, presented in Table I, highlight the performance of various models used to predict Bitcoin prices. The evaluation metrics, Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE), reveal significant insights. The Artificial Neural Network (ANN) model exhibits moderate performance with an RMSE of 2620.93 and a MAPE of 3.38%, while the Gradient Boosting Machine/XGBoost (GBM/XGBoost) model shows the highest RMSE and MAPE, indicating less effectiveness in predicting Bitcoin prices compared to other models. This

could be due to the complexity and non-linear nature of Bitcoin price movements, which the GBM/XGBoost model may struggle to capture. The Random Forest model, with an RMSE of 3258.47 and a MAPE of 3.82%, shows moderate performance as well, being outperformed by models like LSTM and the ensemble methods. The Long Short-Term Memory (LSTM) model, with an RMSE of 2560.15 and a MAPE of 3.42%, demonstrates improved performance, likely due to its ability to capture temporal dependencies in the data. The LSTM-GRU Ensemble model, without Twitter sentiments, dramatically outperforms the individual models, achieving a significantly lower RMSE of 1182.86 and a MAPE of 1.64%, suggesting that combining the strengths of both LSTM and GRU leads to more accurate predictions. Moreover, the LSTM-GRU ensemble model with Twitter sentiments achieves the best performance, with the lowest RMSE of 311.67 and a MAPE of 0.158%, indicating that integrating sentiment analysis further enhances predictive accuracy and highlights the impact of social media sentiment on Bitcoin price movements.

V. CONCLUSION

These findings imply that advanced machine learning techniques and sentiment analysis can provide more accurate and reliable predictions, which can lead to better-informed decision-making and potentially higher returns on investments. However, several limitations must be considered, such as the dependency on data quality, the computational complexity of ensemble models, the representativeness of sentiment analysis data, and the influence of additional market factors beyond historical prices and social media sentiment. Future work should focus on enhancing sentiment analysis by incorporating additional data sources, exploring more machine learning models and ensemble techniques, developing real-time prediction systems, and integrating other market factors like trading volumes, regulatory news, and macroeconomic indicators. In conclusion, this study demonstrates that ensemble models, particularly those incorporating sentiment analysis, significantly enhance the accuracy of Bitcoin price predictions, underscoring the importance of considering both historical data and market sentiment in predictive modeling. Future research should aim to address the identified limitations and further improve models to better capture the complexities of the cryptocurrency market.

REFERENCES

[1] Hamayel, M.J.; Owda, A.Y. ANovel Cryptocurrency Price Prediction Model Using GRU, LSTM and bi-LSTM Machine Learning Algorithms. AI 2021, 2, 477–496.

[2] Ho A, Vatambeti R, Ravichandran SK (2021) Bitcoin Price Prediction Using Machine Learning and Artificial Neural Network Model. Indian Journal of Science and Technology 14(27): 2300-2308. K. Elissa, “Title of paper if known,” unpublished.

[3] Gautam, K.; Sharma, N.; Kumar, P. Empirical Analysis of Current Cryptocurrencies in Different Aspects. In Proceedings of the ICRITO 2020—IEEE 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions), Noida, India, 4–5 June 2020; pp. 344–348.

[4] Killer, C.; Rodrigues, B.; Stiller, B. Security Management and Visualization in a Blockchain-based Collaborative Defense. In Proceedings of the ICBC 2019—IEEE International Conference on Blockchain and Cryptocurrency, Seoul, Korea, 14–17 May 2019; pp. 108–111.

[5] Nizzoli, L.; Tardelli, S.; Avvenuti, M.; Cresci, S.; Tesconi, M.; Ferrara, E. Charting the Landscape of Online Cryptocurrency Manipulation. IEEE Access 2020, 8, 113230–113245.