Empirical Evaluation of Machine Learning Performance in Forecasting Cryptocurrencies

Lauren Al Hawi [1], Sally Sharqawi [1], Qasem Abu Al-Haija [2,*], and Abdallah Qusef [3]

[1] Department of Business Intelligence Technology, Princess Sumaya University for Technology, Amman, Jordan;
Email: lor20208050@std.psut.edu.jo (L.A.H.), sal20208071@std.psut.edu.jo (S.S.)
[2] Department of Cybersecurity, Princess Sumaya University for Technology, Amman, Jordan
[3] Department of Software Engineering, Princess Sumaya University for Technology, Amman, Jordan;
Email: a.qusef@psut.edu.jo (A.Q.)
*Correspondence: q.abualhaija@psut.eud.jo (Q.A.A.)

***Abstract*—Cryptocurrencies like Bitcoin are one of today's financial system's most contentious and difficult technological advances. This study aims to evaluate the performance of three different Machine Learning (ML) algorithms, namely, the Support Vector Machines (SVM), the K Nearest Neighbor (KNN), and the Light Gradient Boosted Machine (LGBM), which seeks to accurately estimate the price movement of Bitcoin, Ethereum, and Litecoin. To test these algorithms, we used an existing continuous dataset extracted from Kaggle and coinmarketcap.com. We implemented models using the Knime platform. We used auto biner for volume and market capital. Sensitivity analysis was performed to match different parameters. The F and accuracy statistics were used for the evaluation of algorithm performances. Empirical findings reveal that the KNN has the highest forecasting performance for the overall dataset in our first investigation phase. On the other hand, the SVM has the highest for forecasting Bitcoin and the LGBM for Ethereum and Litecoin in the individual dataset in the second investigation phase.**

***Keywords*—cryptocurrency, machine learning, Support Vector Machines (SVM), K Nearest Neighbor (KNN), Light Gradient Boosted Machine (LGBM), Bitcoin, Ethereum, Litecoin**

I. Introduction

The world is undergoing a digital revolution, which has impacted many parts of people's lives. One of the most contentious and ambiguous inventions in the current global economy is the fast rise of digital currencies during the last decade [1]. Today, the Internet regulates most aspects of life in a virtual environment. Furthermore, according to Lahmiri and Bekiros [2], the banking industry has shifted away from traditional working methods and toward methods that prioritize technology and speed, as the economy's structure, financial markets, and payment systems are all changing as technology changes. Financial markets worldwide have grown more digitized than ever, and a society less dependent on physical currency is upon us. People may now use digital technologies to build their money (digital cryptocurrency) and central bank operations [1].

As stated by Pabuçcu and Ongan *et al*. [3], "Cryptocurrencies" have become one of the most intriguing and maybe most misunderstood phenomena of the early twenty-first century since their inception in 2008. The financial revolution's growth has improved because the interest in cryptocurrencies has gradually increased over the last decade, as cryptocurrency has been instilled in people's lives at all levels and has become better understood [3]. The most trending technological advancements in the modern world are cryptocurrency and Artificial Intelligence (AI) [4]. However, both innovations have a tremendous impact and enhance how things are done in many areas of our lives [5]. Forecasting the movement or prices of cryptocurrencies using different machine-learning algorithms has limited research studies and is considered a new field [1].

A. Cryptocurrency

Cryptocurrency has rapidly grown in popularity over the Internet and several online platforms. Cryptocurrencies are forms of digital money that are encrypted (currency) [6]. They can be traded for something with value. Cryptocurrencies, like any used currency, are used to pay for any product or service. They can also be exchanged and preserved for investment [7].

B. Block Chain

The term comes from how the system's information and files are organized in the blockchain, which consists of single entries called blocks connected on a list called chains [4]. The blockchain was developed to alleviate the dual spending issue and undermine centralized organizations' power in asset transactions, Bitcoin's most significant invention [6]. Most cryptocurrencies are built on "blockchain technology". A blockchain is a special form of database that allows most cryptocurrencies, such as Bitcoin, to exist. It acts as a Peer-to-Peer (P2P) public ledger, keeping track of bitcoin transactions [8].

Manuscript received February 27, 2023; revised March 24, 2023; accepted April 7, 2023; published July 11, 2023.

C. Bitcoin (BTC)

As of October 2020, there were 7378 cryptocurrencies, with a combined market capitalization of about 359.7 billion in USD [9]. Bitcoin is a Peer-to-Peer (P2P) payment cash system that allows internet payments to be transmitted directly from one to another without a need to go to a banking institution [10]. It was created in 2008 by Satoshi Nakamoto and became well-known in 2009. Bitcoin is the most successful blockchain application and is cryptographic digital money. It was a non-regulated digital currency with no legal standing. It is classified as a cryptocurrency because of its cryptographic role in generating and transferring funds [11]. According to Krause and Pham [12], Bitcoin has been the most prominent currency in volume trading in recent years, making it the most promising financial medium for investors. As stated by Farell [13], it secures the transactions by encrypting the sender, receiver, and transaction volume. It is regarded as the most popular cryptocurrency. Cryptocurrencies have become a popular investment objective worldwide since their introduction [6]. However, Nakamoto is regarded as the "Father of Cryptocurrencies" [4].

D. Ethereum (ETH)

Ethereum is a Turing-complete decentralized blockchain-based framework for creating and executing smart contracts or distributed systems, where the coin's value is referred to as ether [14, 15]. As mentioned by Seys and Decaestecker [15] in their research, it was founded in 2013 by "Vitalik Buterin" and financed a year later with 18 million USD in bitcoins gained from a crowded online public auction. Ether's circulation is unrestricted; it allows anybody to build their own rules and transaction formats and can be exchanged in cryptocurrency trades [16]. Ethereum has full Turing completeness meaning it can do any computation, including loops [16].

E. Litecoin (LTC)

According to Chuen and Guo *et al.* [14], Litecoin was created by Charles Lee and published in October 2011, utilizing a similar technology to Bitcoin. The block production time has been reduced by up to 4 times every block (from ten minutes to two and a half minutes per block), and the highest limit has been increased to 84 million, comparable to four times that of Bitcoin. Litecoin, regarded as the cryptocurrency silver standard, has become the second most widely accepted by exchanges and users (miners) [1].

F. Research Motivation

Due to the investment risk of cryptocurrencies, anticipating the price fluctuation trend of cryptocurrencies is critical [11]. Minimal studies, as will be discussed in the literature, have employed K Nearest Neighbor (KNN), Support Vector Machines (SVM), and Light Gradient Boosted Machine (LGBM) together to forecast the price movement of cryptocurrencies. Most of the studies used different ML algorithms, and some highlighted that SVM outperforms LGBM or KNN. Because cryptocurrency exchange rates are notoriously volatile, and to address the issue of people being unable to decide which way to invest in cryptocurrency and to mitigate the risk of the uncertainty in investing, instead of using traditional economic models that only work time series, in this paper, we aim to conduct a comparative performance study of machine learning algorithms to answer the question of which of the proposed models is the most effective in forecasting the movement "high," "low" of three cryptocurrencies and to understand which model outperforms the other in different circumstances, by testing the algorithm's performance on the three cryptocurrencies historical data. Bitcoin, Litecoin, and Ethereum cryptocurrencies are just a few examples. However, other cryptocurrencies should be considered more in the coming literature. The reason is that according to Ref. [16, 17], today's most well-known and valued cryptocurrencies are Bitcoin and Ethereum. However, there has been a quick rise of several cryptocurrencies recently, Litecoin, Ripple, and Stel-lar. This allows us to better comprehend the general price dynamics on a few cryptocurrency markets rather than multiple digital currencies.

G. Research Contribution

To the best of our knowledge, this study differs from other studies because most research studies have not used "auto biner" technology to bin any attributes. Most studies also used the data as it is, while we have computed the difference between high/low and open/close price attributes. Moreover, to justify each classifier parameter, input sensitivity analysis was done. The research will answer the following questions: RQ1: Which models SVM, KNN, and LGBM, can accurately forecast the movement of cryptocurrencies? RQ2: which model of the proposed models is the best in forecasting the movement of BTC, ETH, and LTC?

H. Machine Learning Algorithms

The goal of modeling SVM, KNN, and LGBM is that first, SVM is a common and effective approach for solving classification problems that have been widely applied in a variety of domains especially forecasting the movement of financial markets [18–20], second, a K-Nearest Neighbor model can be used for both classification and regression problems and that are simple to construct [21]. KNN is an instance-based learning approach [22]. Instance-based learning differs from model-based learning because it does not require any parameter tuning or training and may be used to produce predictions immediately. The main idea is to create and use a classification hyper-plane to divide the data into two pieces. The nearest distance between the hyper-plane and the data points obtains the largest value with the greatest possible margin [9]. Third, LGBM is an excellent way to process large-scale data and features. Forecasting accuracy and robustness are higher using the LGBM model than in different ML algorithms [6]. Moreover, Sun and Liu *et al.* [11] used a Gradient Boosting Decision Tree (GBDT) method called Light Gradient Boosting Machine (LGBM) to anticipate the price trend of

cryptocurrencies, claiming that the LGBM model is more robust than SVM and RF models.

II. Literature Review

According to Kostková and Omelina *et al.* [23] and Mohtasham [24], for more than two decades, Machine Learning (ML) has been widely known as a significant model in the industry of forecasting as classical statistics. Artificial Neural Networks (ANNs) and Support Vector Machines (SVM) are two of the most extensively used methods for anticipating financial assets and cryptocurrencies [25], and each has its unique learning patterns [26, 27]. Cryptocurrencies like Bitcoin and Litecoin and any transactions conducted with them are difficult to track. However, the benefits of cryptocurrencies greatly exceed any dangers because they are quick, inexpensive, and extremely secure [1].

The largest issue for traders and regular users is the exchange rate volatility in Bitcoin. Thus, in the financial sector, the ability to forecast asset price movements is a practical consideration that heavily impacts a trader's choice to purchase or sell an investment instrument. The number of studies on the Bitcoin exchange rate time series is growing, although it is still relatively new, as stated by Mallqui and Fernandes [28]. According to Żbikowski [29], to deal with the cryptocurrency stock market, we believe it is vital to use ML technologies to predict the movement of cryptocurrency prices. Due to its significant volatility and price swings, and according to Chowdhury and Rahman *et al.* [6], few publications anticipate price variations. Also, according to Akyildirim and Cepni *et al.* [30], the price of bitcoin has risen at an exponential rate. Other cryptocurrencies, such as Ethereum, Ripple, and Litecoin, followed suit, with values rising by thousands of percent in 2017.

Different ML techniques are pitted against each other to illustrate which MLA performs better at predicting the price movements of cryptocurrencies. Astronautica [24] employed the SVM learning algorithm to see if it could forecast Bitcoin price changes. They found that SVM can anticipate Bitcoin prices five steps ahead in the short, medium, and long term, as well as the total Bitcoin price level. In another study conducted by Hitam and Ismail [31], the findings of the classifiers that performed the best on the test dataset were tested. They revealed that the SVM classifier is more effective for Ethereum and Litecoin. At the same time, ANN functions best for Bitcoin, Nem, and Ripple. Boosted NN has the highest performance accuracy for Ripple and Stellar. However, compared to other classifiers, the SVM classifier performed the best, with a performance accuracy of 95.5%.

Based on some studies highlighted by Akyildirim and Cepni *et al*. [30] that support vector machines (SVM), as well as logistic regression (LR) techniques, they outperform random forest (RF) and artificial neural network (ANN) algorithms. SVMs are known for their capacity to generalize to varied timelines and market situations and their robustness in the face of noisy input. SVMs, as well as LR, perform admirably over a variety of timelines and cryptocurrencies.

In another experiment made by Saadah and Whafa [32], where Long Short-Term Memory (LSTM) approach outperformed the K-Nearest Neighbor (KNN) and SVM. The outcome of the mistake is lesser when using the LSTM algorithm than other methods. However, LSTM takes a lengthy time to execute. SVM prediction, on the other hand, has a lower error than KNN. Meanwhile, SVM outperforms LSTM in terms of time execution. And as a result, if there are resource constraints, the SVM may be used to monitor financial stability [32]. For bitcoin investing, Barnwal and Bharti *et al*. [33] presented a stacking with NN, XGB, SVM, K-NN, and LGBM used as discriminate classifiers to build stacks optimized over one neural network layer to model the direction of cryptocurrency price. The models showed accuracies of 57%, 48%, 59%, 61%, and 52%, respectively. Qa and Alnabhan *et al*. [34] looked into the LR, RF, SVM, and GBM machine learning approaches for predicting price trends. The best outcome came from group classifiers with an accuracy of 59.3%. In a recent work done by [35] where he chose to investigate the data set of three crypto-currencies (Bitcoin, XRP, and Ethereum) using two nonlinear algorithms: Decision Tree Regressor (DTR) and the (KNN) [36]. However, little attention is paid to using and comparing nonlinear methods such as the DTR and KNN regression [35]. Mallqui and Fernandes [28] used ANN and SVM to forecast Bitcoin price direction and daily exchange rates. The SVM had the greatest performance regarding price trend change (classification problem, accuracy 59.45%) and exchange rate forecasts (regression problem, MAPE within 1.52–1.58%). For anticipating the direction of price fluctuations in Kumar and Gopal [18] examined the forecasting abilities of Long Short-Term Memory (LSTM) and Multi-Layer Perception (MLP). They analyzed daily, hourly, and minute data and found that while LSTM needs much longer training time than MLP, it does not outperform it significantly. Chen and Wei *et al*. [9] used various learning models for Bitcoin 5-minute intervals and daily pricing, including RF, Extreme Gradient Boosting (XG Boost), Quadratic Discriminant Analysis, SVM, and LSTM. Statistical approaches (average accuracy of 65%) produced better results for daily pricing than ML methods, which was an unexpected finding (average accuracy of 55.3%). The SVM was the most accurate of the ML models, with a 65.3% accuracy. According to Akyildirim and Cepni *et al*. [30], they used six different ML algorithms, KNN, LR, Naive Bayes, RF, SVM, and XG Boost, in forecasting the movement of cryptocurrencies. Their findings show that the KNN approach and the RF algorithms produce the highest in and out-of-sample accuracy rates at different frequencies. For example, the RF algorithm in-sample success rate can reach 87%. SVM, on the other hand, achieves the greatest out-of-sample success rates. To forecast the cryptocurrency market's price movement, Sun and Liu *et al*. [11] offered three models: SVM, RF model, and Light Gradient Boosting Machine (LGBM). LGBM stands out as a superior

technique to the others [6]. And, to anticipate the price movement of Digital Cash, Bitcoin, and Ripple, Lahmiri and Bekiros [37] suggested two deep learning techniques: Deep Learning Neural Networks (DLNN) and Generalized Regression Neural Networks (GRNN). The Root Means Square Error (RMSE) values produced using DLNN and GRNN are extraordinarily high in the Bitcoin and Digital Cash case. However, this value is considerably lower in the case of Ripple. Greaves and Au [38] advocated using transaction graph data to forecast Bitcoin price changes by collecting Bitcoin transactions. Baseline, SVM, LR, and NN models are the four classification models they employed. The models' accuracy was 53.4%, 53.7%, 54.3%, and 55.1%, respectively. Sun and Liu *et al*. [11] used a modification of GBM to construct a unique approach for predicting the price trend of the crypto-currency market (LGBM). They examined multiple periods (two days, two weeks, two months). They found that the two-week forecasting time horizon produced the best results: accuracy ranging from 0.52% (RF) to 0.61% (SVM, LGBM). For XRP, SGBM achieves the greatest prediction performance of 0.92%, while RF achieves the best outcome of 1.84%. In their study of forecasting price constituents, Chowdhury and Rahman *et al*. [6] found that the maximum predicting performance using the LGBM model is 0.905 for two weeks in the first category of training sets and 0.952 for two weeks in the second category of training sets, indicating that accuracy is 0.924 using the ensemble learning approach, and 0.952 was the highest accuracy of LGBM in the second category of training sets [11].

However, according to Chowdhury and Rahman *et al*. [6], unlike other models, the K-NN model could have performed better when it comes to predicting; this is owing to severe volatility and noisy random characteristics. They compared the ensemble learning method, gradient boosted trees model, NN model, and (K-NN) model to models from literature-based state-of-the-art and found that these models performed better and were more competitive. The ensemble learning strategy, regarded as the best among all the models tested in this work, yielded 92.4% accuracy. This is not to claim that ML algorithms always outperform traditional models. However, with large data sets, ML approaches, built expressly to address certain issues, may give better results. All of this has increased the popularity of ML algorithms among academics. Much empirical research compares conventional models with ML algorithms in terms of forecasting performance. Traditional models outperform ML algorithms in certain research, while the latter outperforms others in other studies [15]. For example, rewired publications by Zhu *et al.* [39], Yao *et al.* [40] concentrated on forecasting exchange rates (prices) using RF and GBM, which looked at the prediction of price trend changes (classification problem, regression problem). It also employed the stochastic GBM (SGBM), which has several benefits, such as a shorter learning period, less memory usage, and higher accuracy.

Traditional models analyze causal correlations using the complete data set, whereas ML approaches divide the data set into training and testing sets. ML gives computers the ability to "learn" and generate predictions consequently. Even though both strategies aim to enhance accuracy by reducing various loss functions, ML uses nonlinear algorithms [41, 42]. Due to the rising economic-political uncertainty, and like any financial market, cryptocurrency changes may be difficult to anticipate for investors, as stated by Seys and Decaestecker [15], and since the goal of this research is to compare the performance of three ML models Support Vector Machine (SVM), Light Gradient Boosted Machine (LGBM) and K Nearest Neighbor (K-NN) in forecasting the movement of three cryptocurrency considering a high degree of accuracy which rely on flexible assumptions to examine which algorithm performs well.

A. K-Nearest Neighbor (KNN)

According to Hitam and Ismail [31], the K-NN algorithm is a non-parametric approach comparing a data set to K training instances that are the dataset's example's closest neighbors. This approach aims to create a K-Nearest Neighbor model that may be used for regression and classification [37]. As a result, it is easy to see how the K-NN algorithm might be useful in projecting the price of components and cryptocurrencies [6]. Furthermore, the K-NN technique is a lazy learning algorithm that uses instances to learn complex class functions rapidly without any information loss [43]. KNN uses a similarity function for a new data instance to locate the most comparable K examples in the training dataset [35].

B. Support Vector Machine (SVM)

The SVM algorithm is a model-based learning technique that is now one of the most effective classifiers according to [44]. In Hitam and Ismail [31], the Support Vector Machine (SVM) classifier or algorithm was established as an initiative principle to avoid the issue of over-fitting in the data when modeling the training dataset, and it is known for its flexibility in creating clear and accurate frontiers [45, 46]. SVM is effective in a wide range of applications due to its ease of use, and the quick training results it provides [47]. It provides a solid and nonlinear solution by mapping the input space onto a maximum dimensional feature using core functions, as stated by Hacib [48]. The SVM has several advantages, including the ability to outperform generalization models and operate well with minimal datasets, which will remain to deliver strong generalization results, including pattern classification [37].

C. Light Gradient Boosted (LGBM)

The LGBM algorithm is described as a relatively new algorithm used by Sun and Liu *et al.* [11]. LGBM is a new GBDT (Gradient Boosting Decision Tree) technique published by Ke and his colleagues in 2017 that has been applied to various data mining applications, including classification, regression, and ordering [49]. Light GBM, unlike standard GBDT-based techniques like XG Boost and GBDT, grows the tree vertically, while other algorithms build trees horizontally. Moreover, Machines for Increasing Gradient and NN and SVM boosting have

been among the most prominent ML approaches in the last two decades. Boosting, unlike bagging, uses a weighted voting system rather than a basic one. One of the most appealing aspects of boosting is how simple it is to create computationally efficient weak learners. A shallow decision tree, or a decision tree with a minimum depth limit, is a popular weak learner [3].

III. Methodology

This paper's major goal is to compare the performance of the most accurate forecast of the following day's movement for three cryptocurrencies BTC, ETH, and LTC. Our study is based on their initial values and other parameters. Since we focus on three ML techniques, SVM, KNN, and LGBM, for predicting cryptocurrency data, these parameters include digital coins and their link between transaction volume and market capitalization in USD volume. In this section, we discuss our research model (shown in Fig. 1). The work is divided into three phases: (1) dataset data preprocessing phase, (2) building and training various models using the Knime tool, and (3) sensitivity analysis.

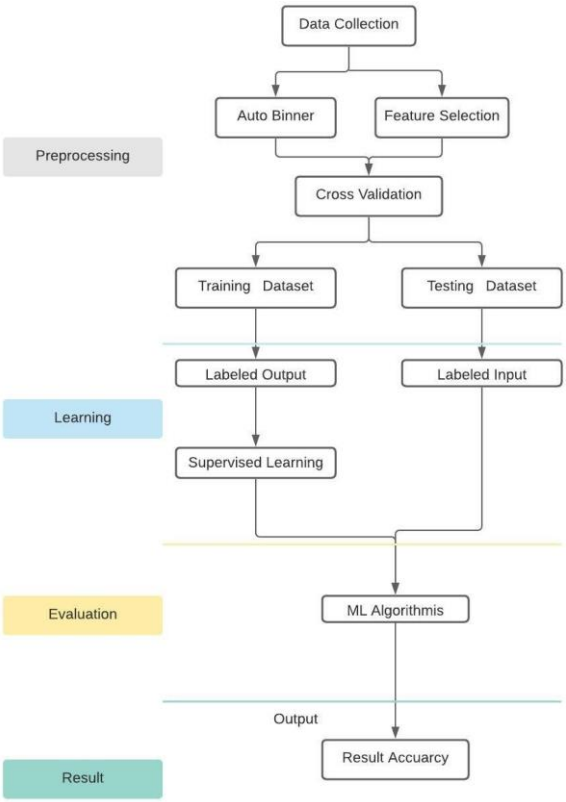

Figure 1. The proposed research model.

A. Dataset

The dataset was mainly extracted and merged from Kaggle, www.kaggle.com/sudalairajkumar/crypt, and coinmarketcap.com, used by Chowdhury and Rahman *et al.* [6]. The dataset represents historical BTC, ETH, and LTC prices from December 23rd, 2021, till January 9th, 2022. We chose different data sources because the first data set covers the interval from December 23rd, 2013, till July 6th, 2021, and the second data set covered the rest till January 9th, 2022. Since Kaggle is a familiar public resource that provides support confidence interval of 99% for the data as to [50], cryptocurrency data on Kaggle is published on daily intervals, which were the most retail transactions data. In contrast, due to the interval, which didn't cover the entire period, we found the rest of the data on coinmarketcap.com, the website we had access to. The data set was valid and clean, with no missing, duplicates, mismatched, or irrelevant values. We have implemented the "Knime" tool during the overall phases. Moreover, the data was highly balanced, with total records of 8217 classified as 52% "High" and = 48% "Low". Table I shows the descriptive statistics for the three coins.

TABLE I. BTC, ETH, LTC Descriptive Statistics

Measure	Bitcoin USD	Ethereum USD	Litecoin USD
Mean	1.819891E+11	7.154723E+10	3.724737E+09
SD	2.892949E+11	1.235381E+11	4.470477E+09
Minimum	2.444379E+09	3.221363E+07	4.117319E+07
Maximum	1.274831E+12	5.690943E+11	2.579652E+10
Sum	5.341381E+14	1.679213E+14	1.093210E+13
Count	2935	2347	2935

The dataset consisted of the seven most used attributes. The attributes before preprocessing are shown in Table II.

TABLE II. Attributes Description

Attribute	Description
Name	Cryptocurrency name
Symbol	Known as
Date-Time	Date and price closing time of each observation
Open	The opening price on the given day
Close	Closing price on the given day
High	The highest price on the given day
Low	Lowest price on the given day
Volume	The volume of transactions on the given day
Market Capital	Market capitalization in USD

B. Data Preprocessing

Feature Extraction: These attributes were hand-selected in our study to their relevance to the problem we're attempting to tackle since most of the studies used such features in predicting movement and even the prices. First, name of the coin: The name of the coin was discretized as "1" for Bitcoin BTC, "2" for Ethereum ETH, and "3" for Litecoin LTC. Second, date: the prices were reflected each day at 12:00:00AM. Third, open-close price difference: We calculated the difference between open-close prices, which indicates the range of increase/ decrease in price. Fourth, high-low price difference: We calculated between high and low prices, indicating the range of increase/ decrease in price. Fifth, we have binned the volume attribute and market capital dis- cussed further in the sensitivity analysis. Lastly, a class representing movement "High", "Low": we have calculated the movement in the price as the following

formula (If ((Today-current "next day")> 0, "Low", "High"). Since this is a classification problem, we need to study the movement in the price of the three mentioned coins, whether it will be "high" or "low" or the next day. The data set was divided using K fold cross validation-random sampling, specifically X partitioner, which eliminates errors according to Barnwal and Bharti *et al*. [33].

Volume and Market Capital Binning: First, all the values equal = 0 from the "volume" dimension were removed since they do not indicate any transaction volume. Second, we use a marginal split to partition the volume data while generating frequency bins with about equal data points. Even if the bins are not of identical size, computing the relative frequencies in each bin is adequate. According to Keskin and Aste [50], binning avoids bias, finding odd-numbered and even-numbered bin counts to provide similar results, suggesting a key benefit in using bins to calculate the selected information. As Davis and Charlton *et al.* [51] stated in their study, they used synthesized data to illustrate the problems with standard binning methods and show that adaptive binning can reduce within-class variance. However, binning was used by Keskin and Aste [50] when they noticed that more bins resulted in a bigger transfer of selective information for the same data, amplifying both the signal and the noise. According to our study, we tested granular partitions from four to ten bins for volume and market capital notations to capture the best output. And as a result, when testing more than bin partitions, depending on the number of bins used, performing the analysis with equal-width bins leads to varied but near results. We presented our results in the sensitivity analysis section for the binned attributes. We chose a partition size equal to eight that produces good and meaningful findings for each studied currency.

C. Sensitivity Analysis

Due to different parameters values for each classifier, such as the binning for volume attribute, k-fold in cross-over validation split, KNN, Radial Basis Function (RBF) in SVM, and the tree depth in LGBM, we had to test different input values for each parameter in each partition and choose the best accuracy result for each classifier as shown below in Tables III–V.

TABLE III. BINNING, K-FOLD, KNN PARAMETERS

Parameter	KNN
Bins Frequency	Vol and MC
4	0.5
6	0.51
8	0.53
10	0.52
K-fold	Bin = 8
5	0.54
7	0.55
10	0.47
20	0.54
KNN	K-fold = 7, Bin = 8
3	0.49
5	0.52
7	0.53
10	0.54
20	0.49

TABLE IV. RBF KERNEL FUNCTION

RBF	Bin = 8, K-fold = 7
1	0.52
10	0.53
100	0.52

TABLE V. LGBM TREE LIMIT

Tree Limit	Bin = 8, K-fold = 7
5	0.53
7	0.54
10	0.51
20	0.53

According to the above accuracy results, we chose bin = 8 as the volume and market capital frequency, starting with the binning value. Followed by several K-fold tests and chose K = 7 random samplings. Tested with KNN algorithm and chose K value = 10. Then we entered Bin = 8, K-fold = 7 for SVM separately without KNN to choose whether polynomial or the RBF among different kernels. We used RBF randomly since it provided the same results in the polynomial. We tested three RBF values and chose ten according to the accuracy measure. And lastly, when entering bin = 8, K-fold = 7 on the LGBM model, individually chose the tree depth for LGBM = 7, which showed the highest accuracy. The final best parameter values obtained by each classifier are shown in Table VI.

TABLE VI. BEST PERFORMANCE CLASSIFIERS

Parameters	Values	Accuracy
Bin for both Volume/MC	8	0.53
K-fold	7	0.55
KNN	10	0.54
RBF for SVM	10	0.53
Tree Limit in LGBM	7	0.54

With all of the abovementioned, we have tested the three model's accuracy together on two inputs with different splits; the first was entering each crypto coin into models. The second was entering the data of the three coins and testing their accuracy, which will be discussed further in the results.

Moreover, an essential part of evaluating different machine learning models is choosing a representative metric that measures the quality of the model and can be easily monitored. Such an evaluation metric is important to differentiate between models' performances and tune their hyper-parameter. Since the data is highly balanced, accuracy and F-measure would be good performance measures in the applied data set. A confusion matrix has been used to evaluate the performance of each ML algorithm KNN, SVM, And LGBM, which can be represented mathematically as mentioned in [52]:

$$\text{Recall P/N} = \text{True Positives} / (\text{True Positives} + \text{False Negatives}) \quad (1)$$

$$\text{Precision P/N} = \text{True Positives} / (\text{True Positives} + \text{False Positives}) \quad (2)$$

$$\text{F-Measure} = (2 \times \text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (3)$$

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN) \quad (4)$$

IV. Results and Analysis

For the classification algorithms, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Light Gradient Boosted Machine LGBM are evaluated here. These three algorithms are selected as they are easy to implement, provide a good result with medium data sets, and can provide a good baseline. Other algorithms and further optimization can be applied here. Tables VII–IX represent the results of Experiment 1, predicting each classifier's Bitcoin, Ethereum, and Litecoin individually. Compared to other studies, the forecasting results using the mentioned models were deemed low. This could have happened due to a number of features being selected; other features should be considered, or other models should be included and more tuning to the models' parameters. However, the data set was split into 420 as a test set for Bitcoin, 336 for Ethereum, and 294 for Litecoin. According to the results, SVM is the best classifier in predicting bitcoin with an accuracy of 0.58%, as it predicts high class with 0.71 F-measure and 0.55 precision. LGBM follows them with an accuracy of 0.54%. While LGBM is the best for predicting Litecoin with the highest accuracy of 0.55%, it also predicts its "high" and "low" classes with a qual F-measure of 0.55. Moreover, LGBM is the best classifier in predicting Ethereum than SVM and KNN, with an accuracy of 0.55%, with equal F-measure for "high" and "low" movements.

TABLE VII. Bitcoin Results

Bitcoin	KNN	SVM	LGBM
Accuracy	0.53	0.58	0.54
Precision (High)	0.58	0.55	0.56
Precision (Low)	0.42	0.43	0.48
Recall (High)	0.07	1	0.68
Recall (Low)	0.31	0.05	0.36
F1(High)	0.63	0.71	0.62
F1(Low)	0.35	0.032	0.41

TABLE VIII. Ethereum Results

Ethereum	KNN	SVM	LGBM
Accuracy	0.46	0.41	0.51
Precision (High)	0.51	0.50	0.51
Precision (Low)	0.50	0.43	0.46
Recall (High)	0.65	1	0.53
Recall (Low)	0.36	0.04	0.43
F1(High)	0.57	0.67	0.51
F1(Low)	0.42	0.51	0.44

TABLE IX. Litecoin Results

Litecoin	KNN	SVM	LGBM
Accuracy	0.53	0.51	0.55
Precision (High)	0.51	1	0.56
Precision (Low)	0.50	0.51	0.54
Recall (High)	0.43	0.01	0.54
Recall (Low)	0.59	1	0.57
F1(High)	0.46	0.013	0.55
F1(Low)	0.54	0.66	0.55

According to the results in Table X, when entering the whole cryptocurrencies dataset. The three classifiers can predict the price movement, whether "high" or "low", but with different accuracies and Pearson correlation values. The test set consisted of 1174 as to k-fold validation for the data split. The results show that KNN outperforms SVM and LGBM with accuracies of 0.53%, 0.51%, and 0.51%, respectively. It also predicts 0.59% of the "high" movements in the F-measure. Due to the numbers, SVM and LGBM work almost the same in predicting movements with equal accuracy of 0.51% and F-measure of 0.54% and 0.53%, respectively. Also, according to the results of investigating the whole data set, LGBM predicts "low" movements better than SVM and KNN. KNN is the best classifier for predicting the lower bound of the movement, as it predicts almost 22% of the downward movements. While SVM and LGBM, on the other hand, are the best classifiers in predicting high bound in the test set with 37% and 33%, respectively. While still, KNN gave good results for high bound with 32% from the test set. More improvements can be made in further studies to alleviate each classifier's output.

TABLE X. All Dataset Results

Measure	KNN	SVM	LGBM
Accuracy	0.53	0.51	0.51
Precision (High)	0.53	0.51	0.52
Precision (Low)	0.52	0.49	0.50
Recall (High)	0.66	0.58	0.54
Recall (Low)	0.38	0.42	0.48
F-Measure (High)	0.59	0.54	0.53
F-Measure (Low)	0.44	0.45	0.49

In addition to the accuracy statistics, Table XI illustrates the confusion matrix for each classifier on the test consisting of 1174 results in the test set from k-fold validation. According to the results of investigating the whole data set. KNN is the best classifier for predicting the low bound of the movement, which means it predicts almost 22% of the downward movements. While SVM and LGBM, on the other hand, are the best classifiers in predicting high bound with 37% and 33%, respectively, from the entire test set. While still, KNN gave good results for high bound with 32% from the test set. More improvements can be made in further studies to alleviate the outputs of each classifier.

TABLE XI. Three Models Confusion Matrix (Horizontal = Predicted, Vertical = Actual)

SVM	Low	High
Low	174	397
High	166	437
KNN	Low	High
Low	254	317
High	222	381
LGBM	Low	High
Low	181	390
High	219	384

TABLE XII. Previous Results

Ref.	Results
[30]	KNN and RF produce the highest accuracy = 87%
[33]	Accuracies of 57%, 48%, 59%, 61% and 52% respectively
[53]	SVM Ethereum = 95.5% and Litecoin = 82.4%. ANN Bitcoin = 79.4%
[9]	SVM was the most accurate at 65.3%
[6]	The accuracy for LSTM = 52.78% and the accuracy for RNN was 50.25p, Ensemble = 92.4% while the highest accuracy is 95p using LGBM model, RMSE for LSTM = 6.87%, RNN = 5.45%, ELM = 0.934, GB = 0.001

[31]	SVM accuracy rate = 95.5%. Bitcoin 78.90%, Ethereum 95.50%, and Litecoin 82.40%
[28]	The accuracy rate of 59.45%, ANN = 53.4%
[25]	Accuracy for each model: ARIMA = 50.05%, RNN = 50.25%, LSTM = 52.78%, LGBM are 90% and 0.924%, using (EML)
[11]	The LGBM model works better than SVM and RF models. LGBM had max accuracy of 91%
Our Study	SVM: BTC = 0.58%, LGBM ETH, LTC= 0.51%, 0.55%. All: KNN = 0.53%, SVM = 0.51%, LGBM = 0.51%

Finally, Table XII highlights several studies conducted on forecasting cryptocurrency price movement. Each study used different algorithms, and each classifier showed different results. Most of the notations in this study are used in these previous studies.

V. Conclusion

Comparing the results in the two experiments to evaluate the performance of the three Machine Learning (ML) algorithms, the Support Vector Machines (SVM), the K Nearest Neighbor (KNN), and Light Gradient Boosted Machine (LGBM) on Bitcoin, Ethereum, and Litecoin. The performance on each classifier differs when entering individual coins than entering the whole dataset; when investigating the whole data set, KNN showed the most accurate classifier among SVM and LGBM with an accuracy of 53% for predicting cryptocurrencies movement "high" and "low", followed by SVM and LGBM, which ensures it's the best classifier among the three classifiers and that's due to the ability of KNN which works better severe volatility and noisy random characteristics. When investigating the coins individually as separate data sets, the results reveal that SVM provided the most accuracy in forecasting Bitcoin, while LGBM is accurate in forecasting Ethereum and Litecoin. This concludes that KNN is the best forecasting set of cryptocurrencies in general, while if we want to forecast each currency individually, then SVM and LGBM show better results. This study used several ML algorithms to classify BTC, ETH, and LTC movements as "high" or "low" Sensitivity analysis to discuss the used hyperparameters for each classifier. Future work may hold more and different features when considering crypto-currency datasets, such as exchange rates, interest rates, oil, and dollar prices. As well as test deep learning algorithms to anticipate the movement of cryptocurrencies. More algorithms can be pitted against the ones we used on a greater number of coins considering several security and privacy issues [54].

References

[1] B. Aiden and O. Mason, "Cryptocurrency and the future currency in the United States of America," *Journal of Finance and Accounting*, vol. 5, no. 2, pp. 10–17, 2021.

[2] S. Lahmiri and S. Bekiros, "Intelligent forecasting with machine learning trading systems in chaotic intraday bitcoin market. Chaos," *Solitons & Fractals*, vol. 133, 109641, 2020.

[3] H. Pabuçcu, S. Ongan, and A. Ongan. "Forecasting the movements of bitcoin prices: An application of machine learning algorithms," *Quantitative Finance and Economics*, vol. 4, no. 4, pp. 679–692, 2020.

[4] A. Ganapathy, M. Redwanuzzaman, M. M. Rahaman, and W. Khan, "Artificial intelligence-driven cryptocurrencies," *Global Disclosure of Economics and Business*, vol. 9, no. 2, pp. 107–118, 2020.

[5] S. Vadlamudi. "Agri-food system and artificial intelligence: Reconsidering imperishability," *Asian Journal of Applied Science and Engineering*, vo. 7, pp. 33–42, 2018.

[6] R. Chowdhury, M. A. Rahman, M. S. Rahman, and M. Mahdy, "An approach to predict and forecast the price of constituents and cryptocurrency index using machine learning," *Physica A: Statistical Mechanics and Its Applications*, vol. 551, 124569, 2020.

[7] H. Paruchuri, "Credit card fraud detection using machine learning: A systematic literature review," *ABC Journal of Advanced Research*, vol. 6, no. 2, pp. 113–120, 2017.

[8] Q. A. Al-Haija, "Time-series analysis of cryptocurrency price: Bitcoin as a case study," in *Proc. 2022 International Conference on Electrical Engineering, Computer and Information Technology (ICEECIT)*, Jember, Indonesia, 2022, pp. 49–53, doi: 10.1109/ICEECIT55908.2022.10030536

[9] B. Chen, F. Wei, and C. Gu, "Bitcoin theft detection based on supervised machine learning algorithms," *Security and Communication Networks*, 6643763, 2021.

[10] A. A. Badawi and Q. A. Al-Haija, "Detection of money laundering in bitcoin transactions," in *Proc. 4th Smart Cities Symposium (SCS 2021), Online Conference*, Bahrain, 2021, pp. 458–464, doi: 10.1049/icp.2022.0387

[11] X. Sun, M. Liu, and Z. Sima, "A novel cryptocurrency price trend forecasting model based on lightGBM," *Finance Research Letters*, vol. 32, 101084, 2020.

[12] W. Bakry, A. Rashid, S. Al-Mohamad, and N. El-Kanj, "Bitcoin and portfolio diversification: A portfolio optimization approach," *Journal of Risk Financial Management*, vol. 14, 282, 2021, doi: 10.3390/jrfm14070282

[13] R. Farell, "An analysis of the cryptocurrency industry," *Wharton Research Scholars Journal*, University of Pennsylvania, USA, 2015.

[14] D. Lee, K. Chuen, L. Guo, and Y. Wang, "Cryptocurrency: A new investment opportunity," *Journal of Alternative Investments*, vol. 20, no. 3, pp. 16–40, 2017.

[15] J. Seys and K. Decaestecker, "The evolution of bitcoin price drivers: Moving towards stability," Master' thesis, University of Ghent, Gent, Belgium, 2016.

[16] D. Vujičić, D. Jagodić and S. Ranđić, "Blockchain technology, bitcoin, and Ethereum: A brief overview," in *Proc. 2018 17th International Symposium INFOTEH-JAHORINA (INFOTEH)*, East Sarajevo, Bosnia and Herzegovina, 2018, pp. 1–6, doi: 10.1109/INFOTEH.2018.8345547

[17] P. Katsiampa, S. Corbet, and B. Lucey, "Volatility spillover effects in leading cryptocurrencies: A BEKK-Mgarch analysis," *Finance Research Letters*, vol. 29, pp. 68–74, 2019.

[18] M. A Kumar and M. Gopal, "Least squares twin support vector machines for pattern classification," *Expert Systems with Applications*, vol. 36, no. 4, pp. 7535–7543, 2009.

[19] A. Soualhi, K. Medjaher, and N. Zerhouni, "Bearing health monitoring based on Hilbert–Huang transform, support vector machine, and re-aggression," *IEEE Transactions on Instrumentation and Measurement*, vol. 64, no. 1, pp. 52–62, 2014.

[20] Y. Geng, J. Chen, R. Fu, G. Bao, and K. Pahlavan, "Enlighten wearable physiological monitoring systems: On-body of characteristics based human motion classification using a support vector machine," *IEEE Transactions on Mobile Computing*, vol. 15, no. 3, pp. 656–671, 2015.

[21] M. Z. Ashi, M. Alnabhan, Q. A. Al-Haija, "Effective one-class classifier model for memory dump malware detection," *Journal of Sensor and Actuator Networks*, vol. 12, no. 1, 5, 2023, doi: 10.3390/jsan12010005

[22] S. Garcia, J. Derrac, J. Cano, and F. Herrera, "Prototype selection for nearest neighbor classification: Taxonomy and empirical study," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 417–435, 2012.

[23] K. Kostková, L. Omelina, P. Kyčina, and P. Jamrich, "An introduction to load management," *Electric Power Systems Research*, vol. 95, pp. 184–191, 2013.

[24] J. Mohtasham, "Renewable energies," *Energy Procedia*, vol. 74, pp. 1289–1297, 2015.

[25] S. McNally, J. Roche, and S. Caton, "Predicting the price of bitcoin using machine learning," in *Proc. 2018 26th Euro Micro International Conference on Parallel, Distributed, and Network-Based Processing (PDP)*, IEEE, 2018, pp. 339–343.
[26] S. Abu-Zaideh, M. A. Snober, Q. A. Al-Haija, "Smart boosted model for behavior-based malware analysis and detection," *Lecture Notes in Networks and Systems*, vol. 528, Springer, 2023, doi: 10.1007/978-981-19-5845-8_58
[27] J. Patel, S. Shah, P. Thakkar, and K. Kotecha, "Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques," *Expert Systems with Applications*, vol. 42, no. 1, pp. 259–268, 2015.
[28] D. Mallqui and R. Fernandes, "Predicting the direction, maximum, minimum, and closing prices of daily bitcoin exchange rate using machine learning techniques," *Applied Soft Computing*, vol. 75, pp. 596–606, 2019.
[29] K. Żbikowski, "Application of machine learning algorithms for bitcoin automated trading," *Machine Intelligence and Big Data in Industry*, vol. 161, no. 8, 2016.
[30] E. Akyildirim, O. Cepni, S. Corbet, and G. S. Uddin, "Forecasting mid-price movement of Bitcoin futures using machine learning," *Annals of Operations Research*, pp. 1–32, 2021.
[31] N. A. Hitam and A. R. Ismail, "Comparative performance of machine learning algorithms for cryptocurrency forecasting," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 11, no. 3, pp. 1121–1128, 2018.
[32] S. Saadah and A. A. A Whafa, "It monitors financial stability based on a prediction of cryptocurrencies price using an intelligent algorithm," in *Proc. 2020 International Conference on Data Science and Its Applications (ICoDSA)*, IEEE, 2020, pp. 1–10.
[33] A. Barnwal, H. P. Bharti, A. Ali, and V. Singh, "I am stacking with neural networks for cryptocurrency investment," in *Proc. 2019 New York scientific Data Summit (NYSDS)*, IEEE, 2019, pp. 1–5.
[34] Q. A. Al-Haija, M. Alnabhan, E. Saleh, and M. Al-Omari, "Applications of blockchain technology for improving security in the internet of things (IoT)," in *Proc. Blockchain Technology Solutions for the Security of IoT-Based Healthcare Systems*, 2023, pp. 199–221.
[35] S. A. Alahmari, "Using nonlinear machine learning algorithms to predict the price of cryptocurrencies," *International Journal of Future Generation Communication and Networking*, vol. 13, no. 1, pp. 745–752, 2020.
[36] M. A. Razi and K. Athappilly, "A comparative predictive analysis of Neural Networks (NNS), nonlinear regression, and classification and regression tree (cart) models," *Expert Systems with Applications*, vol. 29, no. 1, pp. 65–74, 2005.
[37] S. Lahmiri and S. Bekiros, "Cryptocurrency forecasting with deep learning chaotic neural networks," *Chaos, Solitons & Fractals*, vol. 118, pp. 35–40, 2019.
[38] A. Greaves and B. Au, "Using the bitcoin transaction graph to predict the price of bitcoin," *No Data*, vol. 8, pp. 416–443, 2015.
[39] W. Zhu, *et al.*, "Real-time prediction of bitcoin bubble crashes," *Physica A: Statistical Mechanics and Its Applications*, vol. 548, 124477, 2020.
[40] W. Yao, K. Xu, and Q. Li, "Exploring the influence of news articles on bitcoin price with machine learning," in *Proc. 2019 IEEE Symposium on Computers and Communications (ISCC)*, IEEE, 2019, pp. 1–6.
[41] J. E. Butner, A. K. Munion, B. Baucom, and A. Wong, "Ghost hunting in the nonlinear dynamic machine," *PloS One*, vol. 14, no. 12, e0226572, 2019.
[42] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and machine learning forecasting methods: Concerns and ways forward," *PloS One*, vol. 13, no. 3, e0194889, 2018.
[43] T. Cover and P Hart, "Nearest neighbor pattern classification," *IEEE transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.
[44] D. Zhang, G. Yang, F. Li, J. Wang, and A. K. Sangaiah, "Detecting seam-carved images using uniform local binary patterns," *Multimedia Tools and Applications*, vol. 79, no. 13, pp. 8415–8430, 2020.
[45] A. Basudhar and S. Missoum, "An improved adaptive sampling scheme for the construction of explicit boundaries," *Structural and Multidisciplinary Optimization*, vol. 42, no. 4, pp. 517–529, 2010.
[46] K. Tan, J. Zhang, Q. Du, and X. Wang, "GPU parallel implementation of support vector machines for hyperspectral image classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 10, pp. 4647–4656, 2015.
[47] R. Huerta, F. Corbacho, and C. C. Elkan, "Nonlinear support vector machines can systematically identify stocks with high and low future returns," *Algorithmic Finance*, vol. 2, no. 1, pp. 45–58, 2013.
[48] T. Hacib *et al*., "Support vector machines for measuring dielectric properties of materials," *COMPEL-The International Journal for Computation and Mathematics in Electrical and Electronic Engineering*, vol. 29, no. 4, pp. 1081–1089, 2010.
[49] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Y. Liu, "LightGBM: A highly efficient gradient-boosting decision tree," *Advances in Neural Information Processing Systems*, vol. 30, pp. 3146–3154, 2017.
[50] Z. Keskin and T. Aste, "Information-theoretic measures for nonlinear causality detection: Application to social media sentiment and cryptocurrency prices," *Royal Society Open Science*, vol. 7, no. 9, 200863, 2020.
[51] R. A. Davis, A. J. Charlton, J. Godward, S. A. Jones, M. Harrison, and J. C. Wilson, "Adaptive binning: An improved binning method for metabolomics data using the undecimated wavelet transform," *Chemometrics and Intelligent Laboratory Systems*, vol. 85, no. 1, pp. 144–154, 2007.
[52] Y. Kara, M. A. Boyacioglu *et al*., "The pre-dictating direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul stock exchange," *Exper. Systemsms. with Applications*, vol. 38, no. 5, pp. 5311–5319, 2011.
[53] K. Albulayhi and Q. A. Al-Haija, "Security and privacy challenges in blockchain application," in *Proc. The Data-Driven Blockchain Ecosystem*, 2022, pp. 207–226.
[54] G. Bontempi, S. B. Taieb, and Y. A. LeBorgne, "Machine learning strategies for time series forecasting," in *Proc. Business Intelligence, eBISS 2012, Lecture Notes in Business Information Processing*, Springer, Berlin, Heidelberg, 2012, vol. 138, pp. 62–77.

Copyright © 2023 by the authors. This is an open access article distributed under the Creative Commons Attribution License (CC BY-NC-ND 4.0), which permits use, distribution and reproduction in any medium, provided that the article is properly cited, the use is non-commercial and no modifications or adaptations are made.