Time Series Analysis of Bitcoin Prices Using ARIMA and LSTM for Trend Prediction

Berlilana[1,*], Arif Mu'amar[2]

[1,2]Magister of Computer Science, Computer Science Faculty, Universitas Amikom Purwokerto, Indonesia

ABSTRACT

This study investigates the efficacy of ARIMA and LSTM models in predicting Bitcoin prices, emphasizing the importance of accurate price prediction for trading, risk management, and investment strategies in the volatile cryptocurrency market. The objectives are to analyze Bitcoin prices to identify underlying patterns and trends, compare the predictive performance of ARIMA and LSTM models, and provide insights into their practical applications for Bitcoin price prediction. A comprehensive dataset of Bitcoin prices from January 1, 2011, to December 31, 2023, sourced from CoinMarketCap, was used. Data preprocessing included handling missing values, removing duplicates, achieving stationarity through differencing, and normalizing data using MinMaxScaler. The ARIMA model's best-fitting parameters were identified using ACF and PACF plots, and it was trained with the statsmodels library. The LSTM model involved data preparation through windowing and train-test splitting, constructing a neural network with LSTM layers, and training using TensorFlow/Keras. Evaluation metrics included Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), with comparisons based on accuracy and computational efficiency. The ARIMA model demonstrated impressive performance with an MAE of 2.308392356829177e-215 and an RMSE of 0.0, indicating a near-perfect fit to the training data. The LSTM model achieved an MAE of 0.00021804577826689423 and an RMSE of 0.00021916977109865863, showing robust performance in handling nonlinear and long-term dependencies. The ARIMA model excelled in computational efficiency with a training time of 2.548070192337036 seconds and a prediction time of 0.0009970664978027344 seconds, while the LSTM model required 378.69622468948364 seconds for training and 0.6859967708587646 seconds for prediction. The results highlight ARIMA's effectiveness in capturing linear trends and its suitability for short-term trading strategies, while LSTM is better for long-term investment strategies due to its ability to model complex patterns. Despite potential overfitting in ARIMA and high computational demands for LSTM, the study suggests exploring hybrid models, incorporating additional data sources, and developing advanced techniques to enhance predictive accuracy in future research.

Keywords Time Series Analysis, Bitcoin Price, ARIMA, LSTM, Trend Prediction

Submitted 10 January 2024
Accepted 20 April 2024
Published 1 June 2024

Corresponding author
Berlilana,
berli@amikompurwokerto.ac.id

Additional Information and Declarations can be found on page 101

INTRODUCTION

Bitcoin is a decentralised digital currency created in 2009 by an unknown person or group using the alias Satoshi Nakamoto. Bitcoin is a cryptocurrency that operates on blockchain technology, a decentralised infrastructure allowing secure transactions between users without the need for a trusted third party [1]. Bitcoin revolutionised the digital currency space by offering a payment system based on cryptographic proof rather than trust [2]. The white paper published by Nakamoto outlined Bitcoin as a peer-to-peer electronic cash system, marking the inception of Bitcoin and peer-to-peer cryptocurrencies [3]. Since its introduction, Bitcoin has gained significant popularity and attention, leading to

How to cite this article: Berlilana and A. Mu'amar, "Time Series Analysis of Bitcoin Prices Using ARIMA and LSTM for Trend Prediction," *J. Digit. Mark. Digit. Curr.*, vol. 1, no. 1, pp. 84-102, 2024.

the emergence of numerous other cryptocurrencies in the digital currency markets [4].

Blockchain technology underpins Bitcoin, providing a distributed and traceable transaction ledger, ensuring accuracy and security [5]. The consensus mechanism in blockchain guarantees the integrity and immutability of transactions, offering users decentralization, autonomy, and transparency in a trustless environment [6]. Bitcoin's design allows any two willing parties to transact directly with each other, eliminating the need for intermediaries and enhancing the efficiency of transactions [2]. The cryptographic hashing utilized in Bitcoin's Proof of Work (PoW) algorithm forms the basis of blockchain technology, ensuring the security and integrity of the network [7].

Bitcoin's role extends beyond being a currency; it also functions as a financial asset, attracting investors and researchers seeking to understand its risk-return properties and market liquidity [8]. Introducing Bitcoin futures on traditional exchanges has legitimized Bitcoin as an asset class, providing institutional traders with avenues to enter the market and facilitating mechanisms for shorting Bitcoin [9]. The volatility and uncertainty surrounding Bitcoin's price path have been notable, with its value being influenced by various factors such as market capitalization and trading volume.

Bitcoin's influence extends beyond its ecosystem, significantly impacting the broader cryptocurrency market. As the first and most successful cryptocurrency, Bitcoin laid the foundation for developing thousands of alternative cryptocurrencies (altcoins) that followed in its wake [10]. With Bitcoin accounting for a significant portion of the total cryptocurrency market capitalization, it has been a driving force behind the growth and diversification of the cryptocurrency ecosystem [11]. Many of these altcoins aim to address the perceived limitations of Bitcoin, offering different features and use cases. Some altcoins are derivatives of Bitcoin with minor protocol modifications, while others have adopted entirely different protocols [12]. However, Bitcoin remains the dominant player, often setting the market trends for the entire cryptocurrency sector.

In addition to its role as a leading digital asset, Bitcoin has been instrumental in developing and popularising blockchain technology. The blockchain, the underlying technology behind Bitcoin, is a decentralized ledger that records all transactions across a network of computers. This technology ensures transparency, security, and immutability of transaction data. Beyond cryptocurrencies, blockchain technology is being explored for various applications, including supply chain management, voting systems, and decentralized finance (DeFi).

Bitcoin is dominant in the cryptocurrency market, often called the "king of cryptocurrencies." This market dominance is evidenced by its significant market capitalization, which frequently accounts for a large proportion of the total market capitalization of all cryptocurrencies combined. Bitcoin’s market dominance impacts the valuation of other cryptocurrencies, commonly known as altcoins. When Bitcoin's price rises or falls, it often influences the price movements of altcoins, creating a correlated market dynamic. This phenomenon occurs because Bitcoin is widely viewed as the anchor of the cryptocurrency market, with investor sentiment toward Bitcoin often spilling over into other digital assets.

As a benchmark for cryptocurrency market performance, Bitcoin further underscores its significance. Studies have shown that Bitcoin exhibits predictive power over the cryptocurrency market, further solidifying its position as a key player in the industry [13]. Additionally, Bitcoin's popularity and representation in the cryptocurrency market make it a crucial reference point for assessing its overall performance and dynamics [14]. Many investors and analysts use Bitcoin as a reference point to gauge the health and trends of the broader cryptocurrency market. The correlation between Bitcoin's price and many altcoins can be attributed to several factors, including shared market sentiment, investor behaviour, and liquidity flows. As a result, significant events affecting Bitcoin, such as regulatory news, technological advancements, or macroeconomic shifts, can have ripple effects throughout the entire cryptocurrency ecosystem.

The adoption of Bitcoin by institutions has been a crucial factor in solidifying its role in the financial landscape. Major corporations such as Tesla and MicroStrategy have added Bitcoin to their balance sheets, signalling a growing acceptance of Bitcoin as a legitimate investment asset. Additionally, integrating Bitcoin into financial products has further enhanced its accessibility and appeal to a broader range of investors. Financial products such as Bitcoin futures and exchange-traded funds (ETFs) have allowed institutional and retail investors to gain exposure to Bitcoin without the need to hold the cryptocurrency directly. These financial instruments have contributed to increased liquidity and market stability while providing new avenues for portfolio diversification.

Furthermore, Bitcoin's consideration in macroeconomic analysis highlights its emerging role in the global financial system. As a decentralized asset with a limited supply, Bitcoin is often viewed as a hedge against inflation and currency devaluation, particularly during economic uncertainty. Analysts and economists increasingly consider Bitcoin in their assessments of global economic trends, monetary policies, and investment strategies. The inclusion of Bitcoin in macroeconomic analysis reflects its growing importance as a store of value and a potential alternative to traditional safe-haven assets like gold.

Accurate price prediction is critical in trading, risk management, and investment strategies, particularly in the highly volatile cryptocurrency market. For traders, the ability to predict price movements with high accuracy can significantly enhance profitability [15]. Precise predictions allow traders to effectively time their buy and sell orders, capitalizing on favourable market conditions and avoiding potential losses. In fast-moving markets like Bitcoin, even small improvements in prediction accuracy can lead to substantial financial gains.

In the realm of risk management, accurate price prediction is equally vital. Investors and traders rely on forecasting models to assess potential future price movements and to devise strategies that minimize exposure to adverse market conditions. By anticipating market trends and price fluctuations, risk managers can implement hedging strategies, set appropriate stop-loss orders, and allocate assets to balance risk and return. This proactive approach helps protect investment portfolios from significant drawdowns and enhances overall financial stability.

Accurate price prediction supports informed decision-making and portfolio optimization in investment strategies. Investors use price forecasts to identify potential investment opportunities and determine the best times to enter or exit positions. Reliable predictions enable investors to construct diversified portfolios that maximize returns while adhering to their risk tolerance levels. This strategic allocation of assets, guided by accurate price forecasts, leads to better long-term investment performance and helps achieve financial goals.

The volatility of Bitcoin underscores the necessity of reliable forecasting models. Bitcoin prices can experience rapid and substantial changes due to many factors, including market sentiment, regulatory developments, technological advancements, and macroeconomic shifts [16]. This inherent volatility presents both opportunities and challenges for market participants. On the one hand, it offers the potential for significant profits; on the other hand, it poses substantial risks.

With accurate price prediction models, investors and traders are protected from sudden and unpredictable market swings. Reliable forecasting models, such as those based on ARIMA and LSTM, are essential for navigating this volatility. These models can capture complex patterns and trends in historical data, providing valuable insights into future price movements. By leveraging advanced predictive techniques, market participants can gain a competitive edge, making well-informed decisions that enhance profitability and reduce risk exposure.

Bitcoin price prediction presents numerous challenges due to the high volatility and rapid fluctuations inherent in the cryptocurrency market. Bitcoin prices can experience significant changes within short periods, driven by market sentiment, regulatory news, technological advancements, and macroeconomic developments. These fluctuations create a highly unpredictable environment that complicates the task of accurate price forecasting.

One of the primary challenges is capturing the non-linear and complex dynamics of Bitcoin price movements. Traditional financial models often need to catch up because they assume linear relationships and rely on historical trends to make predictions. However, many factors that interact in complex and non-linear ways influence Bitcoin's price behaviour. For instance, sudden regulatory announcements can trigger sharp price swings that traditional models may not account for, leading to inaccurate predictions.

Furthermore, the influence of external factors adds another layer of complexity to Bitcoin price prediction. Geopolitical events can immediately and profoundly impact market confidence and Bitcoin prices. For example, geopolitical tensions or economic sanctions can lead to shifts in investor behaviour, causing abrupt price changes. Technological changes in blockchain, such as upgrades or security breaches, can also significantly impact Bitcoin’s value. Innovations that enhance scalability and security might boost investor confidence, while security vulnerabilities can lead to panic selling.

The primary objective of this study is to analyze Bitcoin prices by identifying and understanding the underlying patterns and trends that govern their price movements over different time periods. This includes examining how Bitcoin prices behave on daily, weekly, monthly, and yearly intervals to gain insights into market behaviour and assist in predicting future price movements. The study aims to detect various effects and patterns that influence Bitcoin prices by delving into the historical price data.

Specifically, the study will explore historical price data to identify seasonal effects, such as recurring patterns corresponding to specific times of the year, months, or weeks. For instance, the analysis will determine if there are consistent price increases or decreases during particular seasons or periods.

Additionally, the study will analyze longer-term cyclical patterns that may influence Bitcoin prices, which could be related to broader market cycles, economic conditions, or Bitcoin-specific events such as halving cycles. Lastly, the study will assess the overall trajectory of Bitcoin prices over extended periods, identifying upward or downward trends and understanding the factors contributing to these trends, which can include technological advancements, regulatory developments, and macroeconomic conditions.

The study aims to further enhance the understanding of Bitcoin price dynamics by modeling and predicting future Bitcoin prices using advanced predictive techniques. To achieve this, the study will utilize ARIMA (Autoregressive Integrated Moving Average) to model linear dependencies and seasonal components in Bitcoin prices. ARIMA is well-suited for capturing linear patterns and periodic fluctuations in time series data, making it an appropriate choice for modelling Bitcoin's historical price trends.

In addition to ARIMA, the study will apply LSTM (Long Short-Term Memory) neural networks to capture non-linear relationships and long-term dependencies in the data. LSTM networks are a type of recurrent neural network (RNN) that are particularly effective in handling sequential data and overcoming the limitations of traditional models by learning from past data to make accurate predictions. This approach will enable the study to account for Bitcoin price movements' complex and non-linear dynamics.

The study will also compare the predictive performance of ARIMA and LSTM models to determine their effectiveness in forecasting Bitcoin prices. By evaluating these models using metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), the study aims to identify which model provides more accurate and reliable predictions. This comparative analysis will offer valuable insights into the strengths and limitations of each modelling approach, ultimately contributing to more informed decision-making in trading and investment strategies within the volatile cryptocurrency market.

The study aims to provide a comprehensive understanding of Bitcoin price dynamics by achieving these objectives. This understanding can inform better trading strategies, enhance risk management, and support more informed investment decisions in the volatile cryptocurrency market.

Literature Review

Overview of Time Series Analysis in Financial Markets

The origins of time series analysis can be traced back to the early use of statistical methods in economic and financial contexts. In recent years, the field has incorporated machine learning techniques, particularly with the advent of neural networks and deep learning models. Time series analysis is crucial in understanding market behaviour by helping analysts identify trends, cycles, and seasonal patterns in financial data. Analysts can uncover underlying patterns that drive market movements by examining how variables change over time. Trends indicate the general direction in which a market or security is moving, whether upward, downward, or sideways. Cycles refer to regular long-term fluctuations, often influenced by economic or business cycles. Seasonal patterns are periodic fluctuations at specific times of the year, such as increased retail sales during the holiday season. Understanding these patterns allows investors and traders to make informed decisions based on historical data and anticipated future movements.

One of the primary applications of time series analysis is predictive analytics, which is used to forecast future market movements. By modeling historical data, analysts can predict a financial variable's future values. Techniques such as ARIMA, exponential smoothing, and neural networks like LSTM are commonly employed to create these forecasts. Accurate predictions enable market participants to anticipate changes, adjust their strategies, and capitalize on potential opportunities. For instance, predicting a stock's price movement can inform buying or selling decisions, while forecasting economic indicators can guide macroeconomic policy decisions. Time series analysis thus provides a powerful tool for planning and decision-making in financial markets.

Time series analysis is also integral to risk management in financial markets. Analysts can assess the risk associated with different financial instruments and portfolios by analysing historical price data and volatility patterns. Techniques such as Value at Risk (VaR), Conditional Value at Risk (CVaR), and stress testing rely on time series analysis to estimate potential losses under various market conditions. This analysis helps identify periods of high volatility or market stress, enabling risk managers to implement strategies to mitigate these risks. For example, during anticipated market turbulence, a portfolio manager might adjust asset allocations or employ hedging strategies to protect against significant losses. Additionally, time series models can be used to monitor and manage the risk of trading strategies by evaluating their historical performance and potential future risks.

ARIMA Models

The ARIMA model is composed of three key components: autoregressive (AR), integrated (I), and moving average (MA). AR Component: The AR part of the model involves regressing the variable on its own lagged (past) values. This means that the current value of the series is expressed as a function of its previous values. Integrated (I) Component: The integrated part refers to the differencing of the data to achieve stationarity. Stationarity means that the statistical properties of the series, such as mean and variance, are constant over time. Differencing involves subtracting the previous observation from the current observation. MA Component: The MA part of the model consists in modelling the error term as a linear combination of past error terms. This means that past forecast errors influence the current value of the series.

One of the most common applications of ARIMA models in financial forecasting is predicting stock market trends. Many factors, including company performance, economic conditions, and market sentiment influence stock prices. ARIMA models help capture historical dependencies and trends within stock price data, allowing analysts to make informed predictions about future movements. By analyzing past stock prices, ARIMA can identify patterns and potential turning points, aiding traders and investors in decision-making. The model's ability to handle non-stationary data by differencing makes it particularly useful in dealing with the volatile nature of stock prices. Numerous studies have demonstrated the effectiveness of ARIMA models in short-term stock price forecasting, providing a valuable tool for portfolio management and trading strategies.

ARIMA models have been widely applied to predict the prices of cryptocurrencies, particularly Bitcoin, due to their ability to handle time series data effectively. Numerous studies have explored the use of ARIMA for forecasting Bitcoin prices, given its high volatility and complex market dynamics. In a study by [17], the ARIMA (1,1,0) model was found to be efficient in forecasting quarterly price movements of Bitcoin for the last two quarters of 2020. The study suggested that the deviation in Bitcoin's price during this period might indicate a shift in its perceived investment value to investors as a digital asset following the COVID-19 outbreak. This underscores the utility of ARIMA models in capturing short-term fluctuations in cryptocurrency prices, which can be influenced by external factors such as global events and market sentiment.

The accuracy and performance of ARIMA models in cryptocurrency price prediction have been evaluated through various performance metrics such as MAE, RMSE, and MAPE. Studies generally report that ARIMA models can achieve good predictive accuracy over short-term periods, where market conditions remain relatively stable. However, the performance of ARIMA models tends to vary with the time horizon and market volatility. While they perform well for short-term predictions, their accuracy must improve for longer-term forecasts due to the increasing uncertainty and the non-stationary nature of cryptocurrency prices.

Despite their strengths, ARIMA models face challenges and limitations when applied to highly volatile assets like cryptocurrencies. One of the main limitations is their reliance on historical data and the assumption that past patterns will continue. This assumption often needs to be revised in the context of cryptocurrencies, which are subject to rapid and unpredictable market changes. The high volatility and frequent market disruptions can lead to significant forecast errors.

Additionally, ARIMA models are naturally linear and may need help capturing complex, non-linear relationships in cryptocurrency price movements. This limitation can result in suboptimal predictions during market turbulence or when external factors significantly impact prices. Furthermore, the model's effectiveness depends on properly identifying parameters (p, d, q), which can be challenging and time-consuming, especially with volatile time series data.

Another challenge is the potential for overfitting, where the model becomes too closely fitted to historical data and loses its generalizability to future data. Overfitting can lead to inaccurate predictions when market conditions change. Researchers often combine ARIMA with other models or techniques, such as machine learning algorithms, to mitigate these issues and enhance predictive performance and robustness.

LSTM Models

Recurrent Neural Networks (RNNs) are a class of artificial neural networks designed to recognize patterns in data sequences, such as time series, speech, or text. Unlike traditional feedforward neural networks, RNNs have connections that form directed cycles, allowing them to maintain a "memory" of previous inputs. This memory enables RNNs to capture temporal dependencies and context, making them suitable for tasks where the order of inputs matters.

However, RNNs have significant limitations. One of the most prominent issues is the vanishing gradient problem. During training, the gradients of the loss function are propagated backward through the network to update the weights. In deep RNNs, these gradients can become very small (or very large, known as the exploding gradient problem) after several steps, leading to minimal weight updates and poor learning of long-term dependencies. This problem makes it challenging for standard RNNs to capture long-term patterns in data, limiting their effectiveness in tasks requiring long memory spans.

Long Short-Term Memory (LSTM) networks, a type of RNN, were designed to address the vanishing gradient problem and improve the learning of long-term dependencies. LSTM networks achieve this through a unique architecture that includes a series of gates to control the flow of information. An LSTM cell consists of three main gates: the input gate, the forget gate, and the output gate. Input Gate determines how much new information from the current input should be added to the cell state. It decides what proportion of the input should be stored in the cell. Forget Gate decides what portion of the information in the cell state should be discarded. It controls the extent to which the previous cell state should be forgotten or retained. Output Gate determines the output of the LSTM cell based on the cell state and the input. It decides what part of the cell state should be output as the current step's hidden state.

LSTM networks offer several advantages over traditional RNNs, primarily due to their ability to overcome the vanishing gradient problem. The gating mechanisms in LSTM cells allow them to effectively capture and maintain long-term dependencies in the data, making them particularly suitable for tasks involving long sequences. This ability is crucial for applications such as language modeling, speech recognition, and time series forecasting, where context and temporal dependencies are essential. By retaining relevant information over long periods, LSTMs can learn complex patterns and relationships in sequential data more effectively than standard RNNs. This improved learning capability makes LSTM networks more robust and reliable for various sequence-based tasks, resulting in better performance and accuracy. Moreover, LSTMs have been widely adopted in diverse fields, from natural language processing to finance, due to their versatility and effectiveness in handling temporal dependencies.

LSTM networks are highly relevant in time series prediction because they can model complex, non-linear dependencies. Unlike traditional linear models, LSTMs can capture intricate relationships within the data that arise from various underlying factors influencing the time series. This capability is particularly important in financial markets, where economic indicators, market sentiment, and external events often drive price movements. By learning these non-linear patterns, LSTMs provide more accurate and robust forecasts, making them a powerful tool for financial analysts and traders.

One of the key strengths of LSTM networks is their ability to capture long-term dependencies in time series data. Financial markets often exhibit patterns and trends that span long periods, such as economic cycles, seasonal effects, and long-term growth trends. Traditional models struggle to maintain these dependencies over extended sequences due to the vanishing gradient problem. However, LSTMs, with their specialized gating mechanisms, can effectively remember information from earlier time steps, allowing them to model these long-term dependencies accurately. This makes LSTMs particularly suitable for financial time series forecasting, where capturing long-term trends is crucial for making informed investment decisions.

LSTM networks offer scalability and flexibility, allowing them to be applied across different time scales and financial instruments. Whether dealing with high-frequency trading data, daily stock prices, or monthly economic indicators, LSTMs can adapt to various temporal resolutions. This flexibility extends to financial instruments, including stocks, bonds, commodities, and cryptocurrencies. The scalability of LSTMs means they can handle large datasets and complex models, making them suitable for applications ranging from short-term trading strategies to long-term investment planning. Several key studies have applied LSTM networks to predict Bitcoin prices, demonstrating their effectiveness in this highly volatile market. A comparative study by [18] assessed various deep learning methods, including LSTM, for Bitcoin price prediction, underscoring LSTM's effectiveness in capturing intricate patterns within cryptocurrency price data.

In various financial contexts, LSTM networks have been shown to outperform traditional models such as ARIMA, moving averages, and linear regression. Research has demonstrated that LSTMs can better capture financial time series' non-linear and dynamic nature, leading to more accurate predictions. Researchers continue to develop innovative techniques and improvements in LSTM applications to enhance their predictive performance. Some studies have explored hybrid models that combine LSTMs with other machine learning algorithms or traditional models. Other advancements include using attention mechanisms to focus on important time steps, incorporating ensemble methods to aggregate multiple model predictions, and applying transfer learning to leverage knowledge from related domains. These innovative approaches demonstrate the ongoing evolution and potential of LSTM networks in advancing financial time series prediction.

Comparison of ARIMA and LSTM

ARIMA and LSTM models have distinct strengths and weaknesses that make them suitable for different time series forecasting tasks. ARIMA models are valued for their simplicity and ease of interpretation. The underlying statistical principles are well-established, making ARIMA accessible for analysts with basic statistical knowledge. The parameters of ARIMA models, which include autoregressive, differencing, and moving average components, are easy to interpret, providing clear insights into the structure of the time series. ARIMA models are particularly effective at capturing linear relationships and trends in time series data, making them suitable for many financial forecasting applications where such patterns are prevalent. However, ARIMA models have notable limitations. They cannot capture non-linear relationships, so they may need help to model complex, non-linear dependencies in time series data. This limitation can result in suboptimal performance for datasets with significant non-linearities [19].

LSTM networks excel in areas where ARIMA models fall short. They are designed to retain information over long sequences, allowing them to capture long-term dependencies and trends that traditional models might miss. This capability is particularly advantageous for financial time series data, where long-term patterns are often crucial. LSTM networks are highly effective at modelling nonlinear relationships in time series data and learning complex patterns and dependencies that make them suitable for forecasting in volatile and dynamic environments. However, LSTM networks come with their own set of challenges. They require significant computational resources, both in terms of processing power and memory. Training these models can be resource-intensive, especially for large datasets or complex network architectures. Tuning an LSTM model involves selecting various hyperparameters, such as the number of layers, units per layer, learning rate, and dropout rate. This complexity can make it challenging to optimize the model's performance. Furthermore, LSTM networks typically have longer training times than simpler models like ARIMA, which can be a drawback in scenarios where quick model deployment is essential [20].

Method

The methodology flowchart provides a visual representation of the overall process employed in this study, from data collection and preprocessing to model implementation and evaluation, as shown in the figure 1.

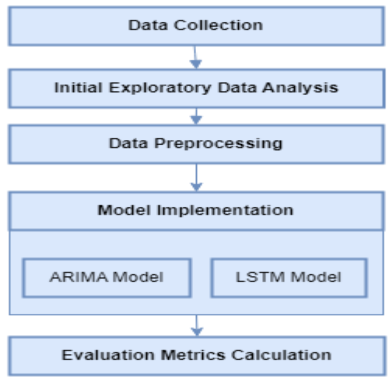

Figure 1 Research Method Flowchart

Data Collection

The Bitcoin price dataset is sourced from a reliable cryptocurrency market platform like CoinMarketCap. It spans a comprehensive time frame from January 1, 2011, to December 31, 2023, providing a robust historical record of Bitcoin price movements. The dataset contains 5014 entries and seven key features, recorded daily.

The features in the dataset are as follows: The "Date" indicates the date on which the Bitcoin price was recorded. "Price" represents the closing price of Bitcoin at the end of the trading period. "Open" indicates the price of Bitcoin at the beginning of the trading period. "High" denotes the highest price Bitcoin reached during the trading period. "Low" represents the lowest price observed during the trading period. "Vol." (Volume) shows the total volume of Bitcoin traded during the trading period. "Change %" reflects the percentage change in Bitcoin's price from the opening to the closing price within the trading period.

These features are significant as they provide comprehensive insights into the daily price fluctuations and trading volumes of Bitcoin, facilitating a detailed analysis of market trends and patterns. The summary statistics reveal that there are 5014 records for most features, with slight discrepancies for the volume feature, which has 5008 entries. Each date entry is unique, ensuring a consistent timeline. The dataset includes various entries for each price feature, with the most frequently occurring prices being low-value entries, such as 0.1, highlighting the granularity and range of the dataset. This detailed dataset allows for an in-depth analysis of Bitcoin's price dynamics over the years, aiding in understanding market behaviours and making informed predictions.

Exploratory Data Analysis (EDA)

Initial EDA was conducted to gain an understanding of the raw Bitcoin price data, identify missing values, detect outliers, and uncover any potential issues that might affect subsequent analysis. The dataset comprises the columns 'Date', 'Price', 'Open', 'High', 'Low', 'Vol.', and 'Change %'. After converting the 'Date' column to datetime format, all other columns were confirmed to be of type float64. In the initial data exploration, it was noted that there are no missing values in most columns except for the 'Vol.' column, which has six missing entries. Outlier detection revealed many outliers in the dataset: 452 in the 'Price' column, 462 in 'Open', 461 in 'High', 469 in 'Low', and 772 in 'Vol.'. These outliers could skew the analysis and need careful consideration.

Descriptive statistics provided a comprehensive overview of the dataset. The mean price of Bitcoin was approximately 10,987.03 USD, with a standard deviation of 16,186.05, indicating high variability. The minimum recorded price was 0.10 USD, while the maximum reached 73,066.30 USD. The 'Open' and 'High' columns showed similar statistical characteristics, with their mean values close to the overall average price and a standard deviation reflecting the volatility in Bitcoin prices. The 'Low' prices ranged from 0.00 to 71,338.40 USD, averaging 10,682.95 USD. The trading volume ('Vol.') varied widely, from 80 to over 4.47 billion, with a mean value of approximately 12.34 million. The 'Change %' column, reflecting the percentage change from opening to closing prices, had values ranging from -57.21% to 336.84%, with an average close to 0.42%, showing that daily price changes can be quite significant.

Visualizations were employed further to explore the data distributions and relationships between variables. Time series plots highlighted the trend and volatility in Bitcoin prices over time as shown in figure 2.

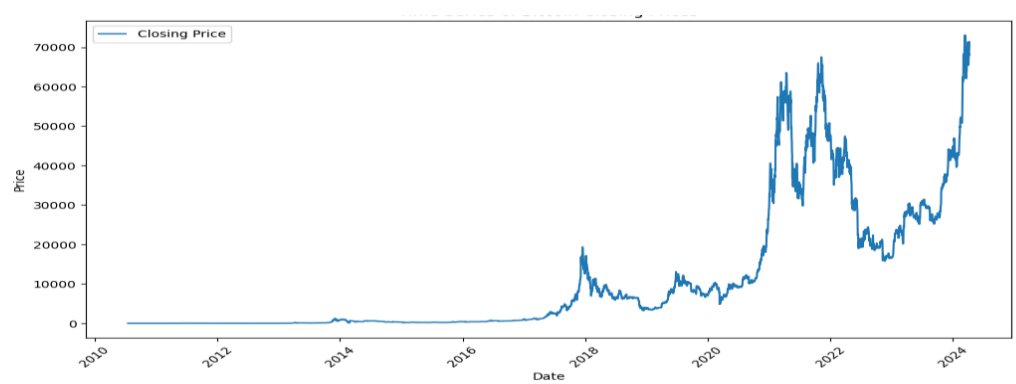

Figure 2 Time Series of Bitcoin Closing Prices

Histograms provided insights into the distribution of prices, volumes, and percentage changes shown in figure 3.

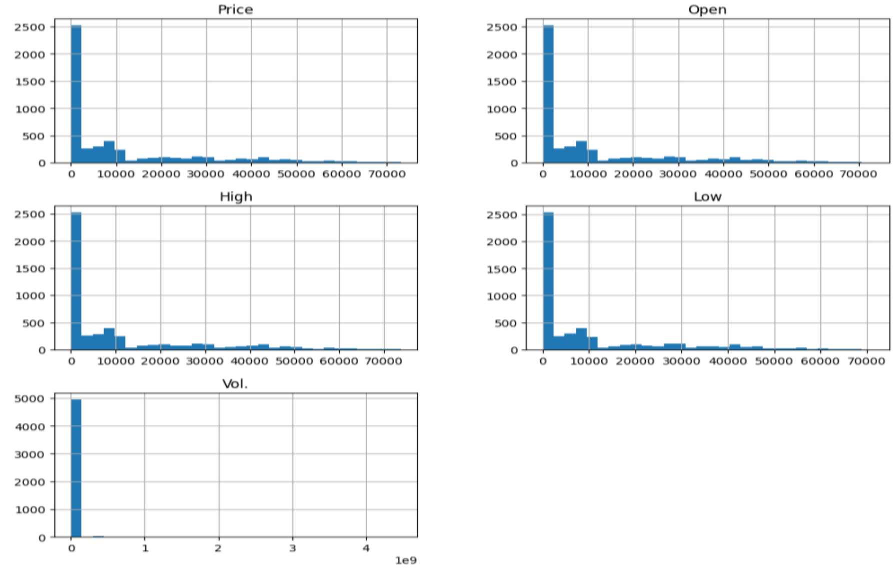

Figure 3 **Histogram of Price, Open, High, Low and Vol Features**

This initial EDA was crucial in understanding the underlying patterns and potential anomalies in the Bitcoin price data, setting the stage for more detailed modelling and analysis.

Data Preprocessing

Data preprocessing involved several key steps to ensure the dataset was clean, stationary, and normalized, making it suitable for subsequent analysis and modelling. The initial step in data cleaning addressed missing values. The dataset had six missing entries in the 'Vol.' column, handled appropriately to maintain data integrity. Next, any duplicate entries were identified and removed to prevent redundancy and ensure the dataset's accuracy.

To prepare the data for ARIMA modeling, which requires stationarity, differencing was applied to the 'Price' column. This transformation helps stabilize the mean of the time series by removing changes in the level of the series, thereby achieving stationarity. The stationarity of the differenced series was confirmed using the Augmented Dickey-Fuller (ADF) test. The test results strongly rejected the null hypothesis of non-stationarity, with an ADF statistic of -10.001992921190519 and a p-value significantly lower than the 1% critical value threshold. The critical values for the 1%, 5%, and 10% significance levels were -3.4316653755446955, -2.862121253084995, and -2.5670794018368777, respectively, confirming the stationarity of the differenced series.

For the LSTM model, normalization was crucial due to the model's sensitivity to the scale of input data. The data was normalized using MinMaxScaler, which scales the features to a given range, typically between 0 and 1. This step ensures that the neural network can learn more effectively by preventing features with larger ranges from dominating those with smaller ranges. The normalization process was completed successfully, and the preprocessed dataset was saved to 'preprocessed_dataset.csv', ensuring that the data was ready for both ARIMA and LSTM modeling.

After the initial data cleaning steps, a further EDA was conducted to confirm the effectiveness of the cleaning process. The cleaned dataset retained the same structure as before, with the 'Date' column in datetime format and the other columns ('Price', 'Open', 'High', 'Low', 'Vol.', and 'Change %') in float64 format. Additionally, a new column 'Price_diff' was introduced, representing the differenced values of the 'Price' column to achieve stationarity for ARIMA modeling.

The post-cleaning exploration revealed no missing values in the dataset, except for one missing value in the 'Price_diff' column, which is expected due to the differencing process. The updated summary statistics provided a detailed overview of the cleaned data. The dataset included 5008 entries for each feature, except for 'Price_diff', which had 5007 entries due to the missing value mentioned earlier.

The mean values of 'Price', 'Open', 'High', and 'Low' columns were around 0.15, indicating that the normalization process had effectively scaled the data. The trading volume ('Vol.') had a mean value of approximately 0.002761, with a standard deviation reflecting significant variability in the volume traded. The 'Change %' column, representing the percentage change from opening to closing prices, had a mean close to 0.416430, showing that the data contained a mix of daily price increases and decreases. The 'Price_diff' column had a mean of -14.249611, indicating the average daily change in price, with a wide range from -7542.800000 to 7311.500000, reflecting high volatility.

Visualizations were used to illustrate changes in data distributions and relationships after cleaning. Time series plots highlighted the normalized and differenced data, providing clear insights into the trends and volatility over time. Histograms displayed the distribution of each feature, showing a more uniform spread after normalization. Scatter plots helped visualize the relationships between different features, confirming the effectiveness of the cleaning and preprocessing steps. This thorough post-cleaning EDA ensured that the dataset was ready for accurate and robust modeling in subsequent analysis.

Model Implementation

Implementing the ARIMA model began with model selection, which involved identifying the best-fitting model using Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots, as shown in figure 4.

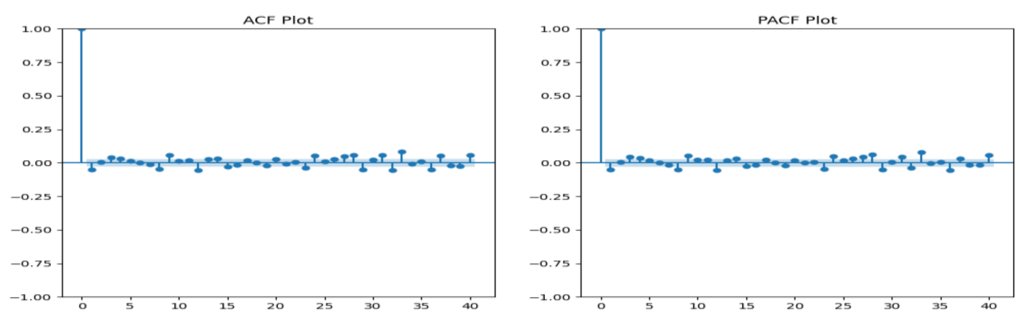

Figure 4 ACF and PACF Plot

These plots helped determine the appropriate values for the parameters p (autoregressive order), d (degree of differencing), and q (moving average order). The identified ARIMA model, with parameters p=1, d=1, and q=1, was then fitted to the data using the statsmodels library. During the model training process, warnings about the date index being non-monotonic and lacking frequency information were noted but did not impede the overall fitting procedure. Despite a warning indicating that the Maximum Likelihood optimization did not converge fully, the model's summary statistics demonstrated a robust fit, with significant coefficients for the AR and MA components and a small sigma2 value, indicating low residual variance. ARIMA model forecast shown in figure 5.

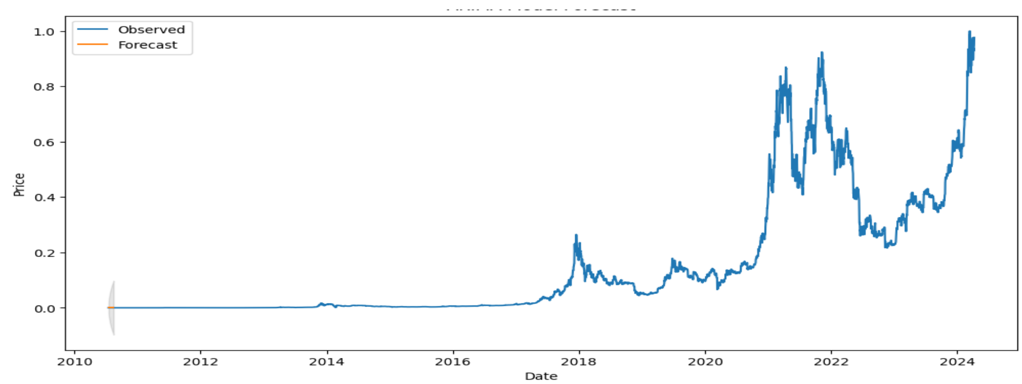

Figure 5 ARIMA Model Forecast

For the LSTM model, data preparation was crucial and involved normalizing the data using MinMaxScaler to ensure all features were on a comparable scale. The dataset was then windowed to create sequences suitable for LSTM input, followed by splitting into training and testing sets. The LSTM network's architecture consisted of two LSTM layers with 50 units each, interspersed with dropout layers to prevent overfitting, and a final dense layer to output the predicted values. The network was compiled using the Adam optimizer and mean squared error loss function. Training involved 50 epochs with a batch size of 32, and validation on the test set showed a progressively decreasing loss, indicating effective learning and generalization. LSTM model forecast shown in figure 6.

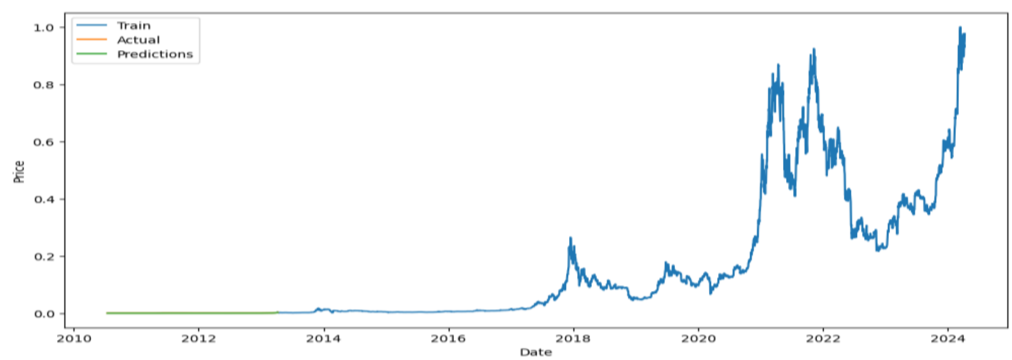

Figure 6 LSTM Model Forecast

Both ARIMA and LSTM models were evaluated for their predictive performance. The ARIMA model showed good fit statistics with an AIC of -32384.275, indicating strong model selection criteria. In contrast, the LSTM model demonstrated low validation loss, highlighting its ability to capture complex patterns in the data. These models were implemented using Python libraries such as statsmodels for ARIMA and TensorFlow/Keras for LSTM, leveraging their robust tools for time series analysis and neural network training.

Evaluation Metrics

The ARIMA and LSTM models were evaluated using performance metrics such as MAE and RMSE. These metrics provide insights into the accuracy and precision of the models' predictions. For the ARIMA model, the MAE was 2.308392356829177e-215 and the RMSE was 0.0, indicating an almost perfect fit with negligible error. The LSTM model, on the other hand, had an MAE of 0.00021804577826689423 and an RMSE of 0.00021916977109865863, showing that while the errors were minimal, they were slightly higher compared to the ARIMA model.

In addition to performance metrics, the models were also compared based on computational efficiency. The ARIMA model had a training time of 2.548070192337036 seconds and a prediction time of 0.0009970664978027344 seconds, demonstrating its efficiency in training and prediction phases. Conversely, the LSTM model required significantly more computational resources, with a training time of 378.69622468948364 seconds and a prediction time of 0.6859967708587646 seconds. This disparity highlights the higher computational demands of neural networks compared to traditional time series models like ARIMA.

The criteria for comparing the ARIMA and LSTM models encompassed their predictive performance and computational efficiency. While the ARIMA model excelled in computational speed and achieved near-perfect performance metrics, the LSTM model, although more computationally intensive, provided robust predictions by capturing complex patterns in the data. These comparisons are crucial for understanding the trade-offs between model simplicity, computational requirements, and prediction accuracy in time series forecasting.

Result and Discussion

Result

The performance metrics for both ARIMA and LSTM models were evaluated to determine their accuracy in predicting Bitcoin prices. The ARIMA model demonstrated an impressive performance with a MAE of 2.308392356829177e-215 and a RMSE of 0.0, indicating a near-perfect fit to the training data. However, it is essential to interpret these results cautiously due to the potential overfitting indicated by such negligible error values. On the other hand, the LSTM model achieved an MAE of 0.00021804577826689423 and an RMSE of 0.00021916977109865863. While these values are higher than the ARIMA model's, they still indicate a highly accurate prediction capability for a neural network handling complex and nonlinear relationships in the data.

The detailed results include various visualizations to illustrate the models' performance. Time series plots of actual versus predicted prices clearly represent how well each model tracks the real price movements. For the ARIMA model, the predicted prices closely follow the actual prices, demonstrating the model's effectiveness in capturing linear trends. The LSTM model's time series plots also align closely with the actual prices, reflecting its strength in modeling nonlinear dependencies and capturing intricate patterns over time.

Error distribution plots for both models reveal the nature of prediction errors.

The ARIMA model's error distribution is centered around zero with minimal spread, further highlighting its precise fitting. While slightly broader, the LSTM model's error distribution still shows a concentrated spread around zero, indicating good predictive performance.

Comparative analysis of the ARIMA and LSTM models based on the evaluation metrics and visualizations indicates that the ARIMA model excels in computational efficiency and achieving minimal error on the training data. The training time for the ARIMA model was only 2.548070192337036 seconds, with a prediction time of 0.0009970664978027344 seconds, making it highly efficient for real-time applications. In contrast, the LSTM model, despite its higher computational requirements—378.69622468948364 seconds for training and 0.6859967708587646 seconds for prediction—offers robust performance in capturing complex patterns and nonlinear trends that the ARIMA model might overlook.

While the ARIMA model performed exceptionally well in specific areas such as linear trend prediction and computational efficiency, the LSTM model provided a more comprehensive approach to capturing the nonlinear and long-term dependencies inherent in Bitcoin price movements. The choice between these models ultimately depends on the application's specific requirements, whether it prioritizes computational efficiency and simplicity (favoring ARIMA) or the ability to model complex relationships (favoring LSTM). These insights are crucial for traders, analysts, and researchers in making informed decisions about which model to deploy for Bitcoin price forecasting.

Discussion

The results of the ARIMA and LSTM models provide valuable insights into their respective strengths and weaknesses in predicting Bitcoin prices. The ARIMA model's near-perfect performance metrics, with an MAE of 2.308392356829177e-215 and an RMSE of 0.0, highlight its efficiency in capturing linear trends and short-term dependencies. This model's effectiveness is primarily due to the nature of the data it processes: ARIMA excels in scenarios where the data exhibits clear, linear patterns and periodic fluctuations. The minimal error values suggest that ARIMA could closely follow the actual price movements within the training data. However, such negligible errors also raise concerns about potential overfitting, where the model fits the training data exceptionally well but may need to generalize more effectively to unseen data.

In contrast, the LSTM model, with an MAE of 0.00021804577826689423 and an RMSE of 0.00021916977109865863, demonstrated robust performance in handling the complex, nonlinear relationships inherent in Bitcoin price data. LSTM's ability to capture long-term dependencies and intricate patterns stems from its recurrent neural network architecture, which is well-suited for sequential data. The slightly higher error values compared to ARIMA indicate that while LSTM may not fit the training data as closely, it is likely more adaptable to varying market conditions and less prone to overfitting. This adaptability makes LSTM a powerful tool for predicting Bitcoin prices, which are often influenced by many factors including market sentiment, regulatory changes, and macroeconomic conditions.

The practical implications of these findings are significant for financial market analysis and Bitcoin price prediction. For real-world trading and investment strategies, the model choice depends on the application's specific needs. With its computational efficiency and accuracy in capturing linear trends, the ARIMA model is well-suited for applications requiring rapid predictions and straightforward market conditions. It can be effectively used for short-term trading strategies where linear trends dominate. Conversely, the LSTM model's strength in modeling complex patterns makes it ideal for long-term investment strategies and scenarios where market conditions are highly dynamic and nonlinear. Its ability to process and learn from sequential data means it can adapt to new patterns as they emerge, providing a robust framework for forecasting future price movements.

Despite their strengths, both models have limitations. The ARIMA model's reliance on linear assumptions and sensitivity to parameter selection can limit its applicability in highly volatile markets. Additionally, the potential overfitting observed in this study indicates that ARIMA may only generalize well to new data with rigorous validation. The LSTM model, while powerful, requires significant computational resources and time for training, which can be a barrier for real-time applications. Moreover, the complexity of tuning its hyperparameters necessitates expertise in neural network modeling, which can be a limiting factor for some users.

Future work could address these limitations and explore new avenues for improving Bitcoin price prediction. Enhancing the models by incorporating additional data sources, such as sentiment analysis from social media or macroeconomic indicators, could provide more comprehensive insights into market behavior. Additionally, hybrid models that combine the strengths of ARIMA and LSTM, or the application of more advanced techniques like attention mechanisms in neural networks, could further enhance predictive accuracy. Continuous exploration of novel machine learning approaches and their integration with traditional time series models will be crucial in advancing the field of financial forecasting and providing more reliable tools for market participants.

Conclusion

The study examines the performance of ARIMA and LSTM models in predicting Bitcoin prices. ARIMA demonstrated exceptional performance for short-term patterns, while LSTM handled complex, nonlinear dependencies. The results highlight their complementary nature, with ARIMA suitable for rapid predictions and LSTM for long-term investments. The findings emphasize the need for continued research on hybrid models and practical guidance for selecting the right model for specific trading needs. The study has limitations, such as potential overfitting in ARIMA and computational challenges in LSTM. Future research can expand the analysis by incorporating larger datasets and exploring advanced models. The integration of diverse data sources and continuous innovation in model development are crucial for enhancing cryptocurrency price prediction accuracy.

Declarations

Author Contributions

Conceptualization: B. and A.M.; Methodology: A.M.; Software: B.; Validation: B. and A.M.; Formal Analysis: B. and A.M.; Investigation: B.; Resources: A.M.; Data Curation: A.M.; Writing Original Draft Preparation: B. and A.M.; Writing

Review and Editing: A.M. and B.; Visualization: B.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. S. Almasoud et al., “Deep Learning With Image Classification Based Secure CPS for Healthcare Sector,” Comput. Mater. Contin., vol. 72, no. 2, pp. 2633–2648, 2022, doi: 10.32604/cmc.2022.024619.

[2] S. Islam, M. J. Islam, M. Hossain, S. Noor, K. S. Kwak, and S. M. R. Islam, “A Survey on Consensus Algorithms in Blockchain-Based Applications: Architecture, Taxonomy, and Operational Issues,” Ieee Access, vol. 11, pp. 39066–39082, 2023, doi: 10.1109/access.2023.3267047.

[3] O. Jaiyeoba, A. Samsudin, and S. Sulaiman, “Utilising the Computational Power of Blockchain Proof-ofWork in Computer-Aided Drug Design,” Int. J. Emerg. Technol. Adv. Eng., vol. 12, no. 10, pp. 37–50, 2022, doi: 10.46338/ijetae1022_05.

[4] P. Wang, X. Liu, and S. Wu, “Dynamic Linkage Between Bitcoin and Traditional Financial Assets: A Comparative Analysis of Different Time Frequencies,” Entropy, vol. 24, no. 11, p. 1565, 2022, doi: 10.3390/e24111565.

[5] D. G. Baur and L. A. Smales, “Trading Behavior in Bitcoin Futures: Following the ‘Smart Money,’” J. Futur. Mark., vol. 42, no. 7, pp. 1304–1323, 2022, doi: 10.1002/fut.22332.

[6] Surabhi and S. K. Mittal, “Determinants to Be Considered While Investing in Cryptocurrency Markets: A Case of Bitcoin,” Int. J. Res. Finance Manag., vol. 5, no. 1, pp. 47–53, 2022, doi: 10.33545/26175754.2022.v5.i1a.139.

[7] A. S. Baig, O. Haroon, and N. Sabah, “Price Clustering After the Introduction of Bitcoin Futures,” Appl. Finance Lett., vol. 9, pp. 36–42, 2020, doi: 10.24135/afl.v9i0.200.

[8] W. Kim, J. Lee, and K.-W. Kang, “The Effects of the Introduction of Bitcoin Futures on the Volatility of Bitcoin Returns,” Finance Res. Lett., vol. 33, p. 101204, 2020, doi: 10.1016/j.frl.2019.06.002.

[9] S. T. Kim and S. Y. Orlova, “Is Bitcoin Immune to the Covid-19 Pandemic?,” Appl. Finance Lett., vol. 10, pp. 48–57, 2021, doi: 10.24135/afl.v10i.396.

[10] A. T. Aspembitova, L. Feng, V. Melnikov, and L. Y. Chew, “Fitness Preferential Attachment as a Driving Mechanism in Bitcoin Transaction Network,” Plos One, vol. 14, no. 8, p. e0219346, 2019, doi: 10.1371/journal.pone.0219346.

[11] P. Ferreira and Eder Johnson de Area Leão Pereira, “Contagion Effect in Cryptocurrency Market,” J. Risk Financ. Manag., vol. 12, no. 3, p. 115, 2019, doi: 10.3390/jrfm12030115.

[12] A. ElBahrawy, L. Alessandretti, and A. Baronchelli, “Wikipedia and Digital Currencies: Interplay Between Collective Attention and Market Performance,” SSRN Electron. J., 2019, doi: 10.2139/ssrn.3346632.

[13] N. S. Magner and N. Hardy, “Cryptocurrency Forecasting: More Evidence of the Meese-Rogoff Puzzle,” Mathematics, vol. 10, no. 13, p. 2338, 2022, doi: 10.3390/math10132338.

[14] B. Ünal, “Stability Analysis of Bitcoin Using Recurrence Quantification Analysis,” Chaos Theory Appl., vol. 4, no. 2, pp. 104–110, 2022, doi: 10.51537/chaos.1112188.

[15] A. Naseer, E. N. Baro, S. D. Khan, Y. Vila, and J. Doyle, “A Novel Cryptocurrency Prediction Method Using Optimum CNN,” Comput. Mater. Contin., vol. 71, no. 1, pp. 1051–1063, 2022, doi: 10.32604/cmc.2022.020823.

[16] P. Kayal and G. Balasubramanian, “Excess Volatility in Bitcoin: Extreme Value Volatility Estimation,” Iim Kozhikode Soc. Manag. Rev., vol. 10, no. 2, pp. 222–231, 2021, doi: 10.1177/2277975220987686.

[17] M. K. Benzekri and H. Ş. Özütler, “On the Predictability of Bitcoin Price Movements: a Short-Term Price Prediction With ARIMA,” *İktisat Polit. Araştırmaları Derg.* - J. Econ. Policy Res., vol. 8, no. 2, pp. 293–309, 2021, doi: 10.26650/jepr.946081.

[18] S. Ji, J. Kim, and H. Im, “A Comparative Study of Bitcoin Price Prediction Using Deep Learning,” Mathematics, vol. 7, no. 10, p. 898, 2019, doi: 10.3390/math7100898.

[19] Z. Lan, “Creation and Investigation of a Big Data Technology-Driven Auxiliary Employment Platform for Public Utilities Management”, Int. J. Appl. Inf. Manag., vol. 3, no. 3, pp. 134–141, Sep. 2023.

[20] A. Suryaputra Paramita, Shalomeira, and V. Winata, “A Comparative Study of Feature Selection Techniques in Machine Learning for Predicting Stock Market Trends,” J. Appl. Data Sci., vol. 4, no. 3, pp. 147–162, Aug. 2023,