**Building of Informatics, Technology and Science (BITS)**
Volume 6, No 3, Desember 2024 Page: 1301−1312
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.5744

# Clustering-Based Stock Return Prediction using K-Medoids and Long Short-Term Memory (LSTM)

**Denny Sofyan*, Deni Saepudin**

School of Computing, Telkom University, Bandung, Indonesia
Email: [1,*]dennysofyan@student.telkomuniversity.ac.id, [2]denisaepudin@telkomuniversity.ac.id
Correspondence Author Email: dennysofyan@student.telkomuniversity.ac.id
Submitted: **06/08/2024**; Accepted: **01/12/2024**; Published**: 03/12/2024**

**Abstract**−This research focuses on predicting stock returns using the K-Medoids clustering method and the Long Short-Term Memory (LSTM) model. The primary challenge lies in forecasting stock prices, which are then converted into return predictions. Clustering is performed to group stocks with similar price movements, facilitating the preparation of data for training the LSTM model within each cluster. This issue is crucial for aiding investors in making more informed investment decisions by leveraging predictions within specific stock clusters. Through clustering with K-Medoids, based on average returns and return standard deviation, the LSTM model is trained to predict daily returns for each stock within different clusters using the average stock price in each cluster. The data is divided into training (2013-2019) and testing (2020-2022) datasets, with model evaluation conducted using Root Mean Square Error (RMSE). The implementation results indicate prediction performance measured by RMSE for each cluster, with Cluster 3 showing the best performance with a testing RMSE of 0.0300, while Cluster 4 exhibited the worst performance with an RMSE of 0.3995. In the formation of an equal weight portfolio, tested from May 2020 to January 2023, the portfolio value grew from 1 to 2.50, with an average return of 0.0014 and a return standard deviation of 0.0158, indicating potential gains with lower risk compared to the *LQ45* index.

**Keywords**: K-Medoids; LSTM; Return Prediction; Portfolio; Equal Weight

## 1. INTRODUCTION

This research addresses the problem of predicting stock returns using K-Medoids clustering and Long Short-Term Memory (LSTM) models [1],[2]. The primary goal is to predict stock prices that are subsequently converted into return predictions, employing a clustering-based approach to identify stocks with similar price movements [3],[4]. This approach is crucial as it enables investors to make better investment decisions by selecting stocks with high return potential and lower risk [5],[6]. Sustaining returns while lowering risk requires investors to manage their financial assets effectively [7]. Investors now embrace portfolio optimization, and one notable accomplishment in this regard is the mean-variance (MV) model that Markowitz proposed in 1952 [8]. Even if portfolio optimization has advanced, performance can still be increased by incorporating machine learning (ML) models [9]. Utilizing predictive models to anticipate stock returns enables investors to enhance portfolio outcomes [10]. Return is a primary determinant for investors, representing the reward for bearing investment risk. In portfolio formation, the main consideration is the return calculated from historical data. A portfolio's performance is deemed good if the return based on historical data does not significantly differ from future returns. However, in reality, future data may differ substantially from historical data. Therefore, when building a portfolio, it is necessary to consider future return predictions [11]. In this context, the K-Medoids method is used to cluster stocks based on similar price movement patterns, which helps identify stocks with comparable performance [12]. Meanwhile, the LSTM model is used to predict prices that are converted into returns based on clustering, expected to provide more accurate return forecasts [13]. Stock return is a key factor in investment decision-making, and by leveraging ML-based return predictions, investors can optimize their portfolio performance [14]. The main problem in this research is predicting stock prices converted into return predictions using clustering techniques to group stocks based on similar price movements, which is highly useful for organizing data for LSTM model training [15]. This issue is highly relevant as it allows investors to make more informed investment decisions by relying on predictions for specific stock clusters [16].

In 2023, Ashrafzadeh et al. studied the use of an LSTM model for k-means clustering and initial stock selection in 2023. 21 randomly chosen stocks from the New York Stock Exchange (NYSE) were included in the data, which covered a ten-year span of daily trading from 2012 to 2021. The training phase asset returns and the average and variety of input variables were used as clustering criteria to group the stocks according to attributes like average return and return standard deviation. The LSTM model was then employed to predict prices for each cluster representative, which were subsequently converted into returns by calculating the price change relative to the previous price. These returns were used as indicators of stock performance and processed through the LSTM model to obtain future return predictions. The study's results indicated that the LSTM model was effective in making predictions, with RMSE values for each cluster being 0.0361, 0.0308, 0.0539, 0.0361, and 0.0332, demonstrating varied prediction accuracy that is beneficial for stock selection in an investment portfolio. The average return predicted by the LSTM model was 0.003, with a return standard deviation of 0.0212, confirming that this model can be utilized for long-term profitable investment strategies [3].

In 2022, Man Li et al. conducted a study predicting stock returns by using features such as average return, return standard deviation, and daily return distribution to cluster stocks from the New York Stock Exchange (NYSE) for the period from 2012 to 2021. The Long Short-Term Memory (LSTM) model was employed to predict stock

Copyright © 2024 **Author**, Page 1301
This Journal is licensed under a Creative Commons Attribution 4.0 International License

**Building of Informatics, Technology and Science (BITS)**
Volume 99, No 99, Bulan 2024 Page: 999−999
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v9i9.999

prices, which were then converted into returns by calculating the price change relative to the previous price. The clustering and prediction results indicated that the LSTM model exhibited varying Root Mean Square Error (RMSE) values for each cluster: 0.0361, 0.0308, 0.0539, 0.0361, and 0.0332, reflecting different levels of prediction accuracy. The average return predicted by the LSTM model was 0.003, with a return standard deviation of 0.0212, demonstrating that this model can be utilized for efficient investment strategies [1]. However, unlike Man Li et al.'s approach, which focused on predicting prices and converting them into returns, our study directly predicts returns for each cluster using LSTM. Additionally, while Man Li et al. used traditional clustering methods, our research employs the K-Medoids method, which is more robust against outliers and better suited for grouping stocks based on average return and return standard deviation.

In 2023, Saenz et al. conducted a study using data from 240 companies listed in the Russell 3000 index, covering the period from 2017 to 2022. In this study, the stocks were grouped using the K-Means technique and various distance metrics, including Dynamic Time Warping (DTW), Fourier decomposition, and Extended Frobenius Norm (EROS). The focus of this clustering was on financial ratios, prices, and daily returns. The predictive model used was Long Short-Term Memory (LSTM). The LSTM model was trained for each cluster with the aim of predicting daily prices, which were subsequently converted into stock returns. The results showed that the LSTM model significantly outperformed in predicting stock returns, with an average return gain of 1.97% and a standard deviation of 2.01%. This success underscores that the use of LSTM not only provides high accuracy but also enhances performance, and that clustering data can improve prediction accuracy [4]. In contrast to Saenz et al.'s approach, which involved complex distance metrics and multiple clustering techniques, our study simplifies the process by using the K-Medoids method and focuses on clustering based solely on average return and return standard deviation. Furthermore, we extend our research by applying the results to portfolio formation using the Equal Weight method, thereby offering a practical investment strategy that was not explored in Saenz et al.'s study.

In this study, stock clustering will be performed using the K-Medoids method to group stocks based on their average return and return standard deviation. Subsequently, stock price predictions, which will be converted into return predictions, will be made using the Long Short-Term Memory (LSTM) method for each cluster. The prediction performance for each cluster will be measured using Root Mean Square Error (RMSE). Stocks with the highest predicted returns from each cluster will be selected to represent their respective clusters in portfolio formation. This selection process will be based on the predicted daily returns for each stock. The portfolio will be constructed using the Equal Weight method, and its performance will be measured by comparing the portfolio's average return and return standard deviation against the LQ45 index portfolio. The research aims to develop a more accurate stock return prediction model through clustering and LSTM, ultimately enhancing investment strategy effectiveness.

# 2. RESEARCH METHODOLOGY

## 2.1 Research Stages

The system design in this research utilizes a flowchart to illustrate the entire process from start to finish. Figure 1 presents the research stages involved in this study, beginning with the input of stock price data and concluding with the portfolio performance evaluation.

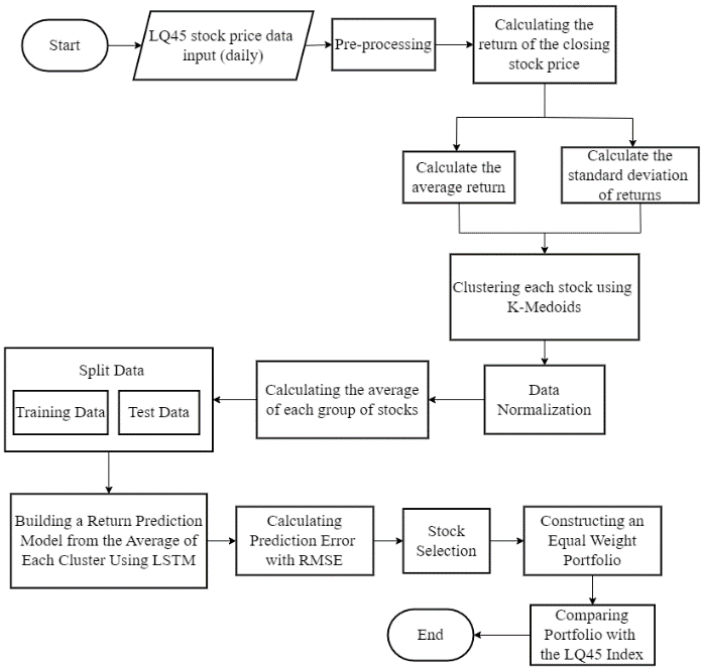


**Figure 1.** Research Stages

Copyright © 2023 **Author**, Page 1302
This Journal is licensed under a Creative Commons Attribution 4.0 International License

**Building of Informatics, Technology and Science (BITS)**
Volume 99, No 99, Bulan 2024 Page: 999−999
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v9i9.999

The flowchart begins with the input of daily LQ45 stock prices, serving as the foundation for the entire process. The output of this stage is a comprehensive dataset containing the daily closing prices for all stocks included in the LQ45 index over the specified period. This dataset then undergoes a pre-processing stage, where it is cleaned and prepared for analysis. During pre-processing, missing values are specifically addressed using linear interpolation, outliers are smoothed, and data consistency is ensured. The refined dataset, now free of inconsistencies, is ready for accurate analysis. Following pre-processing, the system calculates the daily returns of the closing stock prices, measuring the percentage change in stock prices from one day to the next. The output here is a dataset of daily returns for each stock, setting the stage for further statistical analysis.

Next, the system computes the average return and the standard deviation of returns for each stock, providing essential metrics that indicate the stock's performance and volatility. These metrics are then used in the clustering stage, where stocks are grouped using the K-Medoids technique based on their average return and standard deviation. The output of this stage is a set of clusters, each containing stocks with similar return characteristics and volatility. To ensure consistency across these clusters, the data undergoes normalization, resulting in a dataset that is uniformly scaled and ready for predictive modeling. The normalized data is then used to calculate the average of each group of stocks within the clusters, preparing it for the next phase.

The dataset is subsequently split into training and testing sets, with the training data used to build a return prediction model using the LSTM (Long Short-Term Memory) method. The output of this stage is a trained LSTM model capable of predicting future stock returns. The model's performance is evaluated by calculating the prediction error using the Root Mean Square Error (RMSE) metric, producing RMSE values for each cluster that indicate the model's prediction accuracy. Based on these predictions, stocks with the highest predicted returns are selected for portfolio construction. Finally, an equal weight portfolio is constructed from the selected stocks, and its performance is compared against the LQ45 index. The output of this stage includes performance metrics such as average return and standard deviation, which are used to assess the portfolio's effectiveness, concluding the research process.

### 2.2 Stock Data

The stock data utilized in this research comprises the daily closing prices for 39 stocks that are part of the LQ45 index. The data was obtained from finance.yahoo.com, covering a period of 10 years from January 1, 2013, to January 1, 2023. Table 1 provides a detailed list of the stock codes and names of the 39 companies included in this study.

**Table 1.** LQ45 Stock Data

| No. | Code | Stock Name |
|---|---|---|
| **1** | ADRO | Adaro Energy Indonesia Tbk. |
| **2** | AMRT | Sumber Alfaria Trijaya Tbk. |
| **3** | ANTM | Aneka Tambang Tbk. |
| **4** | ASII | Astra International Tbk. |
| **5** | BBCA | Bank Central Asia Tbk. |
| **6** | BBNI | Bank Negara Indonesia (Persero) Tbk. |
| **7** | BBRI | Bank Rakyat Indonesia (Persero) Tbk. |
| **8** | BBTN | Bank Tabungan Negara (Persero) Tbk. |
| **9** | BFIN | BFI Finance Indonesia Tbk. |
| **10** | BMRI | Bank Mandiri (Persero) Tbk. |
| **11** | BRPT | Barito Pacific Tbk. |
| **12** | CPIN | Charoen Pokphand Indonesia Tbk |
| **13** | EMTK | Elang Mahkota Teknologi Tbk. |
| **14** | ERAA | Erajaya Swasembada Tbk. |
| **15** | EXCL | XL Axiata Tbk. |
| **16** | HMSP | H.M. Sampoerna Tbk. |
| **17** | HRUM | Harum Energy Tbk. |
| **18** | ICBP | Indofood CBP Sukses Makmur Tbk. |
| **19** | INCO | Vale Indonesia Tbk. |
| **20** | INDF | Indofood Sukses Makmur Tbk. |
| **21** | INDY | Indika Energy Tbk. |
| **22** | INKP | Indah Kiat Pulp & Paper Tbk. |
| **23** | INTP | Indocement Tunggal Prakarsa Tbk. |
| **24** | ITMG | Indo Tambangraya Megah Tbk. |
| **25** | JPFA | Japfa Comfeed Indonesia Tbk. |
| **26** | KLBF | Kalbe Farma Tbk. |
| **27** | MEDC | Medco Energi Internasional Tbk. |
| **28** | MNCN | Media Nusantara Citra Tbk. |
| **29** | PGAS | Perusahaan Gas Negara Tbk. |
| **30** | PTBA | Bukit Asam Tbk. |

Copyright © 2023 **Author**, Page 1303
This Journal is licensed under a Creative Commons Attribution 4.0 International License

**Building of Informatics, Technology and Science (BITS)**
Volume 99, No 99, Bulan 2024 Page: 999−999
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v9i9.999

| | | |
|---|---|---|
| **31** | SMGR | Semen Indonesia (Persero) Tbk. |
| **32** | TBIG | Tower Bersama Infrastructure Tbk. |
| **33** | TINS | Timah Tbk. |
| **34** | TLKM | Telkom Indonesia (Persero) Tbk. |
| **35** | TOWR | Sarana Menara Nusantara Tbk. |
| **36** | TPIA | Chandra Asri Petrochemical Tbk. |
| **37** | UNTR | United Tractors Tbk. |
| **38** | UNVR | Unilever Indonesia Tbk. |
| **39** | WIKA | Wijaya Karya (Persero) Tbk. |

The stock data utilized in this research comprises the daily closing prices for 39 stocks that are part of the LQ45 index. The data was obtained from finance.yahoo.com, covering a period of 10 years from January 1, 2013, to January 1, 2023. Table 1 provides a detailed list of these 39 stocks, including their respective stock codes and company names. The stocks in the table represent a diverse range of industries, including energy, finance, consumer goods, and telecommunications, reflecting the broad coverage of the LQ45 index.

### 2.3 Pre-processing

Preprocessing is necessary because raw data often has inconsistent formats and may contain missing values. In this research, linear interpolation is used to handle missing values. Linear interpolation is a mathematical algorithm that estimates the median price point by drawing a straight line between two consecutive input points.

### 2.4 Calculate Stock Price Returns

Calculating the return of closing stock prices over the past 10 years using daily data for each stock in the LQ45 index. To calculate the stock price return, the following equation can be used:

$$R_{i,t} = \frac{P_{i,t} - P_{t-1}}{P_{t-1}} \tag{1}$$

The explanation of the above formula is that, $i$ refers to the asset, $R_{i,t}$ shows the stock return $i$ at time $t$, while $P_t$ is value of the stock at time $t$, and finally $P_{t-1}$ is value of the stock at time $t-1$.

#### 2.4.1 Calculate Average Return

The average return is calculated by summing all the daily return values over a specific period and then dividing the result by the number of periods. To calculate the average return, the following equation can be used:

$$\bar{R} = \frac{\sum_{t=1}^{n} R_{i,t}}{n} \tag{2}$$

The explanation of the above formula is that, $i$ refers to the asset, while $\bar{R}$ is the average stock return, $R_{i,t}$ represents the return of stock $i$ at time $t$, and finally $n$ is number of periods.

#### 2.4.2 Calculate Standard Deviation of Returns

The degree to which a stock's return deviates from its average return is determined using the standard deviation of returns. To determine the returns' standard deviation, the following equation can be used:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(R_i - \bar{R})^2}{n}} \tag{3}$$

The explanation of the above formula is that, $\sigma$ refers to the standard deviation, while $i$ denotes the asset, $n$ represents the number of periods, $R_{i,t}$ is return of stock $i$ at time $t$, and finally $\bar{R}$ is average stock return.

### 2.5 Application of K-Medoids Method

At this stage, the data will be processed using the K-Medoids algorithm based on the calculated average and standard deviation of each stock's returns [17], [18]. The experiment involves clustering the data using four variables: the average return for the first five years, the average return for the last five years, the standard deviation of the return for the first five years, and the standard deviation of the return for the last five years. Figure 2 below depicts the flowchart illustrating the process of applying the K-Medoids clustering.

Copyright © 2023 **Author**, Page 1304
This Journal is licensed under a Creative Commons Attribution 4.0 International License

**Building of Informatics, Technology and Science (BITS)**
Volume 99, No 99, Bulan 2024 Page: 999−999
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v9i9.999

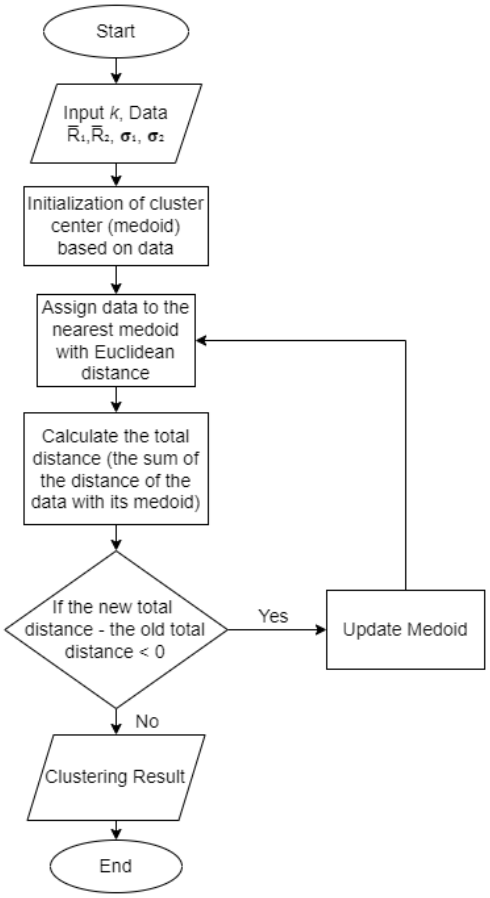


**Figure 2.** Flowchart for Implementing K-Medoids Clustering

Based on Figure 2, the following explains the stages in the process of clustering stocks using the K-Medoids Clustering method:

a. Input $k$, Data $\bar{R}_1, \bar{R}_2, \sigma_1, \sigma_2$: Enter the number of clusters $k$ and the variables $\bar{R}_1, \bar{R}_2, \sigma_1, \sigma_2$ that will be clustered.
b. Initialize cluster centers (medoids) based on the data: Randomly determine the initial cluster centers from the available data.
c. Assign data to the nearest medoid using Euclidean distance: Allocate each data point to the nearest medoid based on the Euclidean distance calculation. The Euclidean distance can be calculated using the following equation:

$$d_{i,j} = \sqrt{(\bar{R}_{1,i} - \bar{R}_{1,j})^2 + (\bar{R}_{2,i} - \bar{R}_{2,j})^2 + (\sigma_{1,i} - \sigma_{1,j}) + (\sigma_{2,i} - \sigma_{2,j})} \quad (4)$$

Where $d_{i,j}$ is the distance between two points $i$ and $j$, and $\bar{R}_1, \bar{R}_2, \sigma_1, \sigma_2$ are the data attributes used.

d. Calculate total distance (sum of the distances between data points and their medoids): Compute the total distance between data points and their respective medoids.
e. If new total distance - old total distance < 0:
   1. Yes: Update the medoid to the new candidate that has been calculated.
   2. No: Retain the current medoid
f. Update Medoid: Re-select the new medoid from the previously calculated non-medoid candidates and return to step 4.
g. Clustering Results: Once the iterations reach convergence (no further changes in medoids), the clustering results are established.

### 2.6 Data Normalization

At this stage, data normalization is performed by dividing each value by its initial price. This normalization process ensures that every stock price has a comparable range of values. Normalizing the data is crucial to avoid disproportionate influence in the prediction model, as stocks with larger value ranges could dominate the prediction outcomes. With normalization, each stock is treated equally, allowing the model to deliver more accurate predictions.

### 2.7 Split Data

After normalizing the data, the dataset is divided into two parts: training data and testing data. The training data is employed to train the algorithm and develop the LSTM model, whereas the testing data is utilized to evaluate the performance of the trained model. The data is divided with 70% designated for training and 30% for testing.

Copyright © 2023 **Author**, Page 1305
This Journal is licensed under a Creative Commons Attribution 4.0 International License

**Building of Informatics, Technology and Science (BITS)**
Volume 99, No 99, Bulan 2024 Page: 999−999
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v9i9.999

### 2.8 Stock Price Prediction with LSTM

The memory units in an LSTM are referred to as cells. The cell state and the hidden state are the two states that are transferred to the following cell. Additionally, LSTM features three types of gates: the forget gate $G_f(t)$, the input gate $G_i(t)$, and the output gate $G_o(t)$. The structure of the LSTM architecture can be seen in Figure 3 [19],[20].

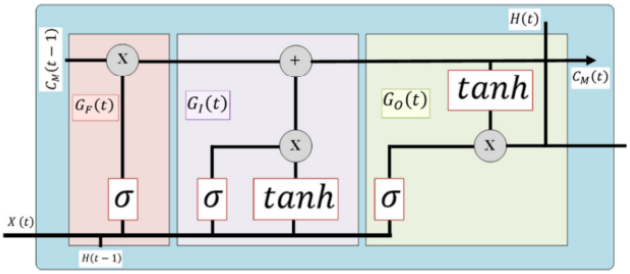


**Figure 3.** Architecture LSTM [3]

The LSTM architecture depicted in Figure 3 showcases the intricate process of handling sequential data, such as stock prices, which often exhibit time-dependent patterns. By effectively controlling the flow of information through the forget, input, and output gates, the LSTM model is capable of maintaining relevant information over long periods while discarding what is no longer necessary.

In this research, stock price prediction is conducted based on the results of the optimal model. The model-building process involves several scenarios and parameter settings to determine the best configuration. To test each scenario, a number of important factors affecting the LSTM model's performance are changed, as outlined in Table 2.

**Table 2.** Default Configuration and Parameter Values of LSTM

| Parameter | Value |
|---|---|
| **Input Features** | Average Stock Price of Clusters |
| **LSTM Layers** | 4 |
| **LSTM Units per Layer** | [50, 100, 150] |
| **Dropout** | [0.2, 0.25, 0.3] |
| **Optimizer** | Adam |
| **Loss Function** | Mean Squared Error |
| **Epochs** | [50, 100, 150] |
| **Batch Size** | [32, 64, 128] |

The LSTM model with 50 units per layer, a dropout rate of 0.2, 50 epochs, and a batch size of 32 was selected because it provides an optimal balance between prediction accuracy and computational efficiency. Models with more LSTM units, such as 100 or 150, and higher dropout rates, such as 0.25 or 0.3, were not chosen despite their potential to improve accuracy because they tend to require longer training times and have a higher risk of overfitting. Additionally, increasing the number of epochs to 100 or 150 did not show significant performance improvements, only increasing the training duration. Therefore, the chosen configuration is considered the most efficient and effective for prediction without sacrificing performance.

### 2.9 Calculation of Prediction Error

After making predictions, the next step is to calculate The error in the prediction results from the developed model is evaluated using Root Mean Squared Error (RMSE). The square root of the average squared differences between the actual and projected values of the model is measured by RMSE. When the RMSE number is low, or nearly zero, it means that the prediction model approximates the real values with accuracy [21]. The following is the RMSE equation for calculating prediction error:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(P_t - \hat{P}_t)^2}{n}} \tag{5}$$

The explanation of the above formula is that, $n$ which is the number of data points, $P_t$ which represents the actual stock price at time $t$, and $\hat{P}_t$ which represents the predicted stock price at time $t$.

### 2.10 Conversion of Predicted Prices to Returns

After building the stock price prediction model, the next step is to convert the prediction results into returns. These predicted returns are then used as indicators in the stock selection process, which will be explained in the next stage. To calculate the predicted stock returns, the following equation can be used:

Copyright © 2023 **Author**, Page 1306
This Journal is licensed under a Creative Commons Attribution 4.0 International License

**Building of Informatics, Technology and Science (BITS)**
Volume 99, No 99, Bulan 2024 Page: 999−999
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v9i9.999

$$\hat{R}_{i,t} = \frac{\hat{P}_t - P_{t-1}}{P_{t-1}} \quad (6)$$

The explanation of the above formula is that, $\hat{R}_{i,t}$ which is the predicted return of stock $i$ at time $t$, $\hat{P}_t$ which represents the predicted stock price at time $t$, and $P_t$ which represents the actual stock price at time $t-1$.

### 2.11 Equal Weight Portfolio

Forming a portfolio with equal weight means distributing an equal portion of the total portfolio value to each asset within the portfolio. To calculate the equal weight (equal weighting) of an asset in a portfolio [22], the following equation can be used:

$$W_i = \frac{1}{n}, i = 1,2,\dots,n \quad (7)$$

The explanation of the above formula is that, $i$ which is the asset, $W_i$ which represents the weight of asset $i$ in the portfolio, and $n$ which represents the total number of assets in the portfolio.

### 2.12 Portfolio Return

Portfolio return is crucial for measuring investment performance and making better future investment decisions. This return can be calculated based on the shift in the portfolio's asset values and the income those assets produce. To calculate the daily portfolio return, the following equation can be used:

$$Rp_t = \sum_{i=1}^{n}(W_i \times R_{i,t}), i = 1,2,\dots,n \quad (8)$$

The explanation of the above formula is that,$i$ which is the asset, $Rp_t$ shows the portfolio return at time $t$, $W_i$ is weight of asset $i$ in the portfolio, $R_{i,t}$ which represents the return of stock $i$ at time $t$, and $n$ which represents the total number of assets in the portfolio.

### 2.13 Portfolio Growth

After calculating the portfolio return, the next step is to conduct tests to evaluate whether the formed portfolio has shown good performance. In this research, the calculation of the wealth value of the created portfolio is performed. The calculated wealth value is then compared with the wealth value from the LQ45 index stock data. The following equation is used to calculate portfolio growth:

$$V_t = V_{t-1} \times (1 + Rp_t) \quad (9)$$

The explanation of the above formula is that, $V_t$ which is the portfolio wealth value at time $t$, $V_{t-1}$ which represents the portfolio wealth value at time $t-1$, and $Rp_t$ which represents the portfolio return at time $t$.

## 3. RESULT AND DISCUSSION

### 3.1 Evaluating Optimal Cluster Number Using the Davies-Bouldin Index

In this study, the K-Medoids method was employed to cluster daily average stock prices, utilizing the average return and standard deviation of returns as key features for grouping stocks based on their volatility and performance. The stock price returns were calculated using Equation 1, which defines the return $R_{i,t}$ of a stock as the change in its closing price from one day to the next, relative to its previous day's closing price. Once the daily returns were computed, the average return over a specific period was determined using Equation 2, providing a measure of each stock's overall performance. Additionally, the standard deviation of returns, calculated using Equation 3, was used to quantify the volatility of each stock by measuring the extent to which the returns deviate from the average return.

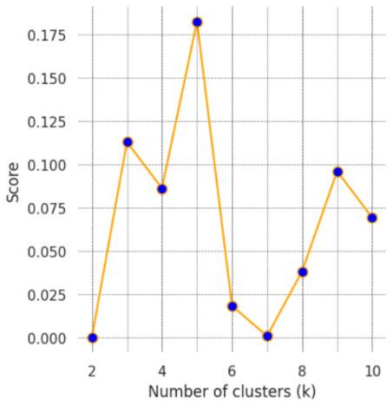


**Figure 4.** Davies-Bouldin Index

Copyright © 2023 **Author**, Page 1307
This Journal is licensed under a Creative Commons Attribution 4.0 International License

**Building of Informatics, Technology and Science (BITS)**
Volume 99, No 99, Bulan 2024 Page: 999−999
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v9i9.999

Figure 4 shows the Davies-Bouldin index values across different numbers of clusters. The analysis revealed that clusters=2 and clusters=7 produced the lowest index values, suggesting these as optimal clustering configurations. Although cluster=2 displayed a very low index value, cluster=7 was selected for further analysis because it offers a more nuanced categorization of the stock data, accommodating a broader set of stock performance characteristics. This decision reflects the importance of capturing diverse stock behavior across different market conditions, which is crucial for making informed investment decisions.

### 3.2 K-Medoids Clustering Results

The K-Medoids clustering method was applied to historical stock data from 2013 to 2022, grouping stocks based on similar price movements. The distance between each stock and the centroid of its assigned cluster was calculated using Equation 4, which measures the Euclidean distance based on features like the average return and standard deviation of returns over both the first and last five years of the study period.

**Table 3.** K-Medoids Clustering Results

| Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 | Cluster 7 |
|---|---|---|---|---|---|---|
| **EXCL** | ASII | CPIN | ANTM | EMTK | ADRO | AMRT |
| **TOWR** | BBCA | JPFA | BRPT | ERAA | BFIN | BBTN |
| | BBNI | MNCN | HRUM | INCO | ITMG | INTP |
| | BBRI | PTBA | TINS | INDY | PGAS | SMGR |
| | BMRI | | | INKP | | TBIG |
| | HMSP | | | MEDC | | TPIA |
| | ICBP | | | | | UNTR |
| | INDF | | | | | |
| | KLBF | | | | | |
| | TLKM | | | | | |
| | UNVR | | | | | |

Table 3 presents the clustering results, where stocks are grouped into seven distinct clusters. Each cluster contains stocks with similar return profiles and volatility, suggesting that the K-Medoids method effectively grouped stocks with related characteristics. The clustering process reveals that certain stocks exhibit more stability and consistent performance, while others are more volatile. This segmentation is valuable for investors as it allows for targeted analysis of stocks within specific clusters, which can be used to tailor investment strategies according to risk tolerance and return expectations.

### 3.3 Stock Price Predictions per Cluster

The predictive model employed in this study leverages a Long Short-Term Memory (LSTM) network to forecast stock prices across various clusters. The LSTM model processes input data through a series of operations involving gates and cell states. Specifically, the input gate determines what new information should be added to the cell state, while the forget gate controls what information should be removed. The cell state is updated by integrating the new information with the retained data, and the output gate decides what part of the cell state should be output. Finally, the hidden state is updated and used for the next time step's prediction. These operations collectively enable the LSTM model to learn and predict sequential patterns in stock prices across different clusters.

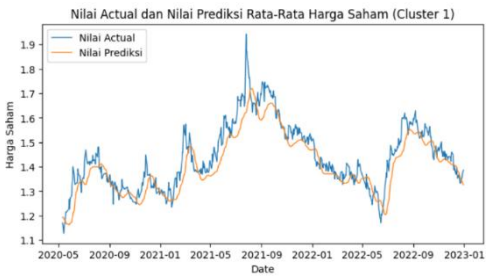


**Figure 5.** Cluster 1 Prediction Results

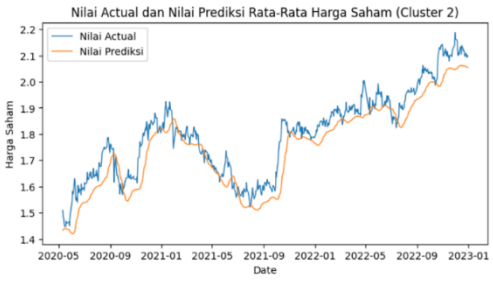


**Figure 6.** Cluster 2 Prediction Results

Copyright © 2023 **Author**, Page 1308
This Journal is licensed under a Creative Commons Attribution 4.0 International License

**Building of Informatics, Technology and Science (BITS)**
Volume 99, No 99, Bulan 2024 Page: 999−999
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v9i9.999

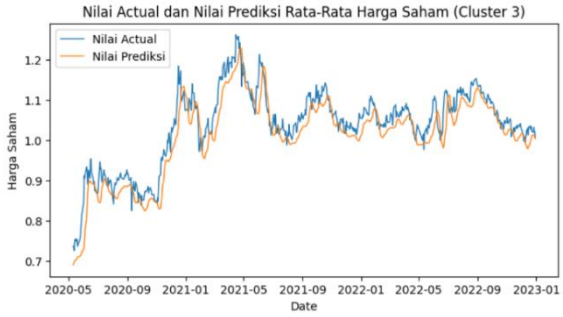


**Figure 7.** Cluster 3 Prediction Results

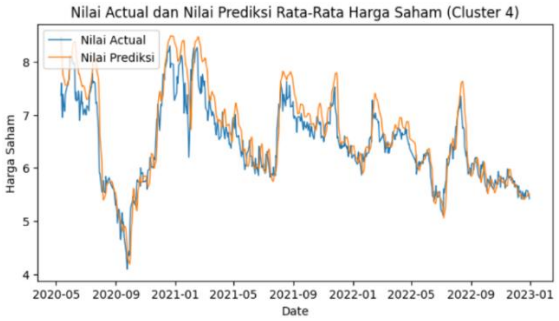


Figure 8. Cluster 4 Prediction Results

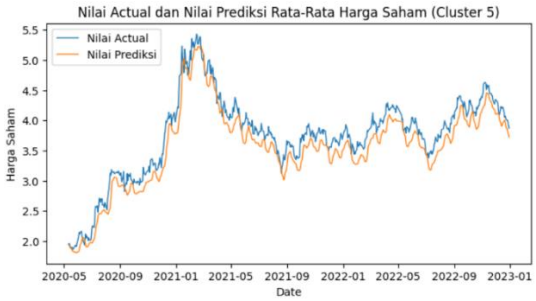


Figure 9. Cluster 5 Prediction Results

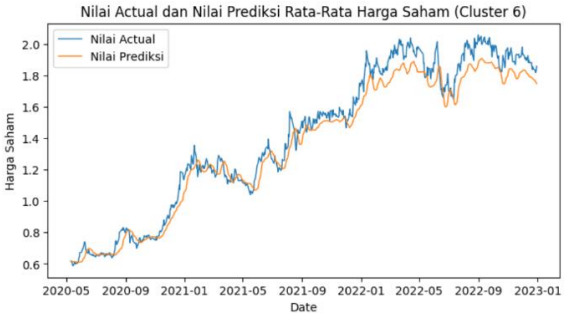


Figure 10. Cluster 6 Prediction Results


**Figure 11.** Cluster 7 Prediction Results

The predictions generally aligned well with actual stock price movements, particularly in Clusters 1, 2, 3, 5, and 6, where the model exhibited relatively high accuracy. However, some discrepancies were observed, with the model tending to underestimate values in these clusters. In contrast, Clusters 4 and 7 demonstrated a tendency for the model to overestimate stock prices, suggesting that these clusters contain stocks with higher volatility or more complex patterns that are challenging for the model to accurately predict. This phenomenon indicates that while the LSTM model is robust, its performance can be influenced by the characteristics of the data within each cluster, particularly in clusters with greater variability or non-linear trends.

### 3.4 Model Performance Evaluation

The performance of the LSTM prediction model across various clusters was evaluated using the Root Mean Square Error (RMSE), which provides a measure of the average prediction error. The RMSE is calculated using Equation 5:

**Table 4.** LSTM Prediction Model Results

| Cluster | RMSE |
|---|---|
| **1** | 0.0710 |
| **2** | 0.0547 |
| **3** | 0.0300 |
| **4** | 0.3995 |
| **5** | 0.1437 |
| **6** | 0.0658 |
| **7** | 0.1870 |

Cluster 1 has an RMSE of 0.0710, indicating good predictive performance with a low error rate. Cluster 2 has an RMSE of 0.0547, which is also low, demonstrating that the LSTM model can predict the average stock prices with

Copyright © 2023 **Author**, Page 1309
This Journal is licensed under a Creative Commons Attribution 4.0 International License

**Building of Informatics, Technology and Science (BITS)**
Volume 99, No 99, Bulan 2024 Page: 999−999
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v9i9.999

high accuracy in this cluster. Cluster 3 has the lowest RMSE of 0.0300, indicating that the LSTM model has very high accuracy in predicting average stock prices. Cluster 4 has the highest RMSE at 0.3995, which is higher compared to the previous clusters, due to the significant price fluctuations within this cluster resulting in a relatively high error rate in predictions. Cluster 5 has an RMSE of 0.1437. Although the RMSE in this cluster is quite high, the LSTM model still manages to provide reasonably good predictions. Cluster 6 shows an RMSE of 0.0658, where the LSTM model demonstrates good predictive performance with a low error rate. Cluster 7 shows an RMSE of 0.1870, indicating a less satisfactory performance of the LSTM model in predicting the average stock prices, due to greater stock price fluctuations than the average within this cluster, leading to a high RMSE in the predictions.

Overall, the LSTM model demonstrates varied capabilities in predicting average stock prices across different clusters. With the lowest RMSE of 0.0300 in Cluster 3, the model shows excellent performance, while the highest RMSE in Cluster 4 at 0.3995 indicates challenges in predicting stock prices in clusters with high fluctuations. Nonetheless, the model generally provides sufficiently accurate predictions in most clusters, which is beneficial for stock analysis and selection in investment portfolio management.

### 3.5 Stock Selection Based on Return Predictions

The selection of stocks within each cluster is based on the predicted returns generated by the LSTM model. The predicted returns are derived by converting the predicted stock prices into returns using Equation 6:

**Table 5.** Illustration of Stock Selection in Cluster 3

| Date | Actual Return CPIN | Predicted Return CPIN | Actual Return JPFA | Predicted Return JPFA | Actual Return MNCN | Predicted Return MNCN | Actual Return PTBA | Predicted Return PTBA |
|---|---|---|---|---|---|---|---|---|
| **2020-05-12** | 0.0311 | -0.0509 | -0.0270 | **0.0061** | -0.0301 | -0.0002 | -0.0293 | 0.1046 |
| **2020-05-13** | -0.0086 | -0.0796 | -0.0111 | **0.0293** | 0.0311 | 0.0246 | -0.0427 | -0.0631 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... |
| **2020-12-30** | -0.0342 | -0.0543 | -0.0077 | 0.0227 | 0.0137 | -0.0942 | 0.0054 | **-0.0168** |

Table 5 provides an illustration of how stock selection is performed based on the highest predicted returns within each cluster. For instance, on May 12, 2020, JPFA was selected because it had the highest predicted return of 0.0061. This process demonstrates that stock selection is carried out by considering the highest predicted return of each stock within the cluster, and the selected stocks are then used to form a portfolio using the Equal Weight method.

### 3.6 Equal Weight Portfolio Growth

The Equal Weight portfolio in this study was constructed by evenly distributing the total investment across all selected stocks. Each stock in the portfolio was assigned an equal weight, calculated using Equation 7, where each stock's weight was $\frac{1}{n}$, with $n$ representing the total number of stocks in the portfolio. The portfolio's overall return was then determined by summing the weighted returns of the individual stocks, as defined by Equation 8. To evaluate the portfolio's performance over time, its cumulative value was calculated using Equation 9, which considers the initial portfolio value and the compounded returns across the observed periods.

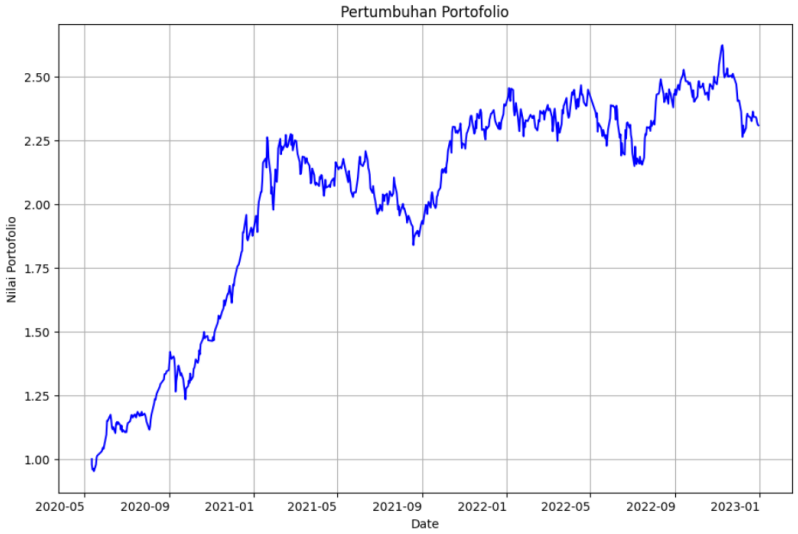


**Figure 12.** Equal Weight Portfolio Growth

Copyright © 2023 **Author**, Page 1310
This Journal is licensed under a Creative Commons Attribution 4.0 International License

**Building of Informatics, Technology and Science (BITS)**
Volume 99, No 99, Bulan 2024 Page: 999−999
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v9i9.999

In Figure 12, the chart illustrates the growth of a portfolio with an Equal Weight allocation from May 2020 to January 2023. The initial value of the portfolio was 1 and peaked at around 2.50 in 2022. Despite experiencing fluctuations and some declines, particularly noticeable towards the end of the period, the overall trend shows a substantial increase, reflecting the resilience and potential benefits of the Equal Weight investment strategy over the observed timeframe. This suggests that the portfolio, while subject to market volatilities, has generally trended upwards, providing a solid return on investment for stakeholders.

### 3.7 Comparison of Equal Weight Portfolio with LQ45 Index

The comparison between the Equal Weight portfolio and the LQ45 index, as shown in Figure 13, reveals a significant difference in performance over the observed period.

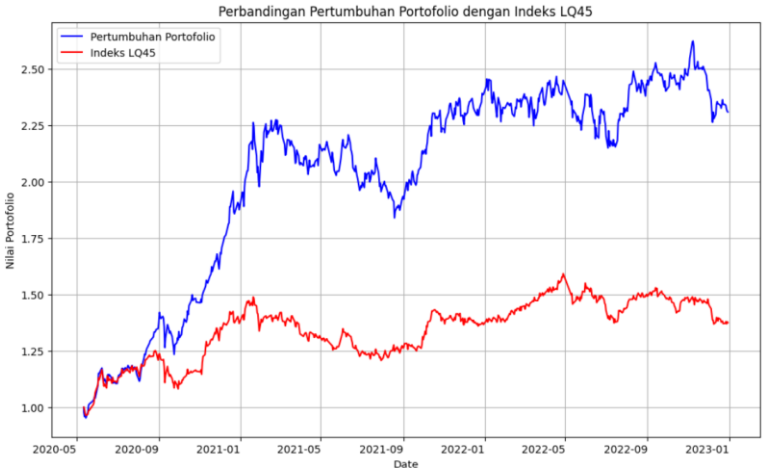


**Figure 13.** Comparison of Equal Weight Portfolio with LQ45 Index

Based on the comparison with the LQ45 index, the Equal Weight portfolio recorded a better performance with higher growth. The Equal Weight portfolio achieved a value of around 2.50, while the LQ45 index only reached a value of about 1.5. Overall, the Equal Weight portfolio constructed provided superior results compared to the LQ45 index.

**Table 6.** Comparison of Average Return and Standard Deviation of Return for Portfolios

| | Average Return | Standard Deviation of Return |
|---|---|---|
| **Portfolio Equal Weight** | 0.0014 | 0.0158 |
| **Index LQ45** | 0.0006 | 0.0119 |

In Table 6, the average return of the Equal Weight Portfolio at 0.0016 indicates that, on average, this portfolio experienced an increase of 0.14%, signifying better performance compared to the LQ45 Index. The standard deviation of return at 0.0158 shows that this portfolio has relatively low fluctuations, indicating lower risk. In contrast, the LQ45 Index has an average return of 0.0006, which means the index saw an increase of 0.06%, showing a lower potential gain. Although the LQ45 Index has a lower standard deviation of return at 0.0119, indicating smaller fluctuations, the overall comparison of lower performance returns suggests that the Equal Weight Portfolio offers more beneficial outcomes with a risk level not significantly different from the standard deviation of return of the LQ45 Index.

## 4. CONCLUSION

Based on the experimental results, the LSTM model demonstrated varied predictive performance across clusters. The best performance was observed in Cluster 3 with an RMSE of 0.0300, indicating a high level of accuracy. In contrast, Cluster 4 had the highest RMSE at 0.3995, showing difficulties in predicting stock price movements in that cluster. These results suggest that the LSTM model is more effective in predicting stock prices in clusters with lower fluctuations, while prediction accuracy decreases in clusters with higher fluctuations. In terms of stock selection, the portfolio formed through a selection process based on the highest predicted daily returns for each stock in each cluster showed better performance compared to the LQ45 index. The Equal Weight portfolio recorded an average return of 0.0014 with a standard deviation of return of 0.0158, indicating a 0.14% increase with low risk. By comparison, the LQ45 index showed a 0.06% growth with lesser risk, with an average return of 0.0006 and a standard deviation of return of 0.0119. This demonstrates that the approach of using K-Medoids clustering to group stocks, followed by employing an LSTM model to predict stock prices which are then derived into returns, can provide better performance in managing potential profits with a risk level not significantly different from the standard deviation of the LQ45 index's returns.

Copyright © 2023 **Author**, Page 1311
This Journal is licensed under a Creative Commons Attribution 4.0 International License

**Building of Informatics, Technology and Science (BITS)**
Volume 99, No 99, Bulan 2024 Page: 999−999
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v9i9.999

# REFERENCES

[1] M. Li, Y. Zhu, Y. Shen, and M. Angelova, "Clustering-enhanced stock price prediction using deep learning," *World Wide Web*, vol. 26, no. 1, pp. 207–232, 2023, doi: 10.1007/s11280-021-01003-0.

[2] N. Naik and B. R. Mohan, *Study of stock return predictions using recurrent neural networks with LSTM*, vol. 1000. Springer International Publishing, 2019. doi: 10.1007/978-3-030-20257-6_39.

[3] M. Ashrafzadeh, H. M. Taheri, M. Gharehgozlou, and S. Hashemkhani Zolfani, "Clustering-based return prediction model for stock pre-selection in portfolio optimization using PSO-CNN+MVF," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 9, p. 101737, 2023, doi: 10.1016/j.jksuci.2023.101737.

[4] J. Vásquez Sáenz, F. M. Quiroga, and A. F. Bariviera, "Data vs. information: Using clustering techniques to enhance stock returns forecasting," *International Review of Financial Analysis*, vol. 88, no. November 2022, p. 102657, 2023, doi: 10.1016/j.irfa.2023.102657.

[5] K. Nakagawa, M. Imamura, and K. Yoshida, "Stock price prediction using k-medoids clustering with indexing dynamic time warping," *Electronics and Communications in Japan*, vol. 102, no. 2, pp. 3–8, 2019, doi: 10.1002/ecj.12140.

[6] D. O. Sunday, "Application of Long Short-Term Memory (LSTM) in Stock Price Prediction," *International Journal of Development and Economic Sustainability*, vol. 12, no. 3, pp. 36–45, 2024.

[7] M. Umer Ghani, M. Awais, and M. Muzammul, "Stock Market Prediction Using Machine Learning(ML)Algorithms," *Advances in Distributed Computing and Artificial Intelligence Journal*, vol. 8, no. 4, pp. 97–116, 2019, doi: 10.14201/ADCAIJ20198497116.

[8] G. N. Mulyono, D. Saepudin, and A. A. Rohmawati, "Portfolio Optimization Based on Return Prediction and Semi Absolute Deviation (SAD)," *International Journal on Information and Communication Technology (IJoICT)*, vol. 9, no. 1, pp. 14–26, 2023, doi: 10.21108/ijoict.v9i1.698.

[9] B. He, E. Gong, L. Li, and Y. Yang, "A Stock Price Prediction Method based on LSTM and K-Means," *Frontiers in Science and Engineering*, vol. 3, no. 6, pp. 44–57, 2023, doi: 10.54691/fse.v3i6.5121.

[10] W. Bessler and D. Wolff, "Portfolio Optimization with Sector Return Prediction Models," *JRFM*, vol. 17, no. 6, p. 254, Jun. 2024, doi: 10.3390/jrfm17060254.

[11] M. Mallikarjuna and R. P. Rao, "Evaluation of forecasting methods from selected stock market returns," *Financial Innovation*, vol. 5, no. 1, 2019, doi: 10.1186/s40854-019-0157-x.

[12] W. Wang, W. Li, N. Zhang, and K. Liu, "Portfolio formation with preselection using deep learning from long-term financial data," *Expert Systems with Applications*, vol. 143, p. 113042, 2020, doi: 10.1016/j.eswa.2019.113042.

[13] Y. Ma, R. Han, and W. Wang, "Portfolio optimization with return prediction using deep learning and machine learning," *Expert Systems with Applications*, vol. 165, no. September 2020, p. 113973, 2021, doi: 10.1016/j.eswa.2020.113973.

[14] O. B. Sezer, M. U. Gudelek, and A. M. Ozbayoglu, "Financial time series forecasting with deep learning: A systematic literature review: 2005–2019," *Applied Soft Computing Journal*, vol. 90, p. 106181, 2020, doi: 10.1016/j.asoc.2020.106181.

[15] T. B. Shahi, A. Shrestha, A. Neupane, and W. Guo, "Stock Price Forecasting with Deep Learning: A Comparative Study," *Mathematics*, vol. 8, no. 9, p. 1441, Aug. 2020, doi: 10.3390/math8091441.

[16] Z. Moeini Najafabadi, M. Bijari, and M. Khashei, "Making investment decisions in stock markets using a forecasting-Markowitz based decision-making approaches," *JM2*, vol. 15, no. 2, pp. 647–659, Nov. 2019, doi: 10.1108/JM2-12-2018-0217.

[17] D. Dwi Aulia and N. Nurahman, "Comparison Performance of K-Medoids and K-Means Algorithms In Clustering Community Education Levels," *j. nas. pendidik. teknik. inform.*, vol. 12, no. 2, pp. 273–282, Jul. 2023, doi: 10.23887/janapati.v12i2.59789.

[18] G. Amato, C. Gennaro, V. Oria, and M. Radovanović, Eds., *Similarity Search and Applications: 12th International Conference, SISAP 2019, Newark, NJ, USA, October 2–4, 2019, Proceedings*, vol. 11807. in Lecture Notes in Computer Science, vol. 11807. Cham: Springer International Publishing, 2019. doi: 10.1007/978-3-030-32047-8.

[19] H. Qian, "Stock Predicting based on LSTM and ARIMA," in *Proceedings of the 2022 2nd International Conference on Economic Development and Business Culture (ICEDBC 2022)*, vol. 225, Y. Jiang, Y. Shvets, and H. Mallick, Eds., in Advances in Economics, Business and Management Research, vol. 225. , Dordrecht: Atlantis Press International BV, 2022, pp. 485–490. doi: 10.2991/978-94-6463-036-7_72.

[20] P. S. Kumar, H. S. Behera, K. Anisha Kumari, J. Nayak, and B. Naik, "Advancement from neural networks to deep learning in software effort estimation: Perspective of two decades," *Computer Science Review*, vol. 38, p. 100288, 2020, doi: 10.1016/j.cosrev.2020.100288.

[21] A. Kumar *et al.*, "Generative adversarial network (GAN) and enhanced root mean square error (ERMSE): deep learning for stock price movement prediction," *Multimedia Tools and Applications*, vol. 81, no. 3, pp. 3995–4013, 2022, doi: 10.1007/s11042-021-11670-w.

[22] A. Chaweewanchon and R. Chaysiri, "Markowitz Mean-Variance Portfolio Optimization with Predictive Stock Selection Using Machine Learning," *IJFS*, vol. 10, no. 3, p. 64, Aug. 2022, doi: 10.3390/ijfs10030064.


This Journal is licensed under a Creative Commons Attribution 4.0 International License