INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN ACCOUNTING, FINANCE AND MANAGEMENT SCIENCES

OPEN ACCESS JOURNAL

Vol 15, Issue 2, (2025) E-ISSN: 2225-8329

Beyond Conventional Methods: Advancing Ethereum Price Prediction through Integrated Technical, On-Chain, and Machine Learning Approaches

Dalia Elbanna, Ema Izati Zull Kepili[1], Nik Hadiyan Nik Azman

School of Management Universiti Sains Malaysia Penang, Malaysia

Corresponding Authors Email: emazull@usm.my

To Link this Article: http://dx.doi.org/10.6007/IJARAFMS/v15-i2/24968 DOI:10.6007/IJARAFMS/v15-i2/24968

Published Online: 07 April 2025

Abstract

Ethereum's anonymity and uncontrolled cryptocurrency attraction have attracted investors. Ethereum's price dynamic inspired this study's prediction analyses. Previous study has focused on either technical analysis or on-chain analysis, leaving investors without the synergistic effects of integrating the two. This study addresses missed insights and lack of cross-comparisons by identifying variable relationships and dependencies and comparing a classical model (ARIMA), a supervised deep learning model (LSTM), and an ensemble machine learning model (XGBoost) in Ethereum price prediction. The dependent variable is Ethereum price and the independent variables are opening, high, low, closing, adjusted closing, volume traded, market capitalization, cumulative return, transactions, blocks, and gas utilized. Prices and market capitalization, traded volume, and volume are strongly correlated, and the LSTM model is the most promising due to its greater prediction accuracy and generality. The analysis reveals the bitcoin market's complexity, affecting investing and risk management.

Keywords: Cryptocurrency, Ethereum, ARIMA, XGBOOST, LSTM, Technical analysis, On-chain Analysis.

Introduction

The cryptocurrency sector has grown significantly since the creation of Bitcoin in 2009. It evolves from a small technological experiment into a global financial ecosystem. The expansion also includes the increasing development of other cryptocurrencies like Ethereum, LiteCoin, and others. Cryptocurrency is evolving into a multifaceted tool and cases, serving not just as a digital currency but also as an investment vehicle, a method for remittances, and a means of payment. Among these digital assets has evolved from being a form of digital currency to having its own decentralized applications (Dapps). The market capitalization of cryptocurrencies has increased dramatically due to numerous use cases. The rise of their use and recognition as digital assets has led to different levels of price volatility. Because of the

[1] Corresponding author's email: emazull@usm.my

market's inherent volatility and unpredictability, it is becoming more difficult to predict cryptocurrency prices. This increase has sparked a rising interest in precise and dependable methods for predicting cryptocurrency prices (Jagannath et al., 2021). Not only that, these developments highlight the urgent need for accurate and robust price prediction models, especially as cryptocurrencies become more integrated into mainstream financial portfolios and investment strategies.

Studies pertaining to price forecasting and prediction have progressed from employing fundamental analysis to technical analysis, on-chain analysis, regression analysis, supervised machine learning, and, at present, deep learning techniques. Predicting cryptocurrency prices, particularly Ethereum, is crucial for a wide range of stakeholders—retail and institutional investors, trading platforms, regulators, and financial analysts. For investors, accurate forecasting can reduce risk and enhance returns. For policymakers and institutions, understanding price behavior can inform regulation and financial stability measures. However, the task is complex. Technical analysis, which use historical price and volume data to recognize patterns and trends, have been extensively used in the financial industry for many years. Yet, its use in cryptocurrency markets has produced varied outcomes. The cryptocurrency market's high volatility and little long-term historical data make standard technical analysis methods less reliable (Akgül et al., 2022). This is due to many reasons such as different interpretation of chart patterns, unreliable historical data to predict future performance, inability to consider fundamental factors and lagging nature of time signal.

A possible alternative to technical analysis is on-chain analysis. On-chain analysis involves examining the data stored on the blockchain itself to gain insights into various aspects of cryptocurrency transactions and network activities such as transaction volume, active addresses, and mining activity to understand the overall health and activity of the cryptocurrency network (Akgül et al., 2022). Through analyzing these on-chain metrics, investors and analysts can possibly extract useful insights about the cryptocurrency's fundamental worth and future price trends. Integrating technical analysis and on-chain analysis into a full prediction model for bitcoin prices is a challenging task. Technical indicators and on-chain data present a challenging obstacle for machine learning algorithms due to their complex relationship.

Multiple regression equation too can be used to create operational analytical programs capable of estimating relationship and real-time forecasting the price movement of ethereum. Akbulaev et al (2020) analyzed the correlation between the prices of Bitcoin and Ethereum, illustrating their strong interdependence through the presentation of a mathematical model. Alahmari (2020) predicting the price of cryptocurrency using support vector regression methods. The use of machine learning in price prediction started to grow with Spilak (2018) proposed a Neural Network framework for price prediction in cryptocurrency markets, utilizing Multilayer Perceptron (MLP), Recurrent Neural Network (RNN), and Long Short-Term Memory (LSTM) models. It compares supervised learning methods for classification tasks and suggests trading strategies based on predictions. Rizwan et al (2019) predicted Bitcoin's USD price using deep learning methods and suggested Bayesian RNN and LSTM networks achieve 52% accuracy and 8% RMSE, outperforming classic regression technique - ARIMA for time series prediction. Demir et. Al (2019) employed machine learning techniques to estimate the price of bitcoin using the Kaggle Bitcoin Dataset

2010-2019. Long-short term memory networks, support vector machines, artificial neural networks, Naive Bayes, decision trees, and the closest neighbor algorithm are among the techniques utilized. The respective obtained accuracy rates are as follows: 91.8%, 86.6%, 85%, and 81.2% - results which are better than classic techniques.

Year 2020 and beyond saw increasing numbers of research on cryptocurrency price prediction using machine learning and deep learning techniques, such as Li et al (2020) who uses attentive LSTM and embedding network, Singh et al (2021), Tanwar et al (2021) and Wang and Yan (2022) who used deep learning approach. In year 2023, the trend moves towards more sophisticated tools within supervised machine learning such as employing XGBoost, Prophet and sentiment analysis were performed on bitcoin data (Ramani et al, 2023) and Time Series Forecasting of Ethereum Price using FB-Prophet (Yuvarani et al, 2023).

The evolution of price prediction analysis using machine learning and deep learning signifies 3 main importance: Firstly, it demonstrates the adaptability of machine learning algorithms in capturing complex patterns within cryptocurrency markets, as evidenced by the utilization of various models including Multilayer Perceptron (MLP), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Bayesian RNN, and LSTM networks. Secondly, it highlights the continuous pursuit of higher accuracy and efficiency in forecasting, as illustrated by the improvement in accuracy rates over time. Thirdly, it emphasizes the diversification of methodologies and tools employed in price prediction research, as indicated by the shift towards utilizing sophisticated techniques like XGBoost, Prophet, The Transformer and sentiment analysis, demonstrating the field's dynamic nature in embracing new advancements and innovations.

In two ways, according to the abovementioned three factors and prior research, the prediction study is deficient. First, prior research that employed technical analysis neglected to include potentially valuable data that is absent because of the segregated approach (missed insights). Previous research has mostly focused on either technical or on-chain indicators in isolation, overlooking the synergistic potential of combining both. Furthermore, studies often compare models within the same methodological family—such as different types of deep learning or machine learning models—limiting a holistic evaluation of which approaches truly offer the best predictive performance. This fragmented approach leaves a critical gap in understanding which methods are most effective in dynamic cryptocurrency environments.

Second, comparison analyses have been confined to its "family" in prior machine learning research (cross-model comparison lackness). Further elaborating on the first issue, the oversight occurred because of prior investigations into predicting cryptocurrency prices exclusively centering on on-chain analysis or technical analysis, thereby depriving investors of a comprehensive understanding of the market. The isolation of this approach prevents it from recognizing the beneficial outcomes that can result from the integration of technical and on-chain analysis. Regarding the second issue, it is necessary to conduct the underutilized "cross-family" analysis, as additional outcomes can ascertain the resilience of different forms of analysis on cryptocurrencies. Even though Kim et al. (2021), Politis et al. (2021), and Hamayel and Owda (2021) have conducted research on Ethereum prices, additional analyses are recommended in order to strengthen the reliability of forecasts. This is because evaluations

of the performance of numerous machine learning models on cryptocurrencies have predominantly concentrated on comparisons within the same "family" of techniques. For instance, they might conduct a comparative analysis of established models such as auto-regressor integrated moving average (ARIMA) and other conventional techniques. Alternatively, they might evaluate more recent and sophisticated models like support vector machine (SVM) and extreme gradient boosting (XGBoost) in relation to one another. In the same way, they might conduct a comparison of various deep learning models, including Prophet and long-short term memory (LSTM).

Nevertheless, a crucial aspect that is frequently overlooked is the comparison of models belonging to distinct "families" or categories. An illustration of this could be the comparison between a supervised model under deep learning (LSTM) and a traditional model (ARIMA) or an ensemble learning under machine learning method (XGBoost) and a deep learning method (Deep AR) in the context of cryptocurrency price prediction, with a specific focus on Ethereum. Although cross-family comparisons offer vital insights into the most effective model types for predicting cryptocurrency prices, they have received less attention from researchers compared to comparisons conducted within the same family.

This study aims to address these gaps (of missed insight & cross-comparison lackness) by i) identifying relationships and dependencies among variable and ii) comparing a classical model (ARIMA) with supervised model under deep learning (LSTM) and an ensemble learning under machine learning method (XGBoost) in predicting Ethereum price. The findings of this study will contribute to the advancement of cryptocurrency price prediction methodologies and provide valuable insights for investors, traders, and financial institutions seeking to navigate the dynamic and unpredictable cryptocurrency landscape.

This study focuses on the Ethereum daily price data from 2018 to 2023 as a time-series data and then extract the relevant blocks data directly from Ethereum network through a virtual node. The research evaluates the most important variables of both technical and on-chain analysis and their correlations with the Ethereum price, then apply a comparative analysis utilizing combined variables from technical and on-chain analysis. The processing of transactions data is considered out of scope of this study due to the huge number of transactions on the chain in the presence of hardware processing limitations. However, it’s recommended the full inclusion of all the on-chain variables in the future work.

Literature Review

The increasing variety of cryptocurrencies has caused their prices to fluctuate with varying degrees of volatility. Bitcoin, the pioneer cryptocurrency, functions as a blockchain-powered peer-to-peer electronic payment system and investment instrument. Ethereum, the second most valuable cryptocurrency in terms of market capitalization, provides a platform for decentralized applications and smart contracts, in addition to its primary functions. Investors are intrigued by this evolution of Ethereum owing to its multifaceted utility, which contributes to its unique volatility in contrast to conventional cryptocurrencies. Past studies explored the complex factors that impact the valuations of cryptocurrencies, in line with the increasing attention and importance attributed to these digital assets.

There are numerous approaches to analyzing cryptocurrency assets, the most common of which are fundamental and technical analysis. fundamental analysis looks for disparities between a stock's market price and its intrinsic value. It examines the "fundamentals" of a company's financial reporting, as well as macroeconomic data and variables. Fundamental analysis presupposes a time lag between fundamental driver fluctuates and stock price movements. The method determines a share's true value and looks for opportunities where it differs from the market price (Petrusheva & Jordanoski, 2016). Variables utilized in crypto analysis include market size, volume, tokenomics, total value locked, and whitepaper.

In contrast, technical analysis predicts a share's future market value by statistically analyzing its historical price behavior. It uses the share's price history to forecast future volatility. Technical analysis posits that price patterns are intrinsically linked to elements such as financial statement statistics, and that examining these links helps investors understand how prices react to financial changes. Technical analysts analyze past price behavior to predict future share price changes (Petrusheva & Jordanoski, 2016).

Because of the limits of fundamental analysis in the cryptocurrency asset class, technical analysis is frequently used. Tokens reflect app values rather than firm finances, therefore normal accounting measurements and indicators cannot be created. Prices also shift regardless of genuine economic conditions (Tans and Sosnoff, 2018). During crypto's early years, this evolved as the primary method for evaluating market developments and trying valuations, with fewer traditional analytical anchors than regular assets (Smith & Johnson, 2020). When fundamentals couldn't be applied to the new cryptocurrency, technical analysis stepped in. It provides a framework for assessing market variations that were not reflected by other common valuation methodologies (Lee & Wong, 2019). Nonetheless, the quick volatility of indicators in the cryptocurrency sphere make it difficult to estimate the prices of crypto assets using technical analysis. The rapid pace of market activity may make it difficult for technical tools to identify trends before they significantly reverse, complicating predictions in comparison to more stable traditional markets with successful track records of technical tools over long periods of time (Jagannath et al., 2021).

While technical analysis involves studying historical price movements and volume data to identify patterns and trends that may indicate future price movements, on-chain analysis focuses on extracting insights from blockchain data, such as transaction volume, network activity, and token circulation. This approach provides a deeper understanding of market dynamics by examining real-time data directly from the blockchain. On-chain analysis can reveal patterns of investor behaviour, identify trends in network usage, and detect potential market manipulation or anomalies. On chain analysis gives us public access to the real-time health of a financial system. On-chain metrics derive data from a blockchain network's inherent information regarding aspects like size, blocks number, transactional volume and mining difficulty. They communicate the network's state to interested parties, maintaining the transparency, immutability and decentralization of the underlying technology. Each time-series metric offers historical activity insights. This knowledge benefits law enforcement monitoring illicit behaviour and financial professionals assessing investments. On-chain data aids authorities and the industry by exposing trends and usage over time through blockchain's transparent, immutable nature. It helps determine viability and track suspect transactions (Jagannath et al., 2021). Integrating on-chain data with machine learning provides a strong

technique to collecting insights, forecasting outcomes, and optimizing decision-making processes in the cryptocurrency industry. Machine learning techniques capitalize on blockchain data's inherent transparency and immutability.

Classical models such as Autoregressive Integrated Moving Average (ARIMA), Seasonal Decomposition of Time Series (STL), Exponential Smoothing (ETS), Holt-Winters Method, and others are now used together with machine learning models. The classical time series models can provide interpretable baseline forecasts and combined with machine learning, models can capture complex nonlinear relationships in the data. (Korstanje, 2023).

The ARIMA model is commonly used for time series forecasting due to its ability to capture both autoregressive and moving average components, making it suitable for analyzing data with trends and seasonality. Salam, Alazzam, and Asiri (2019) assessed the effectiveness of the ARIMA model in predicting Ethereum's value, particularly during periods of economic instability like the COVID-19 pandemic. The research analyzes weekly Ethereum value data from January 2017 to December 2020, totaling 208 samples. The findings indicate that the ARIMA model performed poorly in predicting Ethereum's value, with forecasted values significantly deviating from actual values. The Mean Absolute Percentage Error (MAPE) test revealed an accuracy rate of 51.94%. The study attributes this poor performance to economic uncertainty caused by the COVID-19 pandemic and the rise of decentralized finance in early 2021, which led to substantial increases in Ethereum's value and increased forecasting errors. The study suggests that future research explore more advanced models, such as the Autoregressive Fractionally Integrated Moving Average (AFRIMA), to improve forecast accuracy.

Liantoni and Agusti (2020) utilizes double exponential smoothing to predict Bitcoin prices, focusing on minimizing the mean absolute percentage error (MAPE). The dataset comprises Bitcoin prices from 2017 to 2019, sourced from www.cryptocompare.com. Various alpha (α) parameters, ranging from 0.1 to 0.9, are tested to determine the best fit for price forecasting. Results show that the double exponential smoothing method yields the smallest MAPE value (2.89%) when α is set to 0.9. Predictions for Bitcoin's price on January 1, 2020, are generated, with an error rate of 0.0373%. These findings suggest that the developed system can serve as a valuable decision support tool for Bitcoin trading.

The evolution of forecasting techniques has shifted from classical methods towards a fusion with advanced machine learning algorithms. For instance, forecasting study has utilized deep learning models, a subset of machine learning like Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTMs), and Bi-directional Long Short-Term Memory (Bi-LSTMs). Wang, Yao, and Zou (2020) compared the effectiveness of these three models—in predicting short-term (30 days) and long-term (90 days) Ethereum prices. The dataset comprises closing prices from the past 2000 days, sourced from an API in JSON format and updated daily. Their findings concludes that bidirectional LSTM outperforms other models, including RNN and traditional LSTM, in forecasting Ethereum prices. Using the closing price as the key parameter for prediction, the model proves valuable for understanding price trends. It demonstrates scalability and potential for further accuracy improvements through adjustments. While RNN struggles with price prediction, both LSTM and bidirectional LSTM excel, with the latter being the preferred choice. Bidirectional LSTM effectively forecasts price trends with reasonable

accuracy, paving the way for potential enhancements by incorporating additional parameters and optimizing hyperparameters.

Classical technique was also used with artificial neural network such as Multilayer Perceptron (MLP) and Extreme Learning Machine (ELM). Waheeb, Shah, Jabreel, and Puig (2020) compares statistical and machine learning methods for predicting Bitcoin's closing prices. Thirteen forecasting techniques are tested, including simple ones like averages and more complex ones like ARIMA and MLP. These methods forecast Bitcoin's closing prices for the next 14 days. The study reveals three main findings. First, seven methods, including MLP and ELM, outperformed the simple naive method. Second, MLP and ELM demonstrated superior accuracy on both validation and out-of-sample data compared to other methods. Third, the amount of training data significantly impacts the effectiveness of forecasting methods.

Cryptocurrency price prediction has also utilized supervised machine learning models. The models are trained on labeled data, which means that the data has been tagged with the correct answer. This allows the model to learn the relationship between the input features and the output target. Unsupervised machine learning models, on the other hand, are trained on unlabeled data, which means that the data does not have any tags. Supervised machine learning models can be used for time series forecasting by converting the time series data into a supervised learning problem. This can be done by creating a feature vector for each time step, where the feature vector includes the value of the time series at the current time step and the values of the time series at previous time steps. The target variable for the supervised learning problem is the value of the time series at the next time step (Korstanje, 2023).

Kedyr et al (2021) employed both traditional statistical methods and machine learning techniques like Bayesian regression, support vector machines, and neural networks. Raju and Tarif (2020) applied sentiment analysis and supervised machine learning to Tweets (now X) and Reddit posts, analyzing their correlation with Bitcoin price movements. Several supervised learning algorithms are explored to develop a prediction model, providing insights into future market prices. While traditional time series (ARIMA) models face challenges in producing accurate forecasts, Recurrent Neural Networks (RNN) with long short-term memory cells (LSTM) offer improved efficiency. The study compares LSTM's predictability with sentiment analysis of Bitcoin tweets to the standard ARIMA method.

Among the vast array of machine learning models, Random Forests and XGBoost stand out as true gems in the supervised learning. It’s a cutting-edge machine learning model that redefines the landscape of supervised learning, joining the ranks of Random Forests as a true classic(Korstanje, 2023). Drahokoupil (2022) used the XGBoost machine learning algorithm to forecast Bitcoin (BTC) price changes and create an algorithmic trading strategy. Six XGBoost models estimate BTC closing prices for 1, 2, 5, 10, 20, and 30 days. Bayesian optimization is used twice during strategy development: to pick optimal hyperparameters for the XGBoost models and to optimize each model's prediction weight to maximize trading strategy profitability. Despite its shortcomings, the XGBoost model can accurately anticipate BTC price changes over time. The paper examines algorithmic trading during the COVID-19 timeframe, when BTC prices fluctuated greatly. The trading method outperforms the Buy and Hold (B&H) strategy in total profit, Sharpe ratio, and Sortino ratio.

Mahdi (2021) presented a new method to anticipate whether the gold price will be in the first, second, third, or any quantile the next day, unlike cryptocurrency returns by using the support vector machine (SVM) technique to estimate financial returns for six major digital currencies from the top 10 cryptocurrencies based on sensor data - Binance Coin, Bitcoin, Cardano, Dogecoin, Ethereum, Ripple. Before and during COVID-19 are studied. The study proposed the use of a database sensor to update data analysis. The findings suggested the SVM can create profitable trading strategies and offer accurate findings before and throughout the pandemic.

While researchers typically extract relevant features from datasets strongly correlated with bitcoin prices and randomly select data segments for model training and testing, this random selection approach may yield inappropriate results and reduce prediction accuracy. Ali and Shatabda (2020) addresseed this issue by proposing a method for proper data selection to train prediction models. They applied their methodology to train a simple linear regression prediction algorithm and forecast bitcoin prices for 7 days. When the linear regression model is trained with appropriately selected data chunks, the authors observe acceptable prediction results, achieving a 96.97% accuracy rate according to the percentage error method. The manuscript concludes by discussing potential future improvements to their work.

Research Methodology

This section illustrates the process and the methodology that was applied for this project. We used Python as the primary tool for extracting, preprocessing, applying machine learning, and presenting results. Infura was used as the primary endpoint provider to access the Ethereum blockchain, enabling reliable API access, transaction management, and data retrieval without the need to run our own nodes.

Data Collection and Preprocessing

This study employed a two-pronged approach to data collection, utilizing both technical and on-chain variables to predict Ethereum price movements. Technical variables, representing traditional buy and sell signals, were sourced from Kaggle. The dataset used, "Ethereum Price USD (2018-2023)" – which comprises independent variables was validated by the Kaggle community and pre-cleaned, ensuring data quality.

On-chain variables, providing real-time insights into the Ethereum network's health, were extracted directly from the network using Python and the Infura node provider. Due to hardware limitations and the vast volume of transactions (over 2 billion), the study focused on three key on-chain variables: total transactions, total blocks, and total gas used. These variables were processed from blocks (around 20 million) and merged with the technical variables' dataset daily. This combined dataset, encompassing both technical and on-chain indicators, formed the foundation for the predictive models employed in the study. On the data quality and descriptive statistics, this study involved the retrieval and loading of two datasets, with a total of 2038 days spanning six years from 2018 to 2023. The dataset was found to be cleaned in terms of nulls and duplicates, with a count of 64752 duplicate rows, representing a small proportion compared to the total of 17 million records. However, null values accounted for 40.2% of the dataset.

Basic cleaning procedures were applied to address nulls and duplicates, with identical duplicate rows removed, and columns with nulls exceeding 50% were dropped. The merging technique utilized a "Chunks Approach" to handle large datasets efficiently. The merging process involved calculating the total transactions, blocks, and gas used from the blocks dataset and integrating them into the price dataset.

After merging, imputation techniques were applied to address null values in the dataset. The study imputed new columns based on random values within the accurate range for each column, as null values did not exceed 10% of the merged dataset. Market cap and cumulative return variables were added as calculated columns to provide insights into the cryptocurrency's total value and investment performance. Market cap provides an estimate of the size and relative value of a cryptocurrency within the market. It is widely used to compare the value of different cryptocurrencies and understand their overall significance in the market. Cryptocurrencies with higher market caps are generally considered more established and influential within the market.

Positive cumulative returns indicate a profit, while negative cumulative returns indicate a loss. Cumulative return helps investors assess the profitability or performance of a cryptocurrency investment over time (Kim et al., 2021).
The final dataset output included the cleaned and processed data, ready for further analysis and modeling.
This study applies three models for time-series forecasting and regression analysis: ARIMA, LSTM and XGBoost. Before performing these analyses, unit root test is performed using Dickey-Fuller and KPSS tests to ensure all data are stationary.

ARIMA (AutoRegressive Integrated Moving Average)
An ARIMA model is a statistical model used for analyzing and forecasting time series data. ARIMA extends the simpler AutoRegressive Moving Average model by incorporating the concept of integration.

The are three key components of the model, i) AR (AutoRegression): Utilizes the dependent relationship between an observation and a specified number of lagged observations, ii) I (Integrated): Involves differencing the raw observations (subtracting an observation from its previous value) to make the time series stationary, iii) MA (Moving Average): Models the dependency between an observation and residual errors from a moving average model applied to lagged observations. Each component is specified in the model using parameters. The standard notation ARIMA(*p,d,q*) is used, where the parameters are integers that define the specific ARIMA model being applied. The parameters of the ARIMA model are: *p* (lag order): The number of lagged observations included in the model, *d* (degree of differencing): The number of times the raw observations are differenced and *q* (order of moving average): The size of the moving average window.
The ARIMA(*p, q, d*) can be represented as:

$$\Delta d\ y(t) = c + \sum p\ j{=}1\ \alpha j \times y(t-j) + \epsilon(t) + \sum q\ j{=}1\ \beta j \times \epsilon(t-j) \quad (1)$$

Where Δ = (1 − B), B is the 'Backward' operator and $By(t) = y(t-1)$, y(t) is the observation data at time t, c is the constant, $\alpha 1$, …, αp are the auto-regressive parameters, ϵ(t) is the white noise at time t, and $\beta 1$, …, βq are the moving average coefficients.

The determination of the order *q* and *p* of the ARIMA model can be achieved by employing the autocorrelation function (ACF) and the partial autocorrelation function (PACF) of the data (Box and Jenkins, 1976). In addition, several alternative approaches have been suggested to determine the ARIMA order, including those based on MDL (minimum description length), AIC and BIC (Aho et al., 2014), fuzzy systems, or AIC (Akaikes information criterion) (Shibata, 1976; Haseyama and Kitajima, 2001).

LSTM (Long Short-Term Memory)

The LSTM (Long Short-Term Memory) model is a type of neural network designed for handling sequential data with long-term dependencies. It excels at capturing and remembering information over long periods, making it ideal for tasks like time-series forecasting. (Y. Chen & Ng, 2019). It requires feature scaling to ensure sensitivity to variate values, and the Rectified Linear Unit (ReLU) activation function is applied due to sparse data and vanishing gradients. Starting with a single layer is recommended (S. Chen, 2022). The data is reshaped to fit the LSTM model as a series. The same measurement criteria for errors are applied to evaluate the model. Feature scaling applied using scikit-learn library with the min-max scaler to maintain values between 0 and 1.

The choice of an appropriate loss function is crucial for effective training. Mean Squared Error (MSE) is commonly used for regression tasks, such as predicting continuous values like stock prices or sensor data. It assesses and minimizes prediction errors, focusing on larger deviations that significantly impact training. However, MSE is optimal for normally distributed errors and may not be ideal for non-normally distributed data. Alternative loss functions, such as Categorical Cross-Entropy, are more suitable for tasks involving discrete categories. Sensitivity to outliers can disrupt training, so careful analysis of data's characteristics and objectives is necessary to select the optimal loss function (Ferenczi & Bădică, 2023).

The Adam optimizer is valuable for adaptive optimization, particularly in the presence of noisy or sparse data. It dynamically adjusts learning rates for each weight in the LSTM network, addressing vanishing gradients and enabling the unrestricted flow of information during training. Adam is computationally efficient, user-friendly, and excels in handling large datasets and non-stationary problems. However, it has limitations and alternative optimizers may be more suitable for highly structured data or specific convergence requirements (Chen, 2022).

XGBoost (Extreme Gradient Boosting)

While XGBoost is not as concerned with stationarity as LSTM, it does necessitate feature scaling and, similar to the majority of supervised models, is susceptible to variations in value. The exact separation of the target and predictor columns, feature scaling, and training of 80% of the data were implemented, just as in the LSTM model. XGBoost is a highly effective machine learning algorithm that is implemented in tasks involving classification and regression. It addresses intricacies such as multifaceted characteristics and non-linear associations. XGBoost demonstrates exceptional performance in the domain of price prediction owing to its extensive capabilities in managing complex data, regularization, and feature importance (Nayam, 2022).

Measurements Criteria

In this study three error measures were used on the testing data to compare the performance of the three models:

a. RMSE (Root Mean Squared Error). It focuses on large errors - Squares errors before averaging, amplifying the impact of significant deviations. This helps prioritize accurate predictions for critical price points.
b. MAE (Mean Absolute Error). It is robust to outliers and unaffected by extreme values unlike RMSE, providing a more stable measure of typical error for skewed or noisy data. It also represents the average absolute difference between predictions and actual prices.
c. R^2 (Coefficient of Determination). It captures the proportion of price variance explained by the model, indicating its overall fit to the data. Values closer to 1 represent better model fit, providing a high-level overview of accuracy.

Findings and Discussion

The period from 2020 to September 2022 witnessed a significant surge in crypto prices attributed to various factors such as loose money policies, institutional adoption, retail FOMO, Bitcoin's scarcity, and technological advancements. This led to a period of hype and investment, pushing prices to record highs. Major financial institutions like PayPal and Tesla began accepting and investing in cryptocurrencies, lending legitimacy and boosting confidence in the market.(Nayomi, n.d.)

However, the tide turned in late 2022 due to rising interest rates, the UST collapse, major company insolvencies, and increased regulations. Investor confidence plummeted, triggering a mass sell-off and sending prices crashing down by over 70%. The future of crypto remains uncertain, but understanding these factors provides valuable context for navigating this volatile market. (Nayomi, n.d.) This volatility underscores the uncertainty surrounding the future of cryptocurrencies, emphasizing the importance of understanding market dynamics for informed decision-making. Figure 1 shows the closing price bar chart, offering insights into Ethereum's price trends and stationary status.

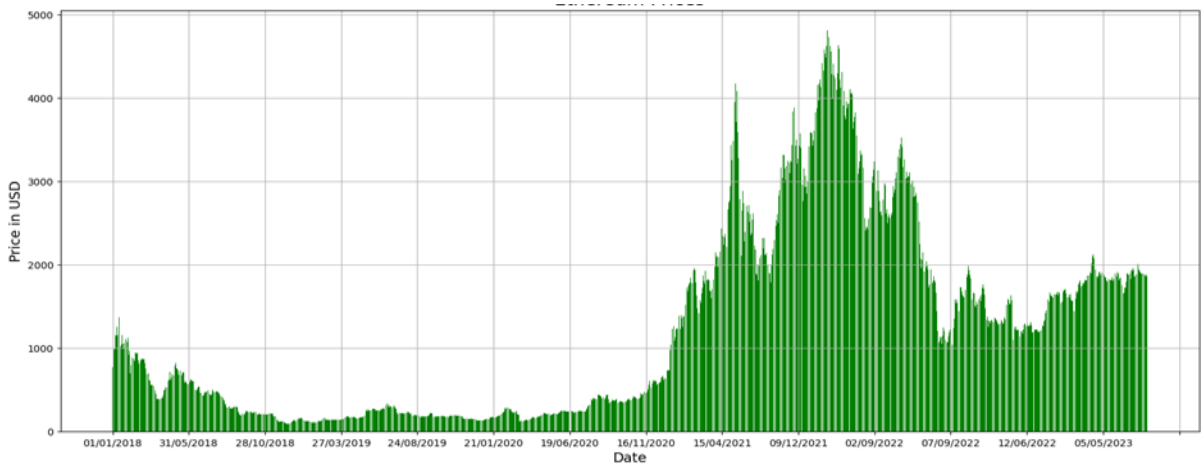

Figure 1. Closing price bar chart

While the data appears non-stationary visually, further tests like the KPSS test can provide quantitative confirmation of stationarity. Additionally, the box plots of close prices by year reveal outliers in specific years, with 2020, 2023 showing more outliers (Figure 2).

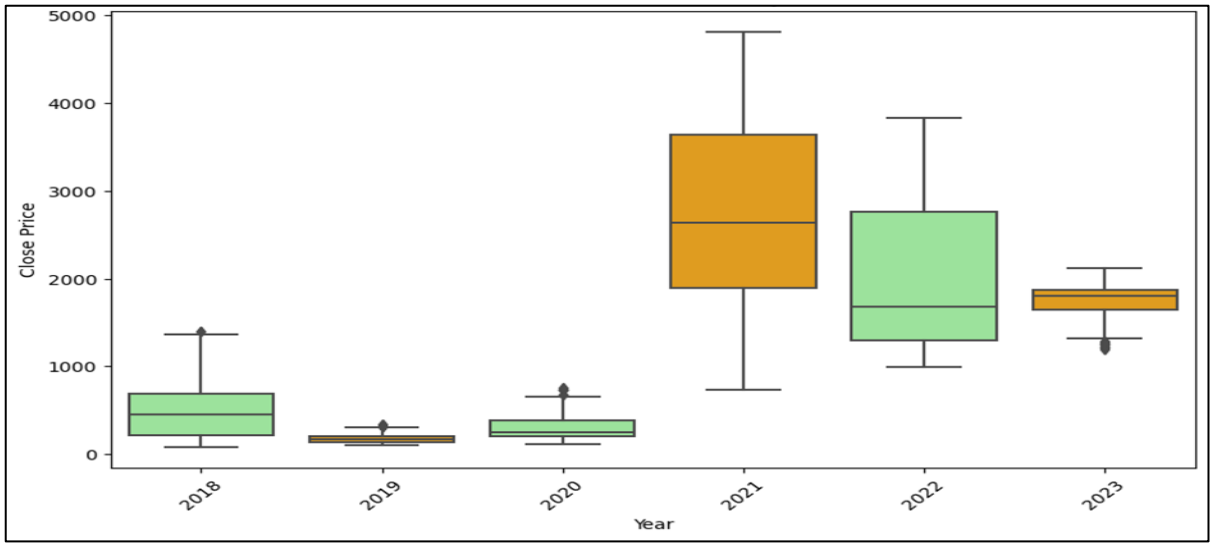

Figure 1. Closing price bar chart

The market capitalization line plot illustrates the fluctuation in market cap over time (Figure 3), with notable peaks and declines reflecting market corrections and crashes. On-chain predictors like total transactions and blocks line plots demonstrate cyclical patterns like closing price movements.

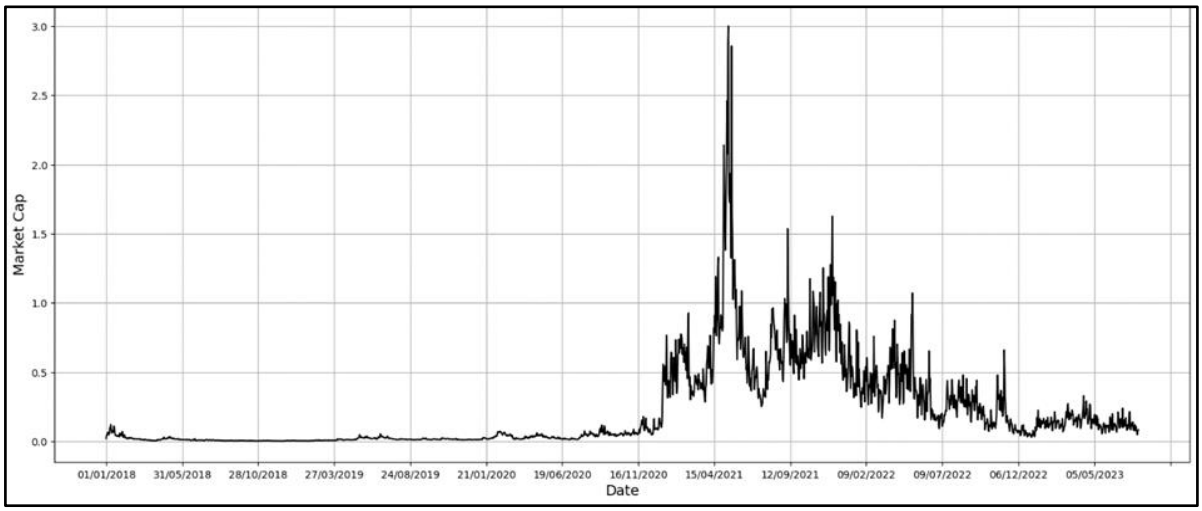

Figure 3. Market Capitalization

To identify the relationships and dependencies among variables (first objective), we performed a correlation analysis (Figure 4). There are several interesting findings from the results. Strong relationship (rho 0.75 and above) can be found in prices–market capitalization, price-volume traded and market cap-volume. Moderate relationships (rho 0.40 to 0.74) can be found in prices-volume, volume-total transaction. Low relationship (below 0.40) can be found in prices-total transaction, prices-total block number, volume traded-total block number. Nearly no correlation was found in relationship involving cumulative return.

Figure 4. Correlation matrix

	Open	High	Low	Close	Adj Close	Volume	Total Txns	Total Gas Used	Total Block Number	Market Cap	Cumulative	Year
Open	1.000	0.999	0.998	0.998	0.998	0.494	0.233	0.060	0.232	0.806	0.009	0.627
High		1.000	0.998	0.999	0.999	0.505	0.236	0.065	0.228	0.814	0.010	0.623
Low			1.000	0.999	0.999	0.473	0.227	0.062	0.235	0.787	0.010	0.633
Close				1.000	1.000	0.491	0.233	0.066	0.231	0.800	0.010	0.627
Adj Close					1.000	0.491	0.233	0.066	0.231	0.800	0.010	0.627
Volume						1.000	0.447	0.059	0.138	0.811	0.023	0.400
Total Txns							1.000	-0.028	0.301	0.309	0.011	0.295
Total Gas Used								1.000	-0.317	0.090	0.001	0.025
Total Block Number									1.000	0.085	0.014	0.491
Market Cap										1.000	0.014	0.406
Cumulative Return											1.000	0.033
Year												1.000

Stationarity Results

Stationarity refers to the constancy of statistical properties in time series data. It is crucial for time series forecasting models as these models assume a stable pattern or behavior. Differencing techniques, such as first differencing and log-differencing, are commonly used to transform non-stationary data into stationary form for ARIMA (Van Greunen et al., 2014). Dicket-Fuller tests detect the presence of a unit root, indicating long-term memory and non-stationarity. A low-test statistic and p-value support rejecting the null hypothesis of non-stationarity. (Heymans et al., 2014). The study shows that our DF test is -1.34. This value is not significant enough to reject the null hypothesis of a unit root at any of the usual significance levels (1%, 5%, or 10%), further the p-value is 0.610135. This suggests that the series may not be stationary. Another stationary test is conducted using KPSS and the result is 3.73 (p-value = 0.01) – suggesting non-stationary at level. After differencing, however, the test statistic of 0.21 (p-value = 0.1) suggest that the data is stationary.

Models Evaluation Visualization

Figure 5 presents the visualization of ARIMA predicted and actual values which shows slight overfitting however its accuracy measures indicate normal performance. Similar trend characteristics are also shown on LSTM-actual values (Figure 6) and XGBoost-Actual values (Figure 7).

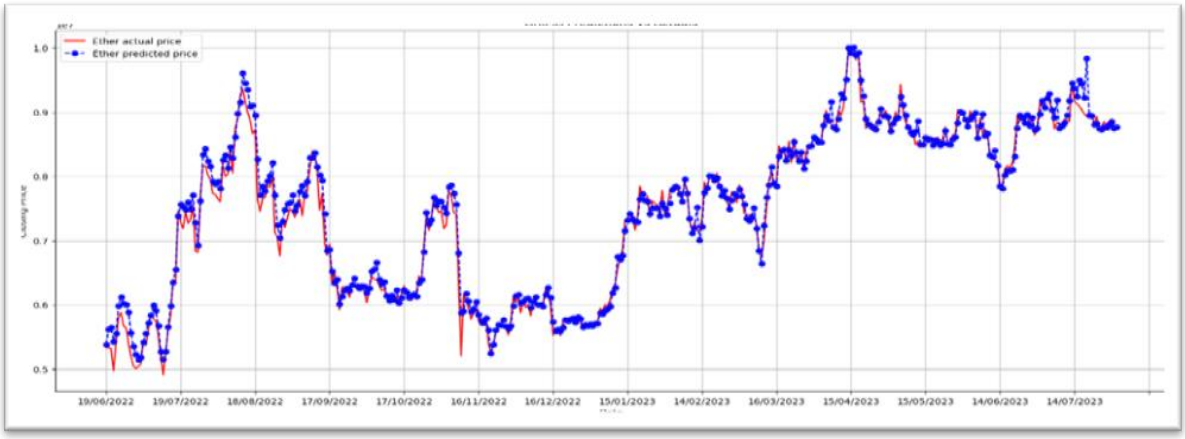

Figure 5. ARIMA predicted vs actual values

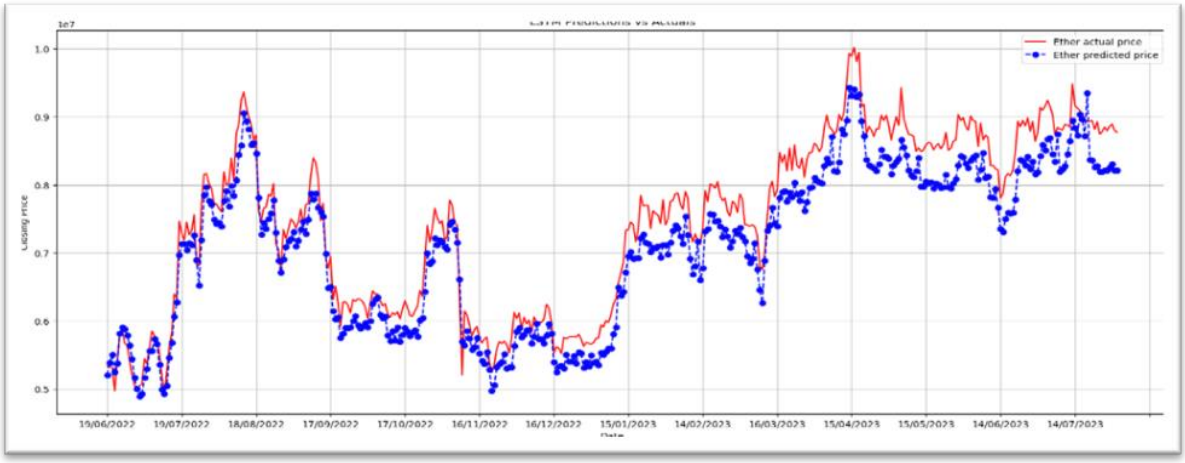

Figure 6. LSTM predicted vs actual values

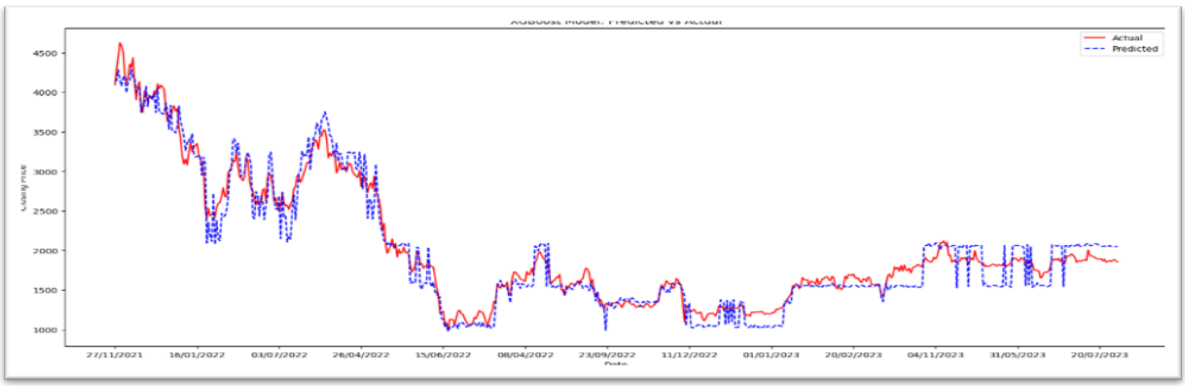

Figure 7. XGBoost predicted vs actual plot

Performance Results

Table 1 shows the performance between ARIMA, LSTM and XGBoost model to help us compare between classical model, supervised model and ensemble learning model (second objective). Based on the RMSE and MAE values (the lowest), LSTM model is suggested to have the best predictive accuracy and makes smaller errors in predictions. In terms of the R^2 value, which measures the goodness of fit, the LSTM model performs slightly better than the other two models, although all three models have R^2 values above 0.9, which is considered high.

Table 1
Performance comparison on test data

Measurements	ARIMA	LSTM	XGBoost
RMSE	62.176	34.505	174.894
MAE	85.244	43.481	143.678
R^2	0.9376	0.9578	0.9522

For the LSTM model in Table 2, the R^2 on the test set (in Table 1) is very close to the R^2 on the training set (0.9578 vs. 0.9811), suggesting that the model generalizes well and is not overfitting. For the XGBoost model, the R^2 values are almost identical for both train and test (0.9918 vs. 0.9522), which is an excellent result, showing robustness in the model. The high train R^2 also indicates that the model can fit the training data very well, and the high test R^2 shows it generalizes well to unseen data.

Table 2
Train and R^2 comparison

R^2	LSTM	XGBOOST
Test	0.9578	0.9522
Train	0.9811	0.9918

Choosing the "champion" model depends on which metric we prioritize:

- If we consider predictive accuracy most important (i.e., the smallest error), the LSTM would be the winner considering its lowest RMSE and MAE.
- If we prioritize generalization (the ability to predict new, unseen data), the LSTM also appears to be very strong, with high R^2 scores that do not degrade from train to test, indicating it is not overfit.

Overall, based on the provided results, the LSTM model seems to be the best performing or "champion" model in this scenario for predicting the price of Ethereum, considering its lower error margins and high R^2 values on both training and testing datasets.

Before addressing non-stationarity, ARIMA achieved an R^2 of 0.7532, 19.6% lower than the final results post stationarity check. Initial ARIMA without a seasonality parameter yielded an R^2 of 0.8921, hinting at SARIMAX's potential for higher accuracy. XGBoost faced initial overfitting with an R^2 of 0.998, later optimized to 0.9522 through parameter adjustments. Further fine-tuning is planned considering the model's high RMSE and MAE. LSTM excelled, scoring an R2 of 0.8721 on 70% of the training data, slightly lower than 0.9578. Data splitting was crucial due to dataset limitations and unique behavior in specific years, ensuring fair comparative analysis across all models.

Discussion

Since Bitcoin's 2009 launch, Ethereum and Litecoin have joined the cryptocurrency market. These digital assets now act as investments, remittances, and payments. Ethereum is a money and a Dapp platform. The rise in cryptocurrency market capitalization and uses has increased price volatility, making accurate price prediction difficult and driving interest in dependable forecasting methodologies (Jagannath et al., 2021). Price prediction studies have progressed from fundamental, technical, on-chain, regression, supervised machine learning, and deep learning. Technical analysis has been utilized by financial experts to identify price and volume

trends. Nonetheless, cryptocurrency volatility and lack of historical data make technical analysis unreliable (Akgül et al., 2022) due to different chart patterns, incorrect past data to predict future performance, failure to address key elements, and lagging temporal signals. Technical analysis alternatives include on-chain analysis.

Prices of Ethereum are the dependent variable, and factors from two other sources are the independent variables. The first source is referred to as 'technical variables', which includes data on the opening price, high price, low price, closing price, adjusted closing price, volume traded, market capitalization, and cumulative return. In the second source, called "on-chain variables," are the number of transactions, blocks, and gas used.

The current methodology, applied to daily price data of Ethereum, will be extended to include more granular data such as hourly and minute-by-minute price data, albeit requiring higher processing power. This study performs prediction in an integrated approach of technical and on-chain analysis. It compares the classical model (ARIMA), supervised model (LSTM) and an ensemble learning under machine learning method (XGBoost).

For our first objective that is to identify relationships and dependencies among variables, several findings are revealed. Strong correlation can be found between prices and market capitalization, prices and traded volume, and market capitalization and volume. Low relationships are found between prices and total transactions, prices and total block number, and traded volume and total block number. The findings highlight that LSTM model as the most promising, showcasing superior predictive accuracy and generalization based on three measurement criteria: RMSE, MAE and R^2 (second objective).

The analysis of the research underlines the continuous effort to enhance predictive models, specifically by investigating creative methods such as SARIMAX and refining established models. Notwithstanding the constraints imposed by hardware, the study's implications are substantial; it advances the field of cryptocurrency price prediction methodologies and lays the groundwork for the community to make well-informed decisions. The research findings have significant effects for investment decision-making and risk management strategies, providing a broader understanding of the complexities of the cryptocurrency market.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process
During the preparation of this work the author(s) used CHATGPT 3.5 in order to assist with generating initial drafts, suggesting improvements, and enhancing clarity. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Data Availability Statement
The data that support the findings of this study are available from the corresponding author, [EIZK], upon reasonable request.

References

Akbulaev, N., Mammadov, I., & Hemdullayeva, M. (2020). Correlation and Regression Analysis of the Relation between Ethereum Price and Both Its Volume and Bitcoin Price. *The Journal Of Structured Finance, 26*(2), 46-56.

Akgül, A., ŞahiN, E. E., & Şenol, F. Y. (2022). Blockchain-based Cryptocurrency Price Prediction with Chaos Theory, Onchain Analysis, Sentiment Analysis and Fundamental-Technical Analysis. *Chaos Theory and Applications, 4*(3), 157–168. https://doi.org/10.51537/chaos.1199241

Alahmari, S. A. (2019). Using Machine Learning ARIMA to Predict the Price of Cryptocurrencies. *ISeCure*, *11*(3).

Alahmari, S. A. (2020). Predicting the price of cryptocurrency using support vector regression methods. *Journal of Mechanics of Continua and Mathematical Sciences, 15*(4), 313-322.

Chen, S. (2022). Cryptocurrency Financial Risk Analysis Based on Deep Machine Learning. *Complexity, 2022*, 1–8. https://doi.org/10.1155/2022/2611063

Chen, Y., & Ng, H. K. T. (2019). Deep Learning Ethereum Token Price Prediction with Network Motif Analysis. *2019 International Conference on Data Mining Workshops (ICDMW)*, 232–237. https://doi.org/10.1109/ICDMW.2019.00043

Drahokoupil, J. (2022). Application of the XGBoost algorithm and Bayesian optimization for the Bitcoin price prediction during the COVID-19 period. FFA Working Papers.

Ferenczi, A., & Bădică, C. (2023). Prediction of Ethereum gas prices using DeepAR and probabilistic forecasting. *Journal of Information and Telecommunication*, 1–15. https://doi.org/10.1080/24751839.2023.2250113

Gkouramanis, A. Γκουραμάνης, Α. (2023). On-chain analysis on the Ethereum blockchain.

Hamayel, M. J., & Owda, A. Y. (2021). A novel cryptocurrency price prediction model using GRU, LSTM and bi-LSTM machine learning algorithms. *AI, 2*(4), 477-496.

Heymans, A., Van Heerden, C., Van Greunen, J., & Van Vuuren, G. (2014). Diligence in determining the appropriate form of stationarity. *Acta Commercii, 14*(1), 14 pages. https://doi.org/10.4102/ac.v14i1.210

Jagannath, N., Barbulescu, T., Sallam, K. M., Elgendi, I., Mcgrath, B., Jamalipour, A., Abdel-Basset, M., & Munasinghe, K. (2021). An On-Chain Analysis-Based Approach to Predict Ethereum Prices. *IEEE Access, 9*, 167972–167989. https://doi.org/10.1109/ACCESS.2021.3135620

John, L. (2021). Machine Learning Classical Models Explained. In *Machine Learning Models Explained*.

Khedr, A. M., Arif, I., El-Bannany, M., Alhashmi, S. M., & Sreedharan, M. (2021). Cryptocurrency price prediction using traditional statistical and machine-learning techniques: A survey. Intelligent Systems in Accounting, Finance and Management, 28(1), 3-34.

Kim, H.-M., Bock, G.-W., & Lee, G. (2021). Predicting Ethereum prices with machine learning based on Blockchain information. *Expert Systems with Applications, 184*, 115480. https://doi.org/10.1016/j.eswa.2021.115480

Korstanje, J. (2023). How to Select a Model For Your Time Series Prediction Task [Guide]. In *How to Select a Model For Your Time Series Prediction Task [Guide]*. https://neptune.ai/blog/select-model-for-time-series-prediction-task

Mahdi, E., Leiva, V., Mara'Beh, S., & Martin-Barreiro, C. (2021). A new approach to predicting cryptocurrency returns based on the gold prices with support vector machines during the COVID-19 pandemic using sensor-related data. Sensors, 21(18), 6319.

Nayam, W. (2022). XGBoost for prediction of Ethereum short-term returns based on technical factor [Master of Science, Chulalongkorn University]. https://doi.org/10.58837/CHULA.THE.2022.108

Nayomi. (2021). World Economic Outlook, April 2021: Managing Divergent Recoveries. *International Monetary Fund*. https://www.imf.org/en/publications/weo?page=2

Petrusheva, N., & Jordanoski, I. (2016). Comparative analysis between the fundamental and technical analysis of stocks. *Journal of Process Management. New Technologies, 4*(2), 26–31. https://doi.org/10.5937/JPMNT1602026P

Raju, S. M., & Tarif, A. M. (2020). Real-time prediction of Bitcoin price using machine learning techniques and public sentiment analysis. arXiv preprint arXiv:2006.14473.

Rizwan, M., Narejo, S., & Javed, M. (2019). Bitcoin price prediction using Deep Learning Algorithm. In *2019 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS)* (pp. 1-7). Karachi, Pakistan. doi: 10.1109/MACS48846.2019.9024772

Salam, M. A., Alazzam, M. B., & Asiri, H. A. (2019). Forecasting the price of Ethereum using ARIMA model. 2019 International Conference on Information and Communication Technology Research (ICTRC).

Van Greunen, J., Heymans, A., Van Heerden, C., & Van Vuuren, G. (2014). The Prominence of Stationarity in Time Series Forecasting. *Studies in Economics and Econometrics, 38*(1), 1–16. https://doi.org/10.1080/10800379.2014.12097260

Wang, J., Yao, J., & Zou, X. (2020). Predicting Ethereum price by Machine Learning Techniques. 2020 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C).

Zanardi, M., & Jaen, H. (2023). Application Of Triple Exponential Smoothing Method For Predicting The Price Movement Of Cryptocurrency Ethereum. *1*(2).

Zhao, H., Crane, M., & Bezbradica, M. (2022). Attention! Transformer with Sentiment on Cryptocurrencies Price Prediction: Proceedings of the 7th International Conference on Complexity, Future Information Systems and Risk, 98–104. https://doi.org/10.5220/0011103400003197