BAREKENG: Journal of Mathematics and Its Applications
September 2026 Volume 20 Issue 3 Page 1855-1868
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol20iss3pp1855-1868

CONSTRUCTING AN OPTIMAL PORTFOLIO USING CLUSTERING LARGE APPLICATION AND VALUE AT RISK ANALYSIS FOR IDX80 STOCKS

Sania Pujianti[1], Hendra Perdana[2], Neva Satyahadewi[3*]

[1,2,3]*Department of Mathematics, Mathematics and Natural Science Faculty, Universitas Tanjungpura*
Jln. Prof. Dr. Hadari Nawawi, Pontianak, 78124, Indonesia

*Corresponding author's e-mail: * neva.satya@math.untan.ac.id*

Article Info

Article History:

Received: 17th June 2025
Revised: 7th October 2025
Accepted: 16th March 2026
Available online: 8th April 2026

Keywords:

Clustering;
MVEP;
Portfolio diversification;
Silhouette coefficient.

ABSTRACT

Investment is a way to manage wealth and achieve financial goals in the future. Stocks are an attractive investment instrument due to their high potential returns, although they also carry significant risks. These risks can be minimized through portfolio diversification. Diversification is carried out by selecting representative stocks from the clustering results. This study aims to construct an optimal portfolio using the Clustering Large Application (CLARA) method and conduct portfolio risk analysis using Value at Risk (VaR). The data used includes IDX80 stock closing prices from November 1, 2024, to January 31, 2025, the financial ratios of IDX80 stocks on December 2024, and the Bank Indonesia (BI-Rate) interest rate from November 2024 to January 2025. The CLARA method produces four stock clusters with a silhouette coefficient of 0.18226. This value indicates a low level of separation between clusters, as there might be overlapping features among the clusters. Representative stocks from each cluster are selected based on the highest Sharpe ratio: SCMA, JPFA, GOTO, and BRIS. The portfolio weights based on MVEP are 15.002% (SCMA), 29.786% (JPFA), 1.858% (GOTO), and 53.354% (BRIS). The VaR calculation shows a potential maximum loss of Rp137,139 in one day, with a 99% confidence level, from an initial investment of Rp10,000,000.

How to cite this article:

S. Pujianti, H. Perdana and N. Satyahadewi., "CONSTRUCTING AN OPTIMAL PORTFOLIO USING CLUSTERING LARGE APPLICATION AND VALUE AT RISK ANALYSIS FOR IDX80 STOCKS," *BAREKENG: J. Math. & App.*, vol. 20, no. 3, pp. 1855-1868, Sep, 2026.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

Investment refers to the allocation of funds in the present with the expectation of generating profits in the future [1]. Investing is done by purchasing assets in the present time and then selling those assets in the future. Investors can invest in real assets and securities or financial assets [2]. Real assets are tangible items such as gold, silver, precious metals, art, property, land, and real estate. In contrast, financial assets consist of various securities such as stocks, bonds, or mutual funds. Among these, stocks are among the most popular investment instruments in the capital market, primarily due to their potential to generate returns.

However, investment has uncertainties regarding the future [1]. The performance of assets can be influenced by many factors, such as changes in global economic conditions, government policies, and market price fluctuations. Due to this uncertainty, it is important for investors to not only focus on the expected returns of stocks but also to understand and manage the risks that may arise. Investors have to build an ideal portfolio to face these uncertainties.

A portfolio can be described as a collection of investments consisting of different kinds of financial instruments, for instance, bonds, stocks, cash, and other instruments. The basis for portfolio construction is to allocate capital to various investment options to reduce investment risk [3]. One way to achieve an optimal portfolio is through a diversification strategy. This strategy involves investing funds in different types of assets to reduce overall risk. A diversification strategy can balance the losses from one asset with the gains from another asset, therefore minimizing the total risk in the portfolio without sacrificing significant return potential.

The Indonesia Stock Exchange 80 (IDX80) is an index consisting of 80 leading stocks on the Indonesia Stock Exchange (BEI) with high liquidity and large market capitalization. Stocks in IDX80 offer attractive investment opportunities, but they also come with market risks that have to be managed well. Therefore, an effective method is required for selecting and grouping stocks as well as managing risks within an investment portfolio. The construction of an optimal portfolio can be achieved by selecting representative stocks from groups (clusters) formed from cluster analysis [4].

Cluster analysis is a method to categorize objects into several groups according to the characteristics that the objects possess. A good cluster is identified by a higher level of similarity in characteristics among the objects within one cluster and has a clear difference from the objects in other clusters [5]. Selecting investment assets from different clusters will enhance diversification and reduce the risk of the formed portfolio because investment assets within the same cluster have similar characteristics. As a result, if one asset suffers a loss, other assets with different characteristics can help minimize the loss.

Several clustering methods have been used in financial data, such as K-Means, hierarchical clustering, and K-Medoids. However, these methods have certain limitations when applied to large datasets or data containing outliers. An outlier is a data that deviates significantly from the other data [6]. K-Means is sensitive to outliers because of the use of the centroid (the average point from one cluster), while hierarchical clustering is computationally intensive for a larger set of data [7]. To overcome these limitations, this study uses CLARA, an extension of K-Medoids that can be used on large time series data and contains outliers [8]. This method works by randomly sampling from large-sized data and searching for one representative sample object (medoids) as the center of the formed cluster. Through the sampling process, the computation time needed for clustering can be reduced [9]. The determination of the best number of clusters can be done using the silhouette coefficient. A method that compares the average distance of a data point within the same cluster to other clusters.

After the formation of clusters, the selection of representative stocks from each cluster for portfolio diversification can be conducted using the Sharpe ratio. This ratio measures investment performance by considering the return relative to the risk incurred [10]. After constructing the portfolio, it is necessary to determine the weights of each stock. One method to obtain an optimal portfolio is the Mean-Variance Efficient Portfolio (MVEP), a method that can provide the highest expected return level for a certain level of risk, on the contrary, has the lowest risk level for a certain expected return level [11].

In addition to constructing a portfolio, risk measurement is also an important aspect of investment. One commonly used measure of risk is Value at Risk (VaR). VaR is a technique used to estimate the maximum potential loss that may occur with a certain level of confidence and over a specific time period [12]. VaR is widely used due to its straightforward interpretability and lw computational complexity.

Given the importance of constructing an optimal and diversified portfolio, this study enhances portfolio construction by integrating a broader set of variables, including expected return, risk, and financial ratios in clustering. Unlike previous studies on clustering-based portfolio construction, which often relied solely on return and risk parameters, this research incorporates multidimensional financial indicators to capture a more comprehensive representation of asset behavior. Furthermore, the distance metric used is specifically adapted to the data's characteristics, ensuring a more precise clustering of assets. The result of this study can be used as a reference for investors aiming to build optimized portfolios within the IDX market.

2. RESEARCH METHODS

2.1 Stock Return

Stock return is the level of investment profit over some period of time [13]. Return can be classified as actual return and expected return. The return that has occurred and is calculated from historical data is called the actual return, while the return that is expected to be received in the future is called the expected return. Actual return can be calculated using Eq. (1)[14].

$$R_{i(t)} = \ln\left(\frac{P_{i(t)}}{P_{i(t-1)}}\right), \tag{1}$$

where:

$R_{i(t)}$: Actual return i-th stock at time t
$P_{i(t)}$: Closing price of i-th stock at time t
$P_{i(t-1)}$: Closing price of i-th stock at time $t-1$

Meanwhile, the expected return can be calculated using Eq. (2).

$$E(R_i) = \frac{\sum_{t=1}^{T} R_{i(t)}}{T}, \tag{2}$$

where:

$E(R_i)$: Expected return of i-th stock
T : Longest observation period

2.2 Risk, Variance, and Covariance

Risk is associated with deviations or divergences of the actual return obtained compared to the expected return [15]. As the deviation increases, the potential profit or loss during the period also increases. Risk can be measured with variance or standard deviation [16]. Variance is calculated using Eq. (3).

$$\sigma_i^2 = \frac{\sum_{t=1}^{T}\left(R_{i(t)} - E(R_i)\right)^2}{T-1}, \tag{3}$$

where:

σ_i^2 : Variance of i-th stock

However, standard deviation is more commonly used to represent the risk. Standard deviation is calculated using Eq. (4).

$$\sigma_i = \sqrt{\sigma_i{}^2} = \sqrt{\frac{\sum_{t=1}^{T}\left(R_{i(t)} - E(R_i)\right)^2}{T-1}}, \tag{4}$$

where:

σ_i : Standard deviation of i-th stock

In the meantime, stock covariance shows the direction of return movement between two assets. If the covariance is positive, it means that both stocks tend to move in the same direction; on the other hand, a

negative covariance means that they move in opposing directions [17]. Stock covariance is calculated with Eq. (5).

$$\sigma_{i;j} = \frac{\sum_{t=1}^{T}\left[\left(R_{i(t)} - E(R_i)\right)\left(R_{j(t)} - E(R_j)\right)\right]}{T-1}, \tag{5}$$

where:

$\boldsymbol{\sigma_{i;j}}$: Stock covariance between *i*-th stock and *j*-th stock

2.3 Multicollinearity Test

Multicollinearity is a statistical condition where two or more independent variables are highly linearly related, potentially making some variables statistically insignificant. It can be detected using the Pearson correlation coefficient using Eq. (6). Multicollinearity happens if the absolute value of the correlation coefficient is above 0.8 [18].

$$r_{x_r,x_l} = \frac{U\left(\sum_{i=1}^{u} x_{i,r}, x_{i,l}\right) - \sum_{i=1}^{u} x_{i,r}\left(\sum_{i=1}^{u} x_{i,l}\right)}{\sqrt{U\left(\sum_{i=1}^{u} x_{i,r}{}^2\right) - \left(\sum_{i=1}^{u} x_{i,r}\right)^2}\sqrt{U\left(\sum_{i=1}^{u} x_{i,l}{}^2\right) - \left(\sum_{i=1}^{u} x_{i,l}\right)^2}}, \tag{6}$$

where:

r_{x_r,x_l} : Correlation coefficient between *r*-th variable and *l*-th variable
U : Total data
$x_{i,r}$: The *i*-th data point in the *r*-th variable
$x_{i,l}$: The *i*-th data point in the *l*-th variable

2.4 Outlier Detection

An outlier is an extreme data point that might reduce the information and lead to distorted results [19] [20]. Outlier detection is performed to help determine the distance metric that best fits the data. The detection is carried out with the boxplot method through Eq. (7).

$$[Q_1 - 1.5 \times IQR\ , Q_3 + 1.5 \times IQR], \tag{7}$$

where:

Q_1 : First quartile
Q_3 : Third quartile
IQR : Interquartile Range $(Q_3 - Q_1)$

Data points falling outside this interval are classified as outliers. These outliers will be retained in the analysis because CLARA is a robust method towards outliers.

2.5 Z-Score Standardization

Data standardization involves transforming data measured on different scales to a uniform scale for comparability. Standardization is necessary because the scale differences between different variables can cause bias in cluster analysis [21]. Standardization can be done using a Z-score using Eq. (8).

$$Z_{i,r} = \frac{x_{i,r} - \mu_r}{\sigma_r}, \tag{8}$$

where:

$Z_{i,r}$: The *i*-th z-score in the *r*-th variable
μ_r : The mean of the *r*-th variable
σ_r : The standard deviation of the *r*-th variable

2.6 Manhattan Distance

The Manhattan distance measures the absolute differences between data [20]. This distance is more robust to outliers compared to the Euclidean distance, which uses squared differences between data. Manhattan distance can be calculated with Eq. (9) [22].

$$d_{i,j} = \sum_{r=1}^{R} \left| x_{i,r} - x_{j,r} \right|, \tag{9}$$

where:

$d_{i,j}$: Manhattan distance between the i-th data with the j-th data

$x_{j,r}$: The j-th data point in the r-th variable

2.7 Clustering Large Application (CLARA)

Cluster analysis is a technique that groups objects into several clusters based on their characteristics. Clustering methods are generally divided into two types: hierarchical methods and partitioning methods. The hierarchical clustering method is generally applied when the number of clusters is not yet determined, whereas the partitioning method needs the initial number of clusters. One example of a partitioning method is CLARA. This method uses medoids as cluster centers, where a medoid is defined as data with minimal average dissimilarity with other data in its cluster [23]. CLARA finds the optimum medoids from multiple samples with a fixed size [24]. Therefore, the CLARA method has a stronger resilience against outliers and is used to handle large-scale data [8].

The minimum sample size for CLARA is $(40 + 2k)$, where k represents the number of clusters [25]. The CLARA method consists of two stages: the build stage and the swap stage. In the build stage, a number of k medoids are selected from the sample to calculate the distances to other data. In the swap stage, medoids are replaced to find the minimum distance among the data. CLARA is an iterative method implemented as follows.

1. Determine the initial number of clusters (k).
2. Select best sample with a size of $40 + 2k$.
3. Randomly select k initial medoids from the best sample.
4. Calculate Manhattan distance between each data and each medoid using Eq. (9).
5. Assign each data point into the appropriate cluster based on the smallest distance to the medoids.
6. Calculate the total Manhattan distance of the first iteration.
7. Select new medoids of k randomly from best sample.
8. Calculate Manhattan distance between each data and each new medoid.
9. Assign each data point into the appropriate cluster based on the smallest distance to the new medoids.
10. Calculate the total Manhattan distance of the second iteration.
11. Calculate the total distance (S) by computing total Manhattan distance of the second iteration − total Manhattan distance of the first iteration. If $S \leq 0$, repeat the process from $7 - 11$. Iteration stops if $S > 0$.

2.8 Silhouette Coefficient

The Silhouette coefficient is used to determine the initial number of clusters. This method calculates the average distance of one data with other data in the same cluster compared to other data in different cluster [26]. The higher the Silhouette coefficient value for k clusters, the more optimal the clusters are. The Silhouette coefficient is calculated as follows [27].

1. Calculating the average distance of the i-th data point to other data in the same cluster, using Eq. (10).

$$a_i(k) = \frac{1}{A_k - 1} \sum_{j=1}^{A_k} \left| d_{i,k} - d_{j,k} \right|, j \neq i, \tag{10}$$

where:

$a_i(k)$: Average distance between the i-th data to other data in one cluster

A_k : Number of data in the same cluster

$|d_{i,k} - d_{j,k}|$: Distance of *i*-th data and *j*-th data within one cluster

2. Calculating the average distance of the *i*-th data point to other data in different cluster, using Eq. (11).

$$d_i(k,C) = \frac{1}{C}\sum_{j=1}^{C}|d_{i,k} - d_{j,C}|, k \neq C. \quad (11)$$

Next, selecting the minimum value using Eq. (12).

$$b_i(k) = \min d_i(k,C), \quad (12)$$

where:

$d_i(k,C)$: Average distance between the *i*-th data to other data in different cluster
C : Number of data in other cluster
$|d_{i,k} - d_{j,C}|$: Distance of *i*-th data to *j*-th data in other cluster
$b_i(k)$: Minimum of $|d_{i,k} - d_{j,C}|$

3. Calculating Silhouette index for *i*-th data using Eq.(13) [28].

$$SI_{i,k} = \frac{b_i(k) - a_i(k)}{\max(a_i(k), b_i(k))}, \quad (13)$$

where:

$SI_{i,k}$: Silhouette index of *i*-th data

4. Calculating Silhouette index for all data in one cluster using Eq. (14).

$$SI_k = \frac{1}{A_k}\sum_{i=1}^{A_k} SI_{i,k}, \quad (14)$$

where:

SI_k : Silhouette index for all data in *k*-th cluster

5. Calculating Silhouette coefficient using Eq. (15).

$$SC = \frac{\sum_{k=1}^{K} A_k \times SI_k}{\sum_{k=1}^{K} A_k}, \quad (15)$$

where:

SC : Silhouette coefficient

2.9 Sharpe Ratio

Sharpe ratio is a measure used to evaluate the performance of an investment in generating returns while considering investment risk [10]. Higher ratio values indicate better asset performance. Sharpe ratio is calculated using Eq. (16)[29].

$$SR = \frac{E(R_i) - R_f}{\sigma_i}, \quad (16)$$

where:

SR : Sharpe ratio
R_f : Daily risk-free rate

2.10 Mean-Variance Efficient Portfolio (MVEP)

Mean-Variance Efficient Portfolio (MVEP) is a portfolio designed to have the lowest level of variance compared to all combinations of portfolios that can be formed [30]. The combination of stocks that produces the lowest variance is obtained by optimizing the stocks' respective weight in the portfolio. Weighting is done to determine the proportion of funds in each stock so that the total portfolio risk can be minimized. The total weight of the stocks equals to 1. The weight of stocks in MVEP is calculated using Eq. (17).

$$w = \frac{\Sigma^{-1} 1_N}{1_N{}^T \Sigma^{-1} 1_N}, \tag{17}$$

where:

w : Stock weight
Σ^{-1} : Inverse of the variance-covariance matrix of stock returns in the portfolio
1_N : One-dimensional vector N × 1
$1_N{}^T$: Transpose of 1_N
N : Number of stocks in portfolio

2.11 Portfolio Return and Value at Risk (VaR)

Portfolio return is one of the main indicators used to measure investment performance. Portfolio return is calculated by combining the individual returns of each asset in the portfolio multiplied by their respective weights. The return can be calculated using Eq. (18) [31].

$$R_{p,t} = \sum_{i=1}^{N} w_i R_{i(t)}, \tag{18}$$

where:

$R_{p,t}$: Portfolio return at time t
w_i : Weight of the i-th stock

Aside from portfolio return, investors must also consider the possible losses that may occur to the portfolio. Value at Risk (VaR) estimates the maximum loss an investor might suffer during a certain period with a certain confidence level [32]. One approach to VaR value is historical simulation. This approach calculates the VaR value based on the past value of an asset without requiring the assumption of a normal distribution of returns [33]. VaR can be calculated using Eq. (19).

$$VaR = V_0 \times P_\alpha \times \sqrt{t}, \tag{19}$$

where:

VaR : Maximum potential loss
V_0 : Initial investment funds
P_α : α-th percentile
$\sqrt{t}$: Investment period

The α-th percentile describes the expected loss incurred by investors at level α. The calculation of the α-th percentile position is obtained from portfolio return data that has been sorted from smallest to largest. It can be calculated using Eq. (20).

$$P_\alpha = \alpha \times T, \tag{20}$$

where:

α : Significance level

3. RESULTS AND DISCUSSION

This analysis uses three sets of secondary data. The first data is IDX80 stocks closing price data for the period 1 November 2024 − 31 January 2025 from the Yahoo Finance site. This data is used to calculate stock returns with Eq. (1). Then, the expected return, risk, variance, and covariance of the stocks will be calculated based on the return calculation obtained. The expected return and risk will be used in cluster analysis. In addition, variance and covariance data are used to calculate the Mean-Variance Efficient Portfolio (MVEP).

The second data is the financial ratio data for IDX80 stocks as of December 2024. The data is used for cluster analysis. Data was obtained from the publication of each company's 2024 Financial Statement on the Indonesia Stock Exchange (IDX) site. The financial ratio consists of Debt to Asset Ratio (DAR), Debt to

Equity Ratio (DER), Return on Asset (ROA), Return on Equity (ROE), Gross Profit Margin (GPM), and Net Profit Margin (NPM).

The third data is the average Bank Indonesia interest rate (BI-Rate) from November 2024 to January 2025, which was 5.917%. This data is used as a risk-free rate in calculating the Sharpe ratio, which is the reference for selecting representative stocks from the clusters formed.

The variables used for cluster analysis are as follows.

Table 1. Variables For Cluster Formation

Code	Variables
X_1	Expected Return
X_2	Risk
X_3	DAR
X_4	DER
X_5	ROA
X_6	ROE
X_7	GPM
X_8	NPM

The expected return and risk values for IDX80 shares are obtained from Eqs. (2) and (4), respectively. The cluster analysis is carried out using RStudio software with the 'cluster' and 'factoextra' packages.

3.1 Multicollinearity Test

Conducting a multicollinearity test is essential to assess the presence of a strong linear correlation among the variables involved in the analysis. If multicollinearity occurs, then the variables need to be removed because the presence of multicollinearity results in several research variables being statistically insignificant. By using Eq. (6), the result of the multicollinearity test is presented below.

Table 2. Correlation Coefficient Values

Variables	X_1	X_2	X_3	X_4
X_1	1	-0.15515	-0.01876	-0.07115
X_2	-0.15515	1	-0.17389	-0.15827
X_3	-0.01876	-0.17389	1	0.77135
X_4	-0.07115	-0.15827	0.77135	1
X_5	-0.16745	-0.02959	-0.25035	-0.19353
X_6	-0.09153	-0.04970	0.25295	0.30118
X_7	0.01758	0.21703	-0.33344	-0.51378
X_8	-0.32732	0.20924	0.08190	0.16384
Variables	X_5	X_6	X_7	X_8
X_1	-0.16745	-0.09153	0.01758	-0.32732
X_2	-0.02959	-0.04970	0.21703	0.20924
X_3	-0.25035	0.25295	-0.33344	0.08190
X_4	-0.19353	0.30118	-0.51378	0.16384
Variables	X_5	X_6	X_7	X_8
X_5	1	0.59048	0.30633	0.39431
X_6	0.59048	1	0.09564	0.18644
X_7	0.30633	0.09564	1	-0.02475
X_8	0.39431	0.18644	-0.02475	1

Table 2 shows that there is no absolute value of the correlation coefficient that is greater than 0.8. Therefore, it can be concluded that there is no multicollinearity, so that the analysis can be continued using the existing research variables.

3.2 Outlier Detection

Outlier detection is done using a boxplot graph. Data that has a value smaller than the lower limit or greater than the upper limit is categorized as an outlier. The lower limit and the upper limit are calculated using Eq. (7). The boxplot visualization is presented in Fig. 1.

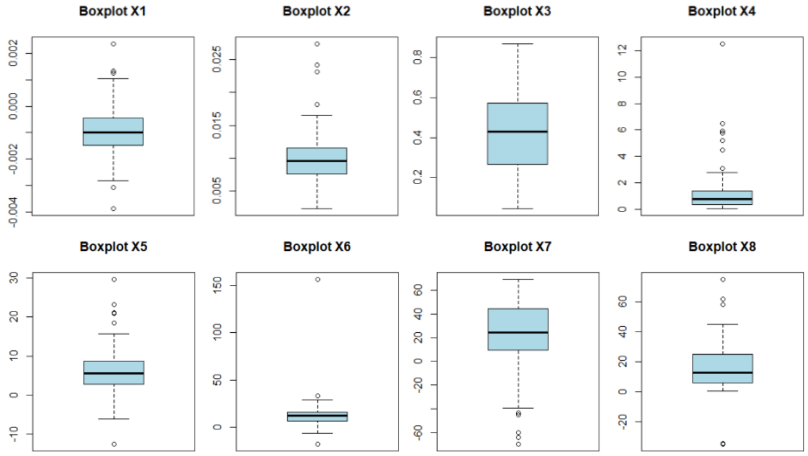

Figure 1. Boxplot Graph

Fig. 1 reveals that there are 7 out of 8 variables that contains outliers. The outliers are identified by points outside the boxplot whiskers (lower limit and upper limit). The variables are $X_{1,}$, $X_{2,}$, $X_{4,}$, $X_{5,}$, $X_{6,}$, $X_{7,}$, and $X_{8,}$. Due to the presence of outlier data, the distance used is the Manhattan distance in Eq. (9).

3.3 Determining the Best Number of Clusters with Silhouette Coefficient

Before conducting clustering analysis, Eq. (8) is applied to standardize the data, matching the scale of measurement across all variables. After standardizing, the best number of clusters is sought.

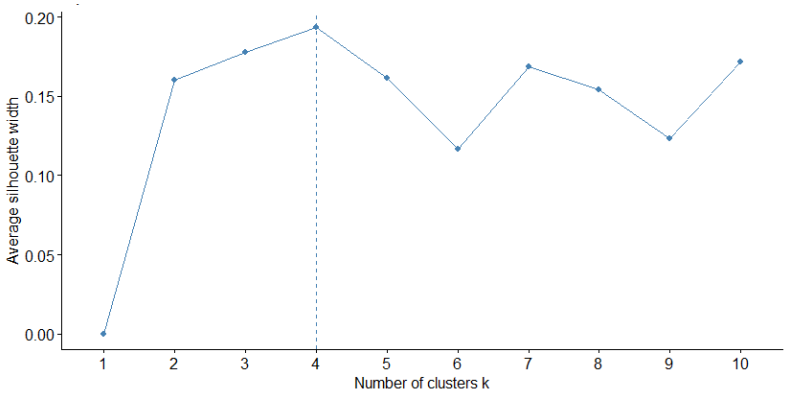

Figure 2. Plot of Best Number of Clusters

Fig. 2 shows the use of the Silhouette coefficient method to get the best number of clusters. Determining the best number of clusters is based on the highest average Silhouette value, indicating 4 clusters as the best cluster number. The value of the Silhouette coefficient is 0.18226. However, this low value indicates a weak separation between clusters, implying a possibility of resemblance among the clusters. This constraint may reduce the effectiveness of clustering to construct a diversified portfolio.

3.4 CLARA

The 'clara.res' function initiates CLARA by selecting the best sample. The best sample size is (40+2k), where k is the number of clusters (k = 4). Therefore, the number of best sample is 48 data.

Table 3. Best Sample

Best Sample							
ADRO	AKRA	AMMN	AMRT	ANTM	ASII	AUTO	AVIA
BBCA	BBNI	BBRI	BBTN	BNGA	BRMS	BRPT	BSDE
BTPS	CMRY	EMTK	ESSA	GJTL	GOTO	HRUM	ICBP
INCO	INDF	INTP	ISAT	ITMG	JPFA	KLBF	MAPA
MBMA	MDKA	MEDC	MIDI	MIKA	MNCN	MPMX	NCKL
PGEO	PNLF	PWON	SMDR	SMGR	SMRA	SRTG	TOWR

Table 3 shows the best sample from RStudio, consisting of 48 stocks, from which four initial medoids will be selected. The initial medoids are ESSA as the first medoid (M_1), GJTL as the second medoid (M_2), SMGR as the third medoid (M_3), and BBRI as the fourth medoid (M_4). After determining the initial medoids, the other stocks will be placed in appropriate clusters based on the nearest Manhattan distance calculated using Eq. (9). The process will be repeated by selecting new medoids until total distance (S) > 0. The CLARA result is presented in Table 4 below.

Table 4. Clustering Result

Cluster	Stocks	Number of Members
1	ACES, ADMR, ADRO, ANTM, AUTO, AVIA, BRMS, BSDE, BTPS, CMRY, CTRA, EMTK, ESSA, INTP, ITMG, KLBF, MAPA. MIKA, MNCN, NCKL, PGEO, PWON, SCMA, SIDO, SRTG, TKIM	26
2	AKRA, AMMN, AMRT, ASII, BFIN, BRPT, CPIN, ELSA, ENRG, ERAA, EXCL, GJTL, HEAL, ICBP, INDF, INDY, INKP, ISAT, JPFA, JSMR, MAPI, MEDC, MIDI, MPMX, MTEL, MYOR, PGAS, PNLF, PTBA, SMDR, SMRA, TLKM, UNTR	33
3	ARTO, BMTR, BUKA, GGRM, GOTO, HRUM, INCO, MBMA, MDKA, PANI, SMGR	11
4	BBCA, BBNI, BBRI, BBTN, BMRI, BNGA, BRIS, NISP, TOWR, UNVR	10

3.5 Choosing Representative Stocks

After obtaining the results of the cluster analysis, the next stage is to select representative stocks from each cluster using the Sharpe ratio in Eq. (16). The average risk-free return is 5.917% (0.05917), which, when converted to a daily rate, becomes 0.00016. Representative stock selection is done by selecting the highest Sharpe ratio for each cluster.

Table 5. Representative Stocks

Cluster	Representative Stocks	Sharpe Ratio	Sector
1	SCMA	0.12135	Consumer Cyclicals
2	JPFA	0.08389	Consumer Non-Cyclicals
3	GOTO	0.06647	Technology
4	BRIS	-0.03751	Finance

Although, the Sharpe ratio of BRIS is a negative value, it still has the highest Sharpe ratio among all stocks in the fourth cluster. Table 5 presents representative stocks from each cluster, reflecting portfolio diversification goals, as indicated by their diverse sector origins.

3.6 Weighting of Stocks in a Portfolio

Calculation of stock weights using the Mean-Variance Efficient Portfolio (MVEP) method begins by finding the variance and covariance of stock returns to form a variance-covariance matrix. Variance is calculated using Eq. (3), while covariance is calculated using Eq. (5). The next process is to calculate the stock weight using Eq. (17). The stock weighting results in percentage is presented in Table 6.

Table 6. Stock Weight in Portfolio

Stock	Weight (%)
SCMA	15.002
JPFA	29.786
GOTO	1.858
BRIS	53.354

Referring to Table 6, it is known that the weight of SCMA stock is 15.002%, JPFA is 29.786%, GOTO is 1.858%, and BRIS is 53.354%.

3.7 Portfolio Return and Value at Risk Calculations

Portfolio return is calculated based on the returns of each share with the proportion or weight of the investment obtained. The return is calculated using Eq. (18). After obtaining the portfolio return for the time

period, the return is sorted from lowest to highest return to get the Value at Risk (VaR). VaR is calculated using Eq. (19) with 1% significance level as follows.

Table 7. Value at Risk

V_0	P_α	$\sqrt{t}$	VaR
Rp 10,000,000	−0.0137139	1	−Rp 137,139

The value of P_α is obtained from Eq. (20), while the value of t is made to calculate VaR for one day ahead of January 31, 2025. As shown in Table 6, the maximum loss with an initial investment fund of Rp 10,000,000 at a 99% confidence level over one day period is Rp 137,139.

4. CONCLUSION

Through the application of CLARA, four stock clusters were formed. From each cluster, a representative stock was selected based on its highest Sharpe ratio to achieve an optimal portfolio. The MVEP calculation showed optimal portfolio weights of 15.002% for SCMA (consumer cyclicals sector), 29.786% for JPFA (consumer non-cyclicals sector), 1.858% for GOTO (technology), and 53.354% for BRIS (finance). As investment is inseparable from risk, the optimized portfolio's maximum potential loss in one day, at a 99% confidence level and an initial investment of Rp10,000,000, is estimated at Rp 137,139.

Future research can consider using alternative distance metrics for datasets containing outliers, such as the Mahalanobis distance, which accounts for correlations between variables and is more robust to the presence of extreme values. In addition, the selection of representative stocks within each cluster could benefit from the use of alternative performance indicators, such as the Sortino ratio and Treynor ratio, which may provide more nuanced assessments of risk-adjusted returns. Moreover, portfolio risk measurement could be enhanced by adopting more complex methods that account for volatile stock market conditions, such as Conditional Value at Risk (CVaR), which provides a more comprehensive estimation of potential extreme losses beyond the traditional Value at Risk (VaR) approach.

Author Contributions

Sania Pujianti: Conceptualization, Formal Analysis, Methodology, Writing - Original Draft. Hendra Perdana: Data Curation, Project Administration, Software, Writing - Review and Editing. Neva Satyahadewi: Formal Analysis, Supervision, Validation, Visualization - Original Draft. All authors discussed the results and contributed to the final manuscript.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Acknowledgment

The author would like to express sincere gratitude to the Study Program of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Tanjungpura, for the guidance, support, and facilities provided throughout the research process.

Declarations

The authors declare no competing interest.

Declaration of Generative AI and AI-assisted Technologies

AI-assisted technology (ChatGPT) was used to support sentence restructuring and improve clarity. The authors affirm that the core ideas, arguments, data analysis, and conclusions are original and not generated by AI. All AI-assisted edits were critically reviewed and validated by the authors.

REFERENCES

[1] Z. Bodie, A. Kane, and A. J. Marcus, *INVESTMENTS*, 11th ed. New York: McGraw Hill Education, 2017.

[2] Z. Bodie, R. C. Merton, and R. T. Thakor, *PRINCIPLES OF FINANCE*. Cambridge: Cambridge University Press, 2025. doi: https://doi.org/10.1017/9781108982610.

[3] N. Jha, R. S. Mishra, and S. M. Bhome, *INVESTMENT ANALYSIS AND PORTFOLIO MANAGEMENT*. Mumbai: Himalaya Publishing House, 2016. [Online]. Available: www.himpub.com

[4] L. Gubu, D. Rosadi, and Abdurakhman, "PEMBENTUKAN PORTOFOLIO SAHAM MENGGUNAKAN KLASTERING TIME SERIES K-MEDOID DENGAN UKURAN JARAK DYNAMIC TIME WARPING," *J. Apl. Stat. Komputasi Stat.*, vol. 13, no. 2, pp. 35–46, 2021, doi: https://doi.org/10.34123/jurnalasks.v13i2.295.

[5] S. J. Sinaga, N. Satyahadewi, and H. Perdana, "DETERMINING THE OPTIMUM NUMBER OF CLUSTERS IN HIERARCHICAL CLUSTERING USING PSEUDO-F," *Euler J. Ilm. Mat. Sains dan Teknol.*, vol. 11, no. 2, pp. 372–382, Dec. 2023, doi: https://doi.org/10.37905/euler.v11i2.23113.

[6] C. C. Aggarwal, *OUTLIER ANALYSIS*. Springer International Publishing, 2017. doi: https://doi.org/10.1007/978-3-319-47578-3.

[7] P.-N. Tan, M. Steinbach, A. Karpatne, and V. Kumar, *INTRODUCTION TO DATA MINING*, 2nd ed. New York: Pearson Education, 2019.

[8] R. Lapiza, Syafriandi, N. Amalita, and D. Fitria, "GROUPING THE DISTRICTS IN SUMATERA REGION BASED ON ECONOMIC DEVELOPMENT INDICATORS USING K-MEDOIDS AND CLARA METHODS," *UNP J. Stat. Data Sci.*, vol. 1, no. 1, pp. 16–22, Feb. 2023, doi: https://doi.org/10.24036/ujsds/vol1-iss1/13.

[9] A. P. Ayuni, D. Kusnandar, and S. Martha, "IMPLEMENTASI ALGORITMA K-MEDOIDS DAN CLUSTERING LARGE APPLICATIONS (CLARA) DENGAN OPTIMASI SILHOETTE COEFFICIENT (STUDI KASUS: PENGELOMPOKAN INDEKS PEMBANGUNAN MANUSIA BERDASARKAN KABUPATEN/KOTA DI INDONESIA)," *Bul. Ilm. Math. Stat. dan Ter.*, vol. 13, no. 2, pp. 191–200, 2024, doi: https://doi.org/10.26418/bbimst.v13i2.76959.

[10] A. N. Pratama, N. Satyahadewi, and E. Sulistianingsih, "ANALYSIS OF OPTIMAL PORTFOLIO FORMATION ON IDX30 INDEXED STOCK WITH THE MEAN ABSOLUTE DEVIATION METHOD," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 18, no. 3, pp. 1753–1764, Jul. 2024, doi: https://doi.org/10.30598/barekengvol18iss3pp1753-1764.

[11] A. Silvia and M. Rosha, "ANALISIS PERBANDINGAN PORTOFOLIO OPTIMAL MODEL MARKOWITZ DAN MODEL MVEP (STUDI KASUS SAHAM LQ-45 DI BURSA EFEK INDONESIA DI MASA PANDEMI COVID-19)," *J. Math. UNP*, vol. 9, no. 2, pp. 125–131, 2024.

[12] D. R. Prihatiningsih, D. A. I. Maruddani, and R. Rahmawati, "VALUE AT RISK (VAR) DAN CONDITIONAL VALUE AT RISK (CVAR) DALAM PEMBENTUKAN PORTOFOLIO BIVARIAT MENGGUNAKAN COPULA GUMBEL," *J. Gaussian*, vol. 9, no. 3, pp. 326–335, 2020, doi: https://doi.org/10.14710/j.gauss.v9i3.28913.

[13] S. Margun, N. Satyahadewi, and H. Perdana, "ANALISIS KINERJA PORTOFOLIO OPTIMAL SAHAM LQ-45 DENGAN METODE MEAN-GIINI MENGGUNAKAN INDEKS SHARPE," *Bul. Ilm. Mat. Stat. dan Ter.*, vol. 11, no. 3, pp. 423–430, 2022, doi: 10.26418/bbimst.v11i3.54959.

[14] M. N. Siddikee, "EFFECT OF DAILY DIVIDEND ON ARITHMETIC AND LOGARITHMIC RETURN," *J. Financ. Data Sci.*, vol. 4, no. 4, pp. 247–272, Dec. 2018, doi: https://doi.org/10.1016/j.jfds.2018.06.001

[15] M. Azis, S. Mintarti, and M. Nadir, *MANAJEMEN INVESTASI, FUNDAMENTAL, TEKNIKAL, PERILAKU INVESTOR, DAN RETURN SAHAM*. Yogyakarta: Deepublish, 2015.

[16] D. Schoenmaker and W. Schramade, *CORPORATE FINANCE FOR LONG-TERM VALUE*. Cham: Springer, 2023. doi: https://doi.org/10.1007/978-3-031-35009-2

[17] T. Latunde, L. S. Akinola, and D. D. Dare, "ANALYSIS OF CAPITAL ASSET PRICING MODEL ON DEUTSCHE BANK ENERGY COMMODITY," *Green Financ.*, vol. 2, no. 1, pp. 20–34, 2020, doi: https://doi.org/10.3934/GF.2020002.

[18] N. Shrestha, "DETECTING MULTICOLLINEARITY IN REGRESSION ANALYSIS," *Am. J. Appl. Math. Stat.*, vol. 8, no. 2, pp. 39–42, Jun. 2020, doi: https://doi.org/10.12691/ajams-8-2-1.

[19] K. Wada, "OUTLIERS IN OFFICIAL STATISTICS," *Japanese J. Stat. Data Sci.*, vol. 3, no. 2, pp. 669–691, Dec. 2020, doi: https://doi.org/10.1007/s42081-020-00091-y.

[20] P. R. Fitrayana and D. R. S. Saputro, "ALGORITME CLUSTERING LARGE APPLICATION (CLARA) UNTUK MENANGANI DATA OUTLIER," *Prism. Pros. Semin. Nas. Mat.*, vol. 5, pp. 721–725, 2022, [Online]. Available: https://journal.unnes.ac.id/sju/index.php/prisma

[21] D. Sawitri, "PERAN DEEP LEARNING DAN BIG DATA DALAM MENDETEKSI MASALAH KEUANGAN," *J. Teknol. Inf.*, vol. 6, no. 1, pp. 193–207, 2025, doi: https://doi.org/10.46576/djtechno.v6i1.6037.

[22] J. Rathee, P. Kaur, and A. Singh, "FUZZY CLUSTERING BASED NOISY IMAGE SEGMENTATION OF MRI/CT SCAN BRAIN TUMOR IMAGES USING DIFFERENT DISTANCE METRICS AS SIMILARITY MEASURE," *SN Comput. Sci.*, vol. 5, no. 777, Aug. 2024, doi: https://doi.org/10.1007/s42979-024-03102-x.

[23] M. Z. Rodriguez *et al.*, "CLUSTERING ALGORITHMS: A COMPARATIVE APPROACH," *PLoS One*, vol. 14, no. 1, pp. 1–34, Jan. 2019, doi: https://doi.org/10.1371/journal.pone.0210236.

[24] V. P. Senthilnathan, M. Singaravelu, S. Rajendran, and S. Srinivas, "A CLUSTERING-METAHEURISTIC-SIMULATION APPROACH TO DETERMINE AIR TAXI OPERATING SITE LOCATION," *Transp. Res. Interdiscip. Perspect.*, vol. 29, Jan. 2025, doi: https://doi.org/10.1016/j.trip.2025.101330.

[25] E. Schubert and P. J. Rousseeuw, "FAST AND EAGER K-MEDOIDS CLUSTERING: O(K) RUNTIME IMPROVEMENT OF THE PAM, CLARA, AND CLARANS ALGORITHMS," *Inf. Syst.*, vol. 101, Nov. 2021, doi: https://doi.org/10.1016/j.is.2021.101804.

[26] A. Et-Taleby, M. Boussetta, and M. Benslimane, "FAULTS DETECTION FOR PHOTOVOLTAIC FIELD BASED ON K-MEANS, ELBOW, AND AVERAGE SILHOUETTE TECHNIQUES THROUGH THE SEGMENTATION OF A THERMAL IMAGE," *Int. J. Photoenergy*, vol. 2020, 2020, doi: https://doi.org/10.1155/2020/6617597

[27] D. A. I. C. Dewi and D. A. K. Pramita, "ANALISIS PERBANDINGAN METODE ELBOW DAN SILLHOUETTE PADA ALGORITMA CLUSTERING K-MEDOIDS DALAM PENGELOMPOKAN PRODUKSI KERAJINAN BALI," *J. MATRIX*, vol. 9, no. 3, pp. 102–109, 2019, doi: https://doi.org/10.31940/matrix.v9i3.1662.

[28] J. Raymaekers and P. J. Rousseeuw, "SILHOUETTES AND QUASI RESIDUAL PLOTS FOR NEURAL NETS AND TREE-BASED CLASSIFIERS," *J. Comput. Graph. Stat.*, vol. 31, no. 4, pp. 1332–1343, 2022, doi: https://doi.org/10.1080/10618600.2022.2050249.

[29] A. H. Manurung, A. Manurung, N. M. Machdar, and J. Sijabat, "STOCK SELECTION USING SEMI-VARIANCE AND BETA TO CONSTRUCT PORTFOLIO AND EFFECT MACRO-VARIABLE ON PORTFOLIO RETURN," *Turkish J. Comput. Math. Educ.*, vol. 15, no. 1, pp. 14–25, 2024, doi: https://doi.org/10.61841/turcomat.v15i1.14350.

[30] K. Hanum, Tarno, and Sudarno, "OPTIMASI VALUE AT RISK REKSA DANA MENGGUNAKAN METODE ROBUST EXPONENTIALLY WEIGHTED MOVING AVERAGE (ROBUST EWMA) DENGAN PROSEDUR VOLATILITY UPDATING HULL AND WEHITE," *J. GAUSSIAN*, vol. 6, no. 3, pp. 375–384, 2017, doi: 10.14710/j.gauss.6.3.375-384.

[31] A. B. Schmidt, "MANAGING PORTFOLIO DIVERSITY WITHIN THE MEAN VARIANCE THEORY," *Ann. Oper. Res.*, vol. 282, pp. 315–329, Nov. 2019, doi: https://doi.org/10.1007/s10479-018-2896-x.

[32] I. E. Etuk, Y. Musa, and U. Gulumbe, "Estimating and Predicting Value at Risk in Selected Banks of Nigeria Stock Market," *Int. J. Stat. Appl.*, vol. 9, no. 4, pp. 117–121, 2019, doi: 10.5923/j.statistics.20190904.03.

[33] A. Solihatun, L. Gubu, Aswani, E. Cahyono, and L. O. Saidi, "PERHITUNGAN VALUE AT RISK (VAR) PADA PORTOFOLIO SAHAM IDX SEKTOR KEUANGAN (IDXFINANCE) MENGGUNAKAN METODE SIMULASI HISTORIS (HISTORICAL SIMULATION METHOD)," *J. Mat. Komputasi dan Stat.*, vol. 3, no. 1, pp. 245–254, 2023, doi: https://doi.org/10.33772/jmks.v3i1.32