DOI: https://doi.org/10.4038/sljas.v26i2.8174

Sri Lankan Journal of Applied Statistics, Vol (26-2)

A Clustering-Based Machine Learning Approach to Forecast Cryptocurrency Prices

S. N. Kalutharage* and H. A. Pathberiya

Department of Statistics, University of Sri Jayewardenepura, Sri Lanka

*Corresponding author: sachinisansala2@gmail.com

Received: 19th May, 2025/ Revised: 29th June, 2025/ Published: 20th July, 2025

ABSTRACT

Forecasting cryptocurrency prices is challenging due to their highly vola–tile and unpredictable nature. This study presents a high-frequency forecasting model that combines changepoint analysis with clustering-based machine lear–ning techniques. The pruned exact linear time (PELT) method was applied to intraday historical data from Bitcoin (BTC), Ethereum (ETH), and Cardano (ADA) for a one-year period, from August 2022 to August 2023, to detect volatility shifts, dividing the data into distinct regions. These regions were then grouped using the K-means algorithm based on characteristics like price and volume variance. Support Vector Regression (SVR), Long Short-Term Memory (LSTM), and Random Forest (RF) models were trained on both clustered and non-clustered data to predict cryptocurrency closing prices. The results showed that using clustered original data improved forecasting accuracy in most cases. For ADA, SVR's Mean Absolute Error (MAE) dropped from 0.0018 to 0.0003, and for RF, MAE improved from 0.0010 to 0.0004. Similar improvements were observed for BTC and ETH. These results show that clustering volatility regions enables machine learning models to more accurately capture price dynamics, resulting in better and more reliable forecasts. This study demonstrates the effectiveness of integrating changepoint-based clustering with machine learning models to improve short-term cryptocurrency price forecasts.

Keywords: Clustering, Cryptocurrency Forecasting, Machine learning

1 Introduction

Digital currencies, commonly known as cryptocurrencies, have emerged fast after the introduction of Bitcoin in 2009. Following its success, numerous alternative digital assets, such as Ethereum, Tether, Binance Coin, Ripple, and

Cardano have entered the market. Currently, thousands of cryptocurrencies exist, each with unique characteristics and applications, significantly transforming the modern financial landscape (Frankenfield, 2023; Yahoo Finance, 2024).

Interest in cryptocurrencies is still rising, even in places like Sri Lanka, where regulations are stricter. The informal use of cryptocurrencies remains common because of the possibility of large returns and quick, inexpensive international transactions (Fidelity, 2023; Handagama, 2021; Yasar, 2023), even though the Central Bank's official stance against recognizing crypto assets (Central Bank of Sri Lanka, 2023; Sewmini, 2023). The Central Bank's consideration of regulatory frameworks further highlights the increasing importance of crypto in both local and global contexts (Chandrasekara, 2024).

In this evolving environment, the increasing popularity and use of cryptocurrencies, especially in areas like investment and online finance, have opened up many new research opportunities. Among them, effective price forecasting is significant in navigating the unpredictable and volatile nature of the market (DB Investing, 2023; Frankenfield, 2023). Making accurate predictions is extremely beneficial for investors, traders, and researchers, as it can support better decision-making, improve risk management, and help develop more effective trading strategies. As the cryptocurrency market continues to grow and change rapidly, having more reliable forecasting methods becomes increasingly important.

In response to this growing need, various studies have explored traditional time series models such as autoregressive integrated moving average (ARIMA) and generalized auto regressive conditional heteroskedasticity (GARCH) for cryptocurrency price prediction. While these models offer a structured framework for analyzing temporal data, their limitations in capturing non-linear and complex relationships have been noted (Bitto et al., 2022; Roy, Nanjiba & Chakrabarty, 2018; Wang, 2022). While some researchers found that ARIMA to outperforms machine learning models in particular contexts (Iqbal et al., 2021; Šťastný et al., 2022), others highlighted the shortcomings of time series models in volatile environments due to their sensitivity to outliers and inability to adapt to sudden changes (Basu, 2020). In contrast, machine learning techniques including, support vector machines (SVM), random forest (RF), and deep learning architectures like long short-term memory (LSTM) and convolutional neural network (CNN) have shown improved performance in capturing dynamic patterns and enhancing forecasting accuracy (Ngai et al., 2023; Oyedele et al., 2022; Pabuccu, Ongan & Ongan, 2020). Despite the success of machine learning methods, few studies have explored clustering to refine prediction accuracy further.

Clustering can uncover hidden structures in volatile time series data by grouping regions with similar volatility behaviors, potentially allowing predictive models to specialize across different regimes (Das et al., 2023; Lorenzo & Arroyo, 2022; Šťastný et al., 2022). However, despite the increasing application of machine learning in cryptocurrency forecasting, the integration of clustering techniques into these models remains relatively underexplored. The gap in the literature highlights the need for further exploration into how clustering can be used to enhance predictive performance in highly volatile markets.

In response to the limited integration of clustering techniques within machine learning models for cryptocurrency forecasting, the present study aims to develop a framework for forecasting high-frequency cryptocurrency prices by analyzing and clustering intraday return volatility patterns. The study focuses on investigating the behavior of return volatility across different time regions. The study identifies suitable features for representing these volatility patterns and defines meaningful clusters based on these characteristics. Finally, it develops a machine learning-based forecasting model that integrates clustering insights.

The remainder of this paper is organized as follows: Section 2 outlines the methodology, detailing the processes of data collection, preprocessing, changepoint detection, clustering, and model development. Section 3 presents results and discusses the results obtained from the analysis and model evaluations. Conclusions from the study are remarked in Section 4.

2 Methodology

This study collected high-frequency cryptocurrency data for Bitcoin, Ethereum, and Cardano. After preprocessing, changepoint analysis was used to segment volatility regions, which were then clustered using K-means. Machine learning models SVR, RF, and LSTM were trained on non-clustered and clustered data to forecast closing prices. Figure 1 presents a summary of the study design.

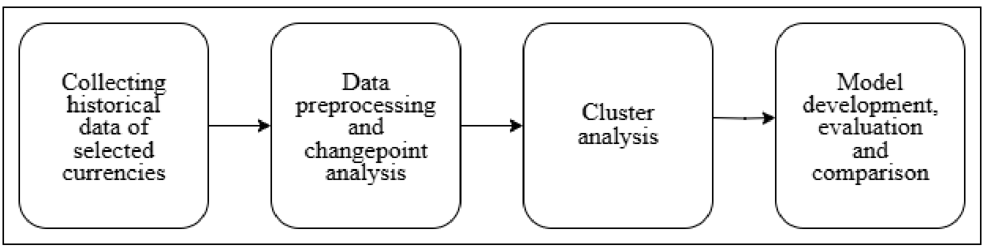

Figure 1: Summary of the methodology framework

2.1 Source of data

Historical high-frequency cryptocurrency data for Bitcoin (BTC), Ethereum (ETH), and Cardano (ADA) were collected from JForex4, a multi-asset trading platform developed by the Swiss bank Dukascopy (JForex4, 2018-2024). The data spans from August 2022 to August 2023 and includes variables such as Time (in UTC), Close, Open, High, Low, and Volume. These datasets were obtained in 10-minute intervals, resulting in 144 data points per asset per day. Over the one-year, this amounted to approximately 52,560 data points per cryptocurrency.

2.2 Data preprocessing

Data were rearranged at fixed 10-minute intervals to capture detailed temporal fluctuations in market behavior. To enable focused intra-day analysis, a ”Day-Index” variable was introduced, segmenting the continuous time series into discrete daily units. This daily segmentation ensures that volatility patterns and changepoints are detected within consistent and comparable timeframes, reflecting the natural rhythm of trading activity.

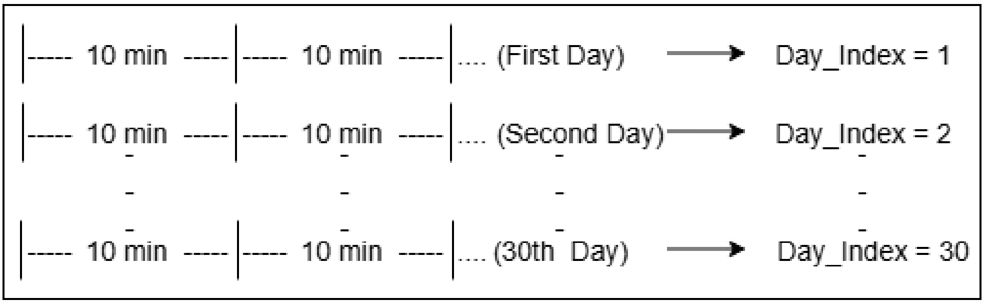

Figure 2: Daily segmentation of high frequency crypto data via Day-Index

Preprocessing also included checking for missing values, which were not present. The ”Date” component was extracted from the Time column, and the time zone was standardized to UTC. A ”Day-Index” variable was created to track daily segments. Return values were calculated using:

$$\text{Return} = \frac{Y(t) - Y(t-1)}{Y(t-1)} \times 100 \tag{1}$$

Where Y(t) is the closing price at time t.

Variables were converted to appropriate data types, with the variable Day-Index set as a character variable and the remaining variables converted to numeric type. These steps laid the foundation for later segmentation and analysis.

2.3 Changepoint Analysis

Changepoint analysis was performed to identify structural shifts in return volatility throughout each trading day. The Pruned Exact Linear Time (PELT) method was employed with a manually specified penalty to guide the change point detection process.

$$1.5 \times \log\left(\text{length(daily\_data\$Return)}\right) \tag{2}$$

The Pruned Exact Linear Time (PELT) method was employed due to its exact optimization capability and superior computational efficiency, making it well-suited for detecting variance changes in large-scale, high-frequency data such as cryptocurrency returns. A custom penalty term was specified to guide the change point detection process. This value helps control the sensitivity of the detection, ensuring the analysis is neither too sensitive to minor fluctuations nor too conservative. The choice enhances the reliability and validity of the detected change points. The multiplier 1.5 was chosen to reduce false positives in a high-volatility environment, after testing showed it provided more interpretable and stable segmentations compared to standard penalties. (Killick, Fearnhead & Eckley, 2012).

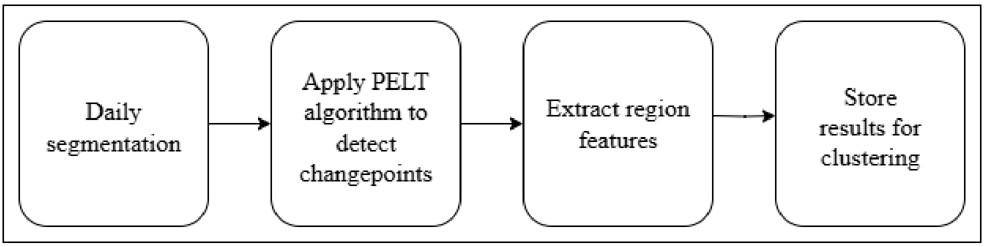

Figure 3: Workflow of Changepoint Analysis Process

Each trading day was divided into regions of stable behavior based on changepoints. For each region, features such as length, return variance, and variances of high, low, and volume were computed and stored the results in a comprehensive data frame. Figure 3 represents the main steps that follow in changepoint analysis.

2.4 Clustering

The regions identified through a changepoint analysis were further examined by calculating key features for each segment, including region length (number of data points/days between two detected change points), return variance, high variance, low variance, and volume variance. These features formed the basis for clustering. K-means clustering was applied to group regions with similar volatility characteristics. Its computational efficiency and alignment with the numeric nature of the features make it well suited for capturing structural similarities in cryptocurrency market regions. The optimal number of clusters was selected using the elbow method. Distance was measured using Euclidean distance. Euclidean distance was chosen for change point detection because it efficiently captures magnitude shifts in univariate cryptocurrency returns and is computationally simple for high-frequency data. When p and q are two vectors of length n, then the Euclidean distance between two points in space is given by,

$$d_{\text{euc}}(p, q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \tag{3}$$

All variables were standardized using z-score normalization. The R packages factoextra and base plotting functions were used for cluster visualization and validation. Figure 4 represents the basic steps followed in the study regarding cluster analysis.

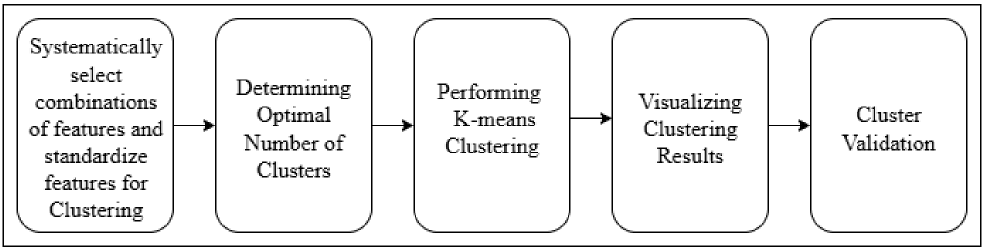

Figure 4: Procedural Steps in Cluster Analysis

2.5 Upgraded Dataset

After clustering, the dataset was incrementally enhanced with newly generated variables from return calculations, changepoint regions, and cluster assignments. This process produced a more comprehensive dataset, allowing for refined model development by including region-based insights and cluster memberships as additional predictive features.

2.6 Machine Learning Model Development

Three machine learning models, namely SVR, RF, and LSTM were used to forecast cryptocurrency closing prices. Each model was trained on:

- **T**he non-clustered dataset – Original dataset with derived features, used as the benchmark.
- **C**lustered datasets – Data segmented into clusters based on volatility features.
- **S**elected largest subsets from each cluster – Selected continuous subsets from clusters to preserve time-series structure.

I. Support Vector Regression (SVR)

Support Vector Regression seeks to find a function that fits the data within a specified margin of tolerance (ε) while simultaneously minimizing model complexity to prevent overfitting. This approach balances accuracy with generalizability, making it suitable for capturing the underlying trends in high-frequency financial data. The objective function is:

$$\min \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}|\xi_i| \tag{4}$$

Constraints:

$$|y_i - w_i x_i| \leq \varepsilon + |\xi_i| \tag{5}$$

Where, ε is the maximum error and ξ_i is the deviation from the maximum error. Implemented using the e1071 package in R, SVR used an RBF kernel (Sharp, 2020). A non-linear kernel was selected to capture the complex, non-linear patterns in volatile cryptocurrency data. The RBF kernel was preferred for its flexibility, generalization ability, and strong performance in high-frequency financial forecasting (Lahmiri, Bekiros & Bezzina, 2022). The dataset was chronologically split, with the most recent 20% reserved as the test set and the preceding 80% used for training. Lagged features were included to account for temporal dependencies.

II. Random Forest (RF)

Random Forest is an ensemble learning method that aggregates predictions from multiple decision trees. Hyperparameters like the number of trees, maximum depth, and minimum samples for splitting were optimized. Feature scaling and lagged variables were included (Cloud Software Group, Inc, 2024). The data was split in the same manner as for SVR, with preceding 80% allocated for training and most recent 20% reserved for testing.

III. Long Short-Term Memory (LSTM)

LSTM, a recurrent neural network variant, is well-suited for time series forecasting. Key hyperparameters included the number of units, dropout rate, and learning rate. Sequential windowing and lagged features were used to model temporal dependencies (Eckhardt, 2018). Implemented in R with keras and tensorflow, the LSTM was trained on preceding 80% of the data and tested on the most recent 20%.

2.7 Model Evaluation

Model performance was evaluated using:

- **M**ean absolute error (MAE)
- **M**ean absolute percentage error (MAPE)
- **C**oefficient of determination (R^2)

Visual inspection and performance metrics were used to compare predictions across the non-clustered, clustered, and cluster-subset datasets. The non-clustered dataset served as the benchmark for evaluating the effect of clustering on prediction accuracy.

This structured and layered methodology will lead to the development of forecasting models that integrated volatility-based clustering insights into high-frequency cryptocurrency price prediction.

3 Results and Discussion

This chapter presents and interprets the results obtained from the study in alignment with the research objectives. The outcomes are arranged based on the stages of analysis.

3.1 Descriptive Analysis of Cryptocurrency Data

Initial visualizations of the closing price and return series for ADA, BTC, and ETH over the study period are provided to understand general price movements and volatility.

Figure 5 presents the time series plots of close prices for BTC, while Figure 6 displays the corresponding return series. Similar figures were obtained for ADA and ETH. The return plots highlight the presence of high-frequency fluctuations and volatility clustering, which justifies the use of changepoint analysis and clustering techniques.

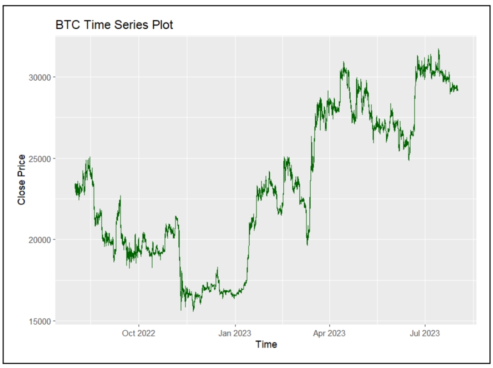

Figure 5: Time Series Plot for BTC Close Price

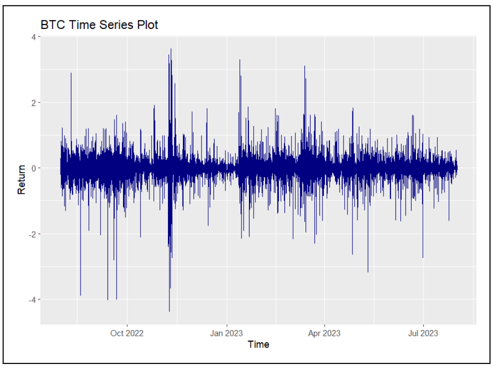

Figure 6: Time Series Plot for BTC Return

3.2 Changepoint Detection Results

Changepoint analysis was applied to the high-frequency return series of ADA, BTC, and ETH to analyze intraday volatility patterns. This technique effectively segmented each asset's time series into distinct regions of volatility, providing a strong basis for subsequent clustering and predictive modeling.

To illustrate the results of the changepoint detection, Figures 7 and 8 display two sample daily return plots for ADA, while Figures 9 and 10 and Figures 11 and 12 present corresponding examples for BTC and ETH, respectively.

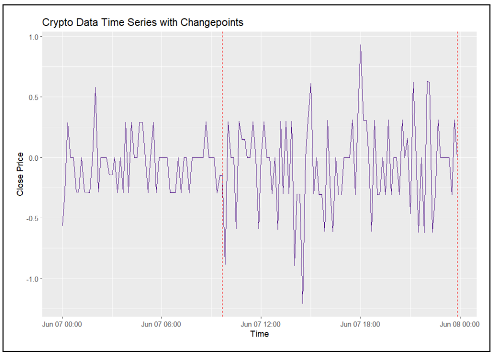

Figure 7: Time Series Plots with Changepoints for ADA on June 07th

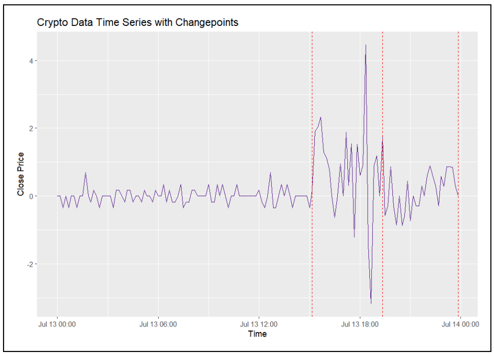

Figure 8: Time Series Plots with Changepoints for ADA on July 13th

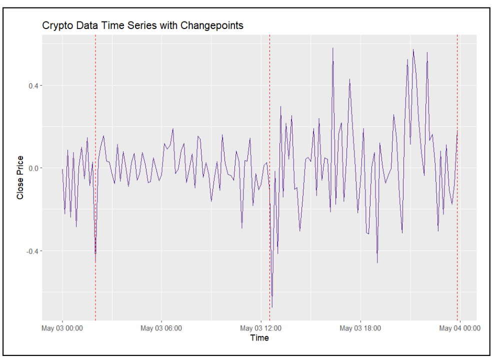

Figure 9: Time Series Plots with Changepoints for BTC on May 3rd

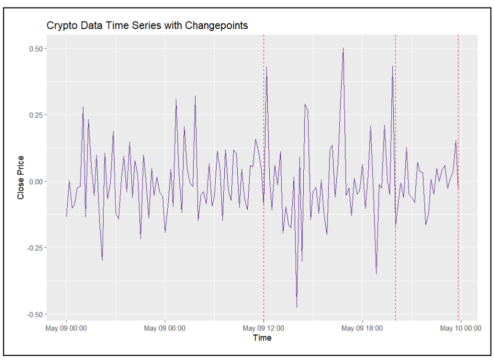

Figure 10: Time Series Plots with Changepoints for BTC on May 9th

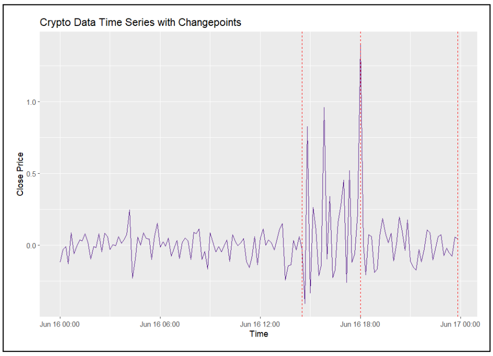

Figure 11: Time Series Plots with Changepoints for ETH on June 16th

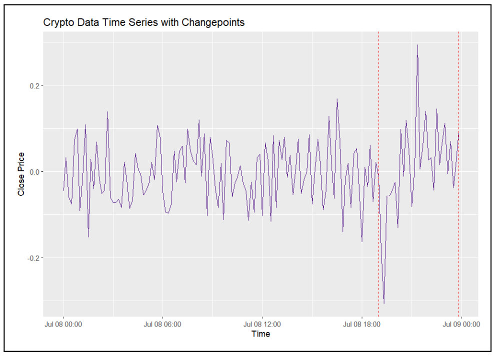

Figure 12: Time Series Plots with Changepoints for RTH on July 8th

Each figure set illustrates how the PELT method identified changepoints in the return series, segmenting each cryptocurrency's intraday data into regions of relatively stable volatility and highlighting abrupt shifts within a single trading day. The segmented regions identified through changepoint analysis served as the foundation for feature extraction. For each region, key volatility characteristics such as region length, return variance, volume variance, high price variance, and low-price variance were calculated, forming the basis for the clustering phase that followed.

3.3 Cluster Analysis

K-means clustering was applied to each dataset using features extracted from the segmented changepoint regions, including return variance, high and low price variance, volume variance, and region length. The elbow method was employed to determine the optimal number of clusters, which was identified as three across all three cryptocurrencies.

The clustering outcomes are summarized in Table 1 below, showcasing the most successful results for each cryptocurrency. Validation metrics such as Total Within-Cluster Sum of Squares (TWSS) and Between-Cluster Sum of Squares (BCSS) further support the robustness of these solutions.

Table 1: Optimal clustering results and evaluation metrics for each cryptocurrency

Cryptocurrency	**Feature Selection**	**Cluster Evaluation**	
		TWSS	**BCSS**
ADA	Volume variance, High price variance	3870.836	6487.164

Continued on next page

Table 1 – continued from previous page

Cryptocurrency	Feature Selection	Cluster Evaluation	
		TWSS	BCSS
	Volume variance, Low price variance	4101.625	6256.375
BTC	Volume variance, Close price variance	1544.712	2583.288
ETH	Close price variance, Region length	1287.907	2732.093
	Volume variance, Region length	1480.872	2539.128

The corresponding visual representations illustrate the most effective cluster formations, emphasizing how volatility patterns were meaningfully grouped using selected feature combinations.

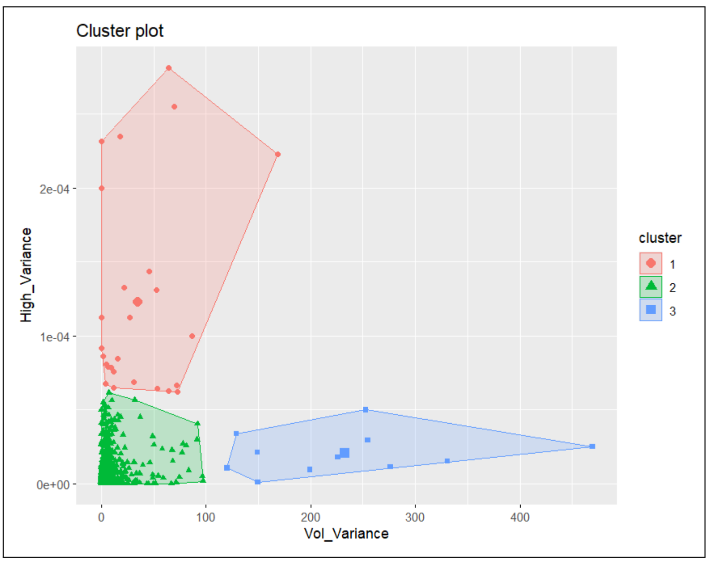

Figure 13: Cluster visualization for ADA using volume variance and high price variance

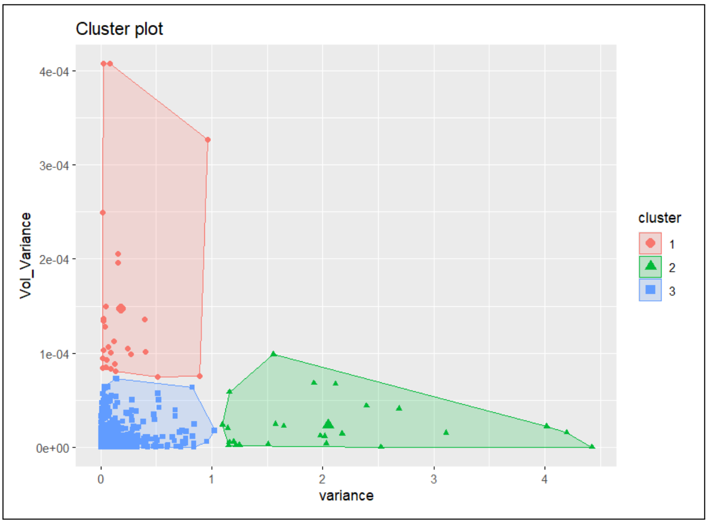

Figure 14: Cluster visualization for BTC using volume variance and close price variance

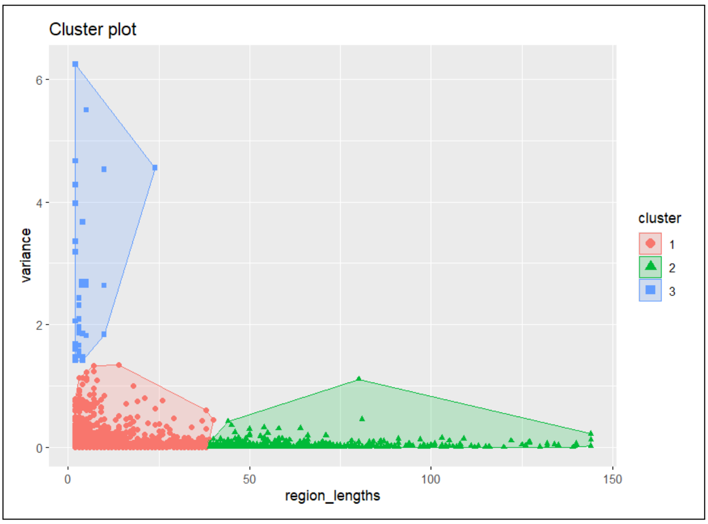

Figure 15: Cluster visualization for ETH using close price variance and region length

These findings confirm that distinct volatility structures can be effectively captured through K-means clustering, offering valuable segmentation for subsequent modeling efforts.

3.4 Forecasting Model Results

This section presents the forecasting outcomes obtained from training machine learning models. The three machine learning models selected for the study, namely SVR, RF, and LSTM, were trained and evaluated on three types of datasets, specifically the non-clustered benchmark datasets, the clustered datasets segmented according to the optimal cluster results, and the selected clustered subsets. While an initial model was trained using the full set of input features, for each dataset, the final models used a subset of features selected based on performance optimization. This approach ensured that the most relevant features were used for each data segment, improving model accuracy while acknowledging that the feature sets may differ across datasets.

The top five model configurations were documented following hyperparameter tuning and feature selection considering each cryptocurrency. Model performance was evaluated on the test data using standard metrics: Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and R-squared (R^2).

SVR models demonstrated notable improvements in predictive accuracy when trained on clustered data for all three cryptocurrencies. Table 2 presents the best predictive results obtained from the SVR model for each dataset across ADA, BTC, and ETH.

Table 2: Evaluation metrics for SVR across ADA, BTC, and ETH

Cryptocurrency	**Data Set**	**MAE**	**R^2**	**MAPE**
ADA	Non-clustered Data Set	0.0018	0.9612	0.5590
	Cluster-1	0.0008	0.9987	0.2304
	Cluster-2	0.0016	0.9948	0.4969
	Cluster-2-Subset 1	0.0009	0.9670	0.2531
	Cluster-3	0.0009	0.9960	0.1575
	Cluster-3-Subset 1	0.0003	0.9977	0.0823
	Cluster-3-Subset 2	0.0004	0.9968	0.1288
BTC	Non-clustered Data Set	27.0674	0.9918	0.0899
	Cluster-1	38.2536	0.9969	0.1332
	Cluster-1-Subset 1	18.3610	0.9498	0.0652
	Cluster-2	131.084	0.9957	0.5100
	Cluster-3	19.5162	0.9971	0.0647
	Cluster-3-Subset 1	42.3240	0.9937	0.2453
	Cluster-3-Subset 2	12.8647	0.9986	0.0426
ETH	Non-clustered Data Set	2.2837	0.9937	0.1204
	Cluster-1	2.6436	0.9961	0.1464
	Cluster-1-Subset 1	3.5403	0.9867	0.2348
	Cluster-1-Subset 2	1.7215	0.8448	0.0866

Continued on next page

Table 2 – continued from previous page

Cryptocurrency	Data Set	MAE	R^2	MAPE
	Cluster-2	4.1991	0.9944	0.2276
	Cluster-2-Subset 1	1.2913	0.9241	0.0747
	Cluster-2-Subset 2	0.3720	0.9908	0.0259
	Cluster-3	11.5008	0.9924	0.7790

In the case of ADA, the Mean Absolute Error (MAE) dropped from 0.0018 (non-clustered) to 0.0003 in the best-performing subset. Similarly, ETH exhibited a sharp decline in MAE from 2.2837 to 0.3720, and BTC from 27.0674 to 12.8647. These improvements were accompanied by increases in R^2, and decreases in MAPE. Collectively these results affirm that SVR models benefit significantly from the data segmentation introduced via clustering, which enables the model to capture distinct volatility behaviors.

Table 3 summarizes the best results achieved using the Random Forest (RF) model across datasets for ADA, BTC, and ETH.

Table 3: Evaluation metrics for RF across ADA, BTC, and ETH

Cryptocurrency	Data Set	MAE	R^2	MAPE
ADA	Non-clustered Data Set	0.0010	0.9484	0.3062
	Cluster-1	0.0116	0.8332	3.1770
	Cluster-2	0.0032	0.9665	0.9010
	Cluster-2-Subset 1	0.0014	0.8956	0.4025
	Cluster-3	0.0009	0.9657	0.2965
	Cluster-3-Subset 1	0.0007	0.9866	0.2032
	Cluster-3-Subset 2	0.0004	0.9972	0.1323
BTC	Non-clustered Data Set	23.2610	0.9843	0.0757
	Cluster-1	79.5847	0.9905	0.2751
	Cluster-1-Subset 1	21.6599	0.9250	0.0769
	Cluster-2	158.5849	0.9771	0.5957
	Cluster-3	25.2589	0.9832	0.0822
	Cluster-3-Subset 1	37.8688	0.9915	0.1934
	Cluster-3-Subset 2	22.4061	0.9833	0.0727
ETH	Non-clustered Data Set	1.4566	0.9937	0.0763
	Cluster-1	3.0968	0.9554	0.1831
	Cluster-1-Subset 1	3.2523	0.9912	0.2150
	Cluster-1-Subset 2	1.7769	0.8454	0.0894

Continued on next page

Table 3 – continued from previous page

Cryptocurrency	Data Set	MAE	R^2	MAPE
	Cluster-2	2.2188	0.9850	0.1268
	Cluster-2-Subset 1	1.5166	0.8899	0.0877
	Cluster-2-Subset 2	0.7359	0.9488	0.0512
	Cluster-3	19.3011	0.9648	1.1988

RF models also exhibited performance gains when trained on clustered data, although the improvements were less pronounced compared to SVR. For ADA, the MAE improved from 0.0010 to 0.0004, with a corresponding increase in R^2 from 0.9484 to 0.9972, and a significant reduction in MAPE from 0.3062 to 0.1323. ETH's MAE decreased from 1.4566 to 0.7359, alongside improved R^2 and a notable drop in MAPE. For BTC, a modest improvement was observed, reducing MAE from 23.2610 to 21.6599, while R^2 remained high and MAPE showed slight improvement. Although Random Forest models exhibit strong predictive capabilities even with unsegmented data, the results demonstrate that isolating and modeling volatility regions through clustering further enhances their performance, leading to more accurate and robust forecasts.

Table 4 records the best-performing results from the LSTM model for each dataset, emphasizing the top subsets selected for ADA, BTC, and ETH.

Table 4: Evaluation metrics for LSTM across ADA, BTC, and ETH

Cryptocurrency	Data Set	MAE	R^2	MAPE
ADA	Non-clustered Data Set	0.0005	0.9986	0.1479
	Cluster-2-Subset 1	0.0008	0.9791	0.2288
	Cluster-3-Subset 1	0.0007	0.9964	0.2016
	Cluster-3-Subset 2	0.0007	0.9835	0.2252
BTC	Non-clustered Data Set	29.2393	0.9947	0.0970
	Cluster-1-Subset 1	10.7100	0.9794	0.0380
	Cluster-3-Subset 1	27.5621	0.9965	0.1387
	Cluster-3-Subset 2	21.6797	0.9980	0.0722
ETH	Non-clustered Data Set	3.0901	0.9895	0.1629
	Cluster-1-Subset 1	16.3096	0.7291	1.0907
	Cluster-1-Subset 2	0.8542	0.9659	0.0430
	Cluster-2-Subset 1	0.6973	0.9813	0.0402
	Cluster-2-Subset 2	0.3532	0.9918	0.0246

LSTM models showed mixed results. For ADA, the non-clustered dataset yielded better performance with the highest R^2 of 0.9986 and lowest MAE

(0.0005) and MAPE (0.1479), indicating strong predictive accuracy without segmentation, whereas clustered data did not improve accuracy. In contrast, ETH showed a significant reduction in MAE from 3.0901 (non-clustered) to 0.3532 (clustered), R^2 improved to 0.9918, and MAPE decreased to 0.0246. Similarly, BTC showed notable improvement in MAE (from 29.2393 to 10.7100) and MAPE (from 0.0970 to 0.0380), although R^2 slightly decreased in some clustered subsets. These findings indicate that while LSTM models can benefit from clustered data in some cases, the effect may vary depending on the cryptocurrency and the structure of the volatility patterns.

Table 5 5 presents a comparative summary of Mean Absolute Error (MAE) results between the benchmark (non-clustered) datasets and the best-performing clustered subsets for each machine learning model (SVR, RF, and LSTM) across ADA, BTC, and ETH. The results demonstrate significant improvements achieved through the clustering-based approach, particularly for ADA and ETH. For these assets, SVR and RF models recorded markedly lower MAE values post-clustering, validating the effectiveness of segmenting volatility patterns. However, in the case of ADA, LSTM did not exhibit improvement with clustering, indicating that the model's performance may be sensitive to the temporal structure of the dataset. In the case of BTC, notable improvements were observed with SVR and LSTM models, while RF showed only modest gains, suggesting that the impact of clustering may vary depending on the asset's volatility behavior and the modeling approach used.

Table 5: MAE Comparison Between Benchmark and Optimized Clustered Models

ML Model	Cryptocurrency	Data Set	MAE
SVR	**ADA**	Non-clustered Data Set	0.0018
		Cluster-3-Subset 1	0.0003
	BTC	Non-clustered Data Set	27.0674
		Cluster-3-Subset 2	12.8647
	ETH	Non-clustered Data Set	2.2837
		Cluster-2-Subset 2	0.3720
RF	**ADA**	Non-clustered Data Set	0.0010
		Cluster-3-Subset 2	0.0004
	BTC	Non-clustered Data Set	23.2610
		Cluster-1-Subset 1	21.6599
	ETH	Non-clustered Data Set	1.4566
		Cluster-2-Subset 2	0.7359
LSTM	**ADA**	Non-clustered Data Set	0.0005
		Cluster-3-Subset 1	0.0007
	BTC	Non-clustered Data Set	29.2393
		Cluster-1-Subset 1	10.7100
	ETH	Non-clustered Data Set	3.0901

Table 5 – continued from previous page

ML Model	Cryptocurrency	Data Set	MAE
		Cluster-2-Subset 2	0.3532

Figures 16, 17, and 18 illustrate visual comparison charts displaying various machine-learning approaches for non-clustered data related to each cryptocurrency.

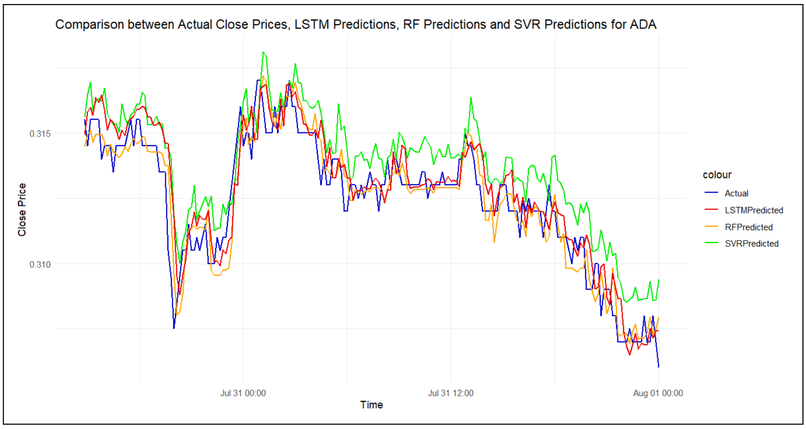

Figure 16: Visual Comparison Between ML Approaches for non-clustered ADA Data

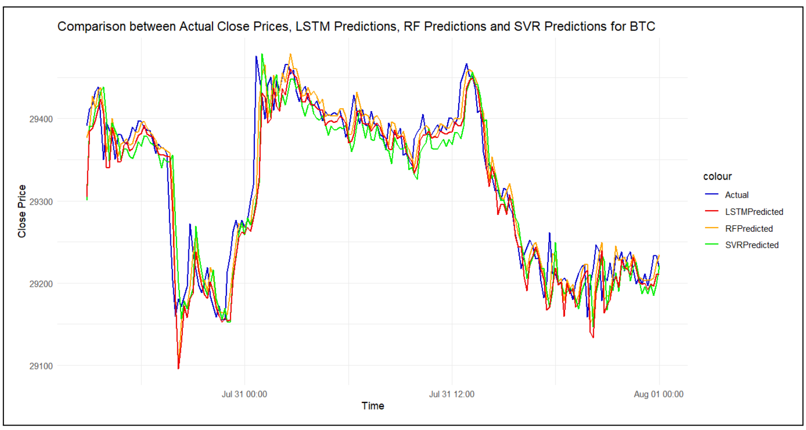

Figure 17: Visual Comparison Between ML Approaches for non-clustered BTC Data

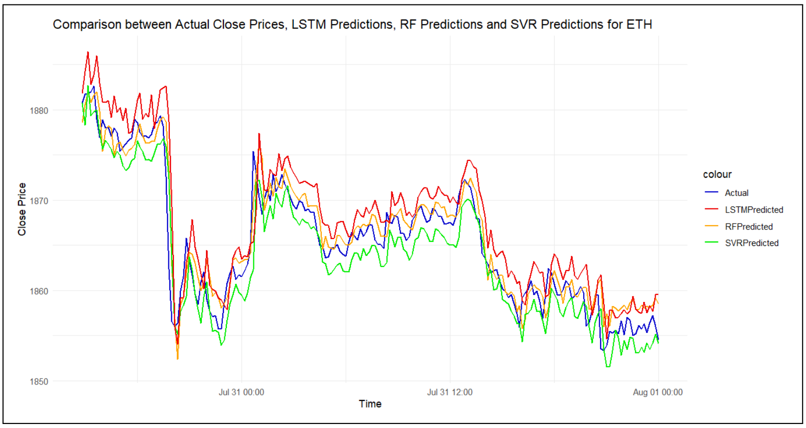

Figure 18: Visual Comparison Between ML Approaches for non-clustered ETH Data

Figure 16 shows a detailed visual representation comparing three machine-learning models for the ADA cryptocurrency using the most recent 200 data points. Similar plots were also created for BTC and ETH to compare the results for these cryptocurrencies, as illustrated in Figures 17, 18, respectively.

Combining clustering techniques with machine learning models led to notable improvements in forecasting accuracy across ADA, BTC, and ETH datasets. Clustering enabled the models to capture more homogeneous volatility patterns, reducing prediction errors. However, results also emphasize the importance of careful cluster validation, as not all cluster divisions consistently enhanced predictive performance compared to the original non-clustered data.

3.5 Discussion

Understanding intraday high-frequency returns is critical in cryptocurrency markets, where volatility shifts rapidly. This study showed that classifying returns into distinct time regions based on volatility provided clear insights into market dynamics that are otherwise hidden at lower frequencies.

The study successfully segmented return data into meaningful regions of stable or volatile behavior by applying changepoint detection and clustering. Region level features such as return variance, volume variance, and region length enabled effective clustering, highlighting patterns such as heightened volatility during market openings and calmer periods during less active times.

Although the segmented models were trained on different time periods and sometimes used varying predictor variables, the comparisons remain valid. Each segment reflects a distinct market regime, and adapting models to these differences provides a clearer picture of performance under varying conditions. This approach highlights the advantage of regime-specific modeling over applying a single model to the full dataset, supporting the overall conclusion that accounting for structural changes improves predictive accuracy.

Machine learning models trained on clustered data demonstrated noticeable improvements in predictive performance. Support Vector Regression (SVR) and Random Forest (RF) models showed substantial reductions in Mean Absolute Error (MAE) after clustering, particularly for ADA and ETH. LSTM models also benefited, especially for BTC and ETH, though results for ADA were mixed, suggesting model specific sensitivity to clustering.

MAE was selected as the primary metric for model comparison due to its straightforward interpretation and direct scaling with the data. In the context of volatile and noisy cryptocurrency returns, MAE offers a more stable and reliable measure of average prediction error. Focusing on MAE made the comparison easier to understand and avoided repeating the same points, while still keeping the results reliable across all the evaluation measures.

The subsets chosen from clustered datasets were selected based on time series continuity within the clusters. Rather than using all clustered data aimlessly, specific subsets were carefully extracted to preserve the temporal structure necessary for high-frequency forecasting. This strategic selection allowed the machine learning models to maintain the sequential relationships critical for accurate cryptocurrency price prediction, further enhancing model performance in clustered scenarios. Clustering isolated volatility patterns, reduced noise, and allowed machine learning models to generalize more effectively to unseen data.

While improvements were most significant for ADA and ETH, BTC showed more variability. This highlights that the effectiveness of clustering can depend on asset-specific volatility structures and market behaviors.

Overall, the changepoint detection, clustering, and machine learning models offer a robust framework for forecasting in high-frequency cryptocurrency markets. The approach strengthens predictive accuracy and model adaptability, contributing valuable insights for algorithmic trading and financial decision-making.

4 Conclusion

This study explored a novel framework for forecasting high-frequency cryptocurrency prices using changepoint analysis, clustering techniques, and machine learning models. The main aim was to improve prediction accuracy by enhancing the structure and quality of input data through volatility-based segmentation.

The research successfully classified intraday high-frequency returns for Bitcoin (BTC), Ethereum (ETH), and Cardano (ADA) into distinct time regions based on return volatility. Applying changepoint analysis identified significant shifts and structural breaks in price behavior, enabling a deeper understanding of market dynamics within trading days.

Using extracted features such as return variance, volume variance, and region length the study grouped volatility regions into homogeneous clusters using the K-means algorithm. The elbow method determined the optimal number of clusters to be three for each cryptocurrency. The best cluster solutions varied by asset. These refined clusters provided more consistent input for forecasting models.

The clustering was found to enhance the forecasting performance of these models in most cases. SVR and RF models benefited significantly from clustering, especially for ADA and ETH, with reduced prediction errors. While BTC results were more variable, some improvements were observed, with clustering improving performance in some models while remaining unchanged or less effective in others. The LSTM model, however, showed mixed outcomes, performing best without clustering for ADA but showing notable improvement for ETH. The results validated that clustering improves the performance of machine learning models in high-frequency price prediction, though the degree of improvement may differ depending on the asset and model type.

Overall, the study confirms that incorporating changepoint-driven clustering enhances the predictive accuracy of machine learning models in high-frequency cryptocurrency forecasting. This method presents promising pathways for creating potent predictive systems in dynamic financial environments.

References

Basu, S. (2020) *What is the difference between the Time series and Machine Learning techniques of prediction?*. Available at: `https://www.linkedin.com/pulse/what-difference-between-timeseries-machine-learning-techniques-basu/` (Accessed: 2024).

Bitto, A., Mahmud, I., Bijoy, H., Jannat, F., Arman, S., Shohug, M. and Jahan, H. (2022) 'CryptoAR: scrutinizing the trend and market of cryptocurrency using machine learning approach on time series data', *Indonesian Journal of Electrical Engineering and Computer Science*, 28(3):1684–1696. DOI: 10.11591/ijeecs.v28.i3.pp1684-1696. (Accessed: 2024).

Central Bank of Sri Lanka (2023) Risks of using and investing in Cryptocurrency. Available at: `https://www.cbsl.gov.lk/en/news/risks-of-using-and-investing-in-cryptocurrency-20230329` [Accessed 2024].

Chandrasekara, G. (2024) *Sri Lanka's Central Bank Governor Discusses Cryptocurrency Regulation Amidst Calls for Relaxation.* Available at: `https://www.lankabusinessonline.com/sri-lankas-central-bank-governor-discusses-cryptocurrency-regulation-amidst-calls-for-relaxation/` (Accessed 2024).

Cloud Software Group, Inc (2024) *What is a random forest?*. Available at: `https://www.spotfire.com/glossary/what-is-a-random-forest` (Accessed 2024).

Das, D., Kayal, P. and Maiti, M. (2023) 'A K-means clustering model for analyzing the Bitcoin extreme value returns', *Decision Analytics Journal*, 6:100–152. DOI:10.1016/j.dajour.2022.100152.

DB Investing (2023) *Cryptocurrencies vs Traditional Finance.* Available at: https://dbinvesting.com/blog/crypto-vs-traditional-finance/ (Accessed 30 August 2023).

Eckhardt, K. (2018) *Choosing the right Hyperparameters for a simple LSTM using Keras.* Available at: https://medium.com/data-science/choosing-the-right-hyperparameters-for-a-simple-lstm-using-keras-f8e9ed76f046 (Accessed 2024).

Fidelity (2023) *Risks and benefits of cryptocurrency.* Available at: https://www.fidelity.com/learning-center/trading-investing/crypto/risks-and-benefits-of-crypto (Accessed 2024).

Frankenfield, J. (2023) *Cryptocurrency Explained With Pros and Cons for Investment.* Available at: https://www.investopedia.com/terms/c/cryptocurrency.asp#:~:text=A%20cryptocurrency%20is%20a%20digital,a%20disparate%20network%20of%20computers (Accessed 2024).

Handagama, S. (2021) *Economic Uncertainty Drives Crypto Growth in Sri Lanka.* Available at: https://www.coindesk.com/markets/2021/05/13/economic-uncertainty-drives-crypto-growth-in-sri-lanka/ (Accessed 2024).

Iqbal, M., Iqbal, M., Jaskani, F., Iqbal, K., Hassan, A. (2021) Time-Series Prediction of Cryptocurrency Market using Machine Learning Techniques. *EAI Endorsed Transactions on Creative Technologies*, DOI:10.4108/eai.7-7-2021.170286.

JForex4 (2018-2024) *Dukascopy.* Available at: https://www.dukascopy.com/land/trading/bank/platforms/jforex4/?lang=en (Accessed 2024).

Killick, R., Fearnhead, P. & Eckley, I. A. (2012) Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590-1598.

Lahmiri, S., Bekiros, S. & Bezzina, F. (2022) Complexity analysis and forecasting of variations in cryptocurrency trading volume with support vector regression tuned by Bayesian optimization under different kernels: : An empirical comparison from a large dataset. *Expert Systems With Applications*, 209(C):0957-4174.

Lorenzo, L. & Arroyo, J. (2022) Analysis of the cryptocurrency market using different prototype-based clustering techniques. *Financial Innovation* DOI:10.1186/s40854-021-00310-9

Ngai, E., Abdullah, S., Nazri, M.Z.A., Sani, N.S. and Othman, Z. (2023) Time Series Prediction of Bitcoin Cryptocurrency Price Based on Machine Learning Approach. *DATA SCIENCE: JOURNAL OF COMPUTING AND APPLIED INFORMATICS (JoCAI)*, 7(2):81-95, DOI:10.32734/jocai.v7.i2-14356

Oyedele, A., Ajayi, A., Oyedele, L., Bello, S., Jimoh, K. (2022) Performance evaluation of deep learning and boosted trees for. *Elsevier*, Volume 213.

Pabuccu, H., Ongan, S. & Ongan, A. (2020) Forecasting the movements of Bitcoin prices: an application of machine. *Quantitative Finance and Economics*, 4(4), DOI:10.3934/QFE.2020031.

Roy, S., Nanjiba, S. & Chakrabarty, A. (2018) Bitcoin Price Forecasting Using Time Series. *21st International Conference of Computer and Information Technology (ICCIT).* DOI: 10.1109/ICCITECHN.2018.8631923.

Sewmini, C. (2023) *The Urgent Need to Regulate the Sri Lankan Cryptocurrency Market.* Available at: `https://www.inss.lk/index.php?id=553` (Accessed 2024).

Sharp, T. (2020) *An Introduction to Support Vector Regression (SVR).* Available at: `https://medium.com/data-science/an-introduction-to-support-vector-regression-svr-a3ebc1672c2#id_token=eyJhbGciOiJSUzI1NiIsImtpZCI6Ijg4MjUwM2E1ZmQ1NmU5ZjczNGRmYmE1YzUwZDdiZjQ4ZGIyODRhZTkiLCJ0eXAiOiJKV1QifQ.eyJpc3MiOiJodHRwczovL2FjY291bnRzLmdvb2dsZS5jb20iLC` (Accessed 2024).

Šťastný, T., Koudelka, J., Bílková, D. & Marek, L. (2022) Clustering and Modelling of the Top 30 Cryptocurrency Prices. *Mathematics*, 10(19), DOI:10.3390/math10193672.

Wang, Y. (2022) Cryptocurrency Price Forecast with Use of Time Series. *International Conference on Financial Innovation, FinTech and Information Technology.*, DOI:10.4108/eai.28-10-2022.2328429.

Yahoo Finance (2024) *Top Cryptos by Volume.* Available at:`https://rb.gy/ryjaen` (Accessed 2024).

Yasar, K. (2023) *Pros and cons of cryptocurrency.* Available at:`https://www.techtarget.com/whatis/feature/Pros-and-cons-of-cryptocurrency` (Accessed 2024).