Jurnal Matematika Integratif.
Vol. 20, No. 2 (2024), pp. 149–172.

p-ISSN:1412-6184, e-ISSN:2549-903
doi:10.24198/jmi.v20.n2.53351.149-172

Construction of Stock Portfolios Based On K-means Clustering of Continuous Trend Features

Hilmi Firmansyah, Dedi Rosadi

Department of Mathematics, Gadjah Mada University
Bulaksumur, Yogyakarta, 55281, Special Region of Yogyakarta, Indonesia
Email: hilmifirman2000gmail.com

Abstract
Optimal portfolio formation to reduce investment risk and increase returns is a concern for investors. There are various problems when investing with portfolio formation. First, it is difficult to select a pool of assets for portfolio formation. When the number of potential assets is relatively large, it will be difficult to select assets that fulfill portfolio formation and appropriate weights. Traditional portfolio theory such as "Markowitz portfolio theory" is only used for the calculation of appropriate weights but cannot be used to automatically select assets from a pool of assets. Secondly, traditional portfolio theory calculates its weights only based on the covariance relationship between different stocks and market data is not taken into account. Thirdly, the sharpe ratio calculation is used to evaluate investment returns but does not consider risk aversion when stocks go down. Therefore, this thesis aims at portfolio formation based on continuous trend features. Utilization of k-means clustering is used to group assets, divide different types of asset pools, and calculation of sharpe ratio based on continuous trend features to avoid downside risk. In addition, it is also combined with the calculation of equal weight for each asset, inverse volatility, risk parity, and Markowitz portfolio theory.

Keywords: *Stock Portfolio, K-means Clustering, Sharpe Ratio, Continuous Trend Features, Portfolio Theory.*

1. Introduction

Investment is an activity of placing a number of funds currently owned in the hope of obtaining future profits, both individually and institutionally. Investment in risky assets such as stocks will not only generate profits (returns), but also have to face losses (risk). Therefore, to reduce risk without sacrificing profits, a portfolio is formed. A portfolio is a collection of assets with certain proportions formed by investors. The formation of a good portfolio can better increase investment returns and reduce investment risks. The portfolio selection published by Markowitz (1952) is generally recognized as the origin of "modern portfolio theory". Since then, modern portfolio research has been largely based on Markowitz's portfolio theory [20]. In Markowitz's Mean-Variance (MV) portfolio model, the mean and standard deviation are used

2000 Mathematics Subject Classification: 62H30, 91G80

Received: 13-2-24, accepted: 12-11-24 .

to measure return and risk. Covariance is used as a measure of risk in the Mean-Variance (MV) portfolio model for quantitative finance. The Mean-Variance (MV) portfolio model became the standard for portfolio performance assessment that facilitates the formation of an efficient portfolio to achieve the highest rate of return at the lowest level of systematic risk or non-systematic risk [30]. Thus, asset diversification in investment can reduce individual risk.

The development of machine learning has combined various methods in machine learning to study Mean-Variance Portfolio Optimization (MVPO). In addition to MVPO, many methods have been developed to calculate portfolio weights based on risk factor investment. Risk factor investment is a type of investment by allocating capital based on risk factors [23], which aims to allocate capital more effectively according to the needs and preferences of investors. One of the risk factor investment methods is the market risk factor of the Capital Asset Pricing Model (CAPM) [27]. Then, risk parity is a diversification method from the point of view of the risk contribution of each stock, that is, the risk contribution of each stock to the portfolio is basically the same. The goal of risk parity is to keep the volatility of the total portfolio unchanged through changes in the external economic environment, which indicates that the volatility level of the overall portfolio remains stable. This means that risk parity refers to the method in which each portfolio stock makes the same contribution to the overall risk, but the determination of the risk budget is different.

There are five risk-based portfolio formation methods including Inverse Volatiliy (IV) [2], Equal Risk Contribution (ERC) [19], Alpha Risk Parity (ARP) [1], Maximum Diversification (MD) [8], and Diversified Risk Parity (DRP) [17]. Other risk parity methods include beta risk parity, systematic risk parity, and inverse variance. The above methods have their own advantages and disadvantages. The mean-variance portfolio is used to form an efficient portfolio with the highest return at low risk. Whereas, the risk parity portfolio tends to produce low return values because it allocates a greater weight to low volatility stocks resulting in lower returns.

The mean-variance and risk parity methods provide weights that are calculated based on the covariance relationship between different stocks. In data processing, covariance is usually represented by the angle of the data vector. In vector spaces, cosine is usually used to determine the similarity between two objects. For each vector a and b in vector space, the cosine equation is given in Equation (1):

$$\cos(a,b)=\frac{\sum_{i=1}^{N}a_ib_i}{\sqrt{\sum_{i=1}^{N}a_i^2\sum_{i=1}^{N}b_i^2}} \tag{1}$$

Equation (2) represents the Pearson correlation coefficient as follows:

$$\begin{aligned}\rho(a,b)&=\frac{cov(a,b)}{\sigma_a\sigma_b}\\&=\frac{\sum_{i=1}^{N}(a_i-\bar{a})(b_i-\bar{b})}{\sqrt{\sum_{i=1}^{N}(a_i-\bar{a})^2}\sqrt{\sum_{i=1}^{N}(b_i-\bar{b})^2}}\\&=\frac{\sum_{i=1}^{N}\tilde{a}_i\tilde{b}_i}{\sqrt{\sum_{i=1}^{N}\tilde{a}_i^2\sum_{i=1}^{N}\tilde{b}_i^2}}\\&=\cos\left(\tilde{a},\tilde{b}\right)\end{aligned} \tag{2}$$

with $\tilde{a}_i=a_i-\bar{a}$ and $\tilde{b}_i=b_i-\bar{b}$. Based on Equation (1) and Equation (2) above, it can be seen that the covariance is equal to the Pearson correlation coefficient and only the Pearson correlation coefficient and variance are related between any two vectors without considering the categorized information of each stock.

The theory of portfolio weight calculation is a static method. Once the weights are calculated, there is no reasonable approach to handle the position. If there are stocks in the portfolio that show continuous decline over a long period of time, then the weight calculation method is only used to calculate the weight through correlation and the situation of stock sales is not taken into account (i.e. avoiding the risk of stock decline that is not considered for all weight calculation methods) as well as increasing the risk of portfolio net worth draw down. Therefore, this study will discuss the formation of stock portfolios based on the continuous trend features. The continuous trend feature refers to a pattern of stock price movement that shows a certain tendency or direction for a certain period of time. This feature can provide an indication of the possible continuation of the direction of price movement in the future with an increasing pattern or a decreasing pattern.

The selection of the k-means clustering method is because this clustering is a distance algorithm-based clustering that meets the requirements according to the distance information between stocks and can determine the stock number of each cluster which also meets the requirements of the characteristics of each stock. The k-means clustering method is quite simple. If a simple algorithm can get good results, it can illustrate its effectiveness better than other methods. The main role of clustering is to provide distance information based on the continuous trend features of each stock. According to the continuous trend features of stocks, prediction models are trained which can effectively predict the changes in the continuous trend. Then, a calculation method using Sharpe ratio and other metrics is performed where the results of the calculation method are compared and the optimal method used to achieve efficient portfolio weighting is obtained.

2. Methodologies

2.1. **Diagram of Portfolio Construction.** The portfolio formation diagram carried out in this study consists of various stages, starting from data processing and segmentation, stock clustering, stock selection for portfolio formation, and portfolio weight calculation as shown in Figure 1. In addition, it also utilizes the RRCT algorithm to calculate the revised return on each stock listed in Algorithm 1 and the PCTM algorithm in the formation of a stock portfolio listed in Algorithm 2.

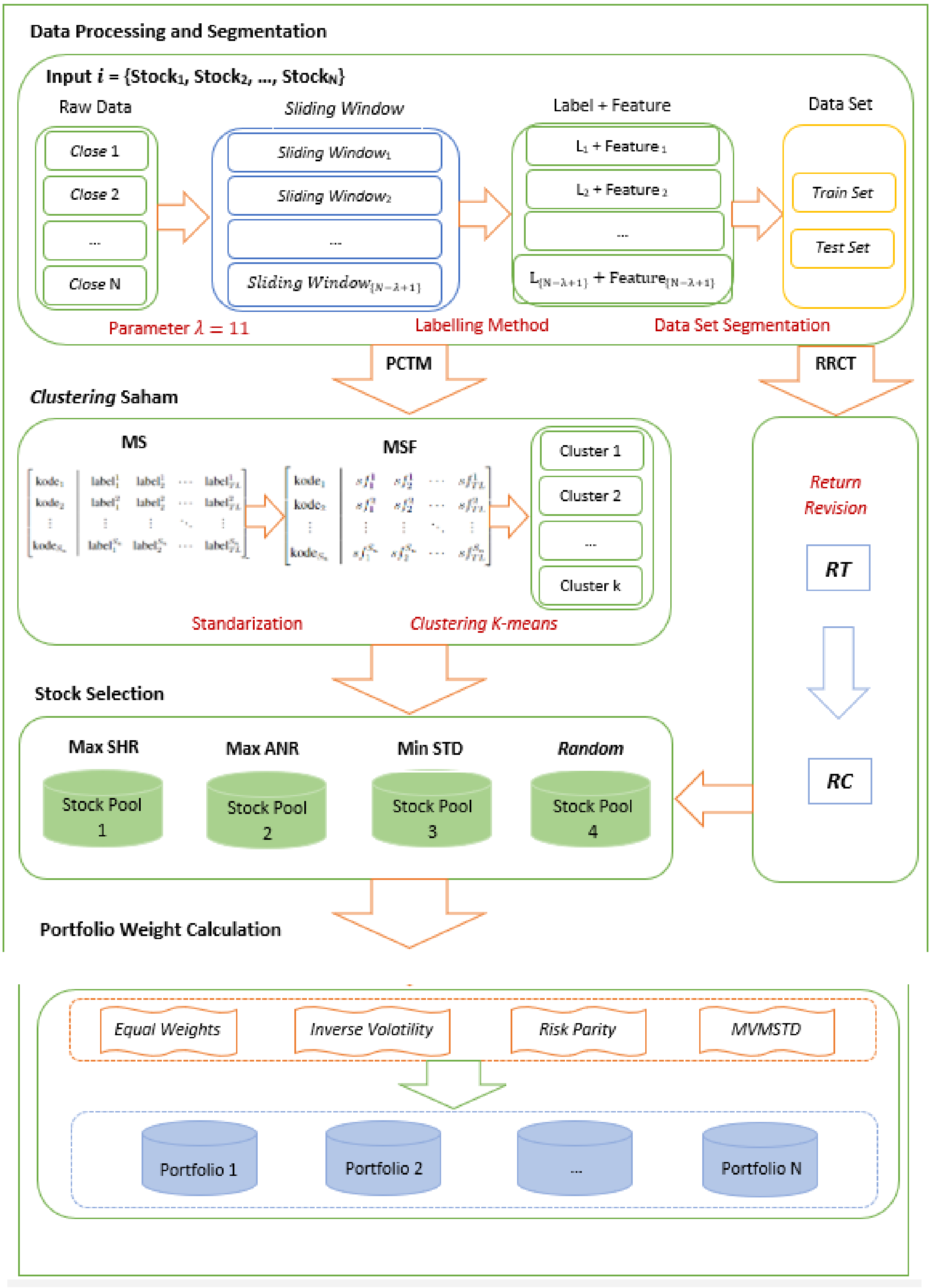

FIGURE 1. Diagram of stock portfolio formation based on feature trend continuous

2.2. **Data Source.** The data used in this study are secondary data in the form of daily transaction data for stocks listed on the LQ45 index in Indonesia at the end of 2022 as many as 45 stocks. Daily stock transaction data was taken starting from January 01, 2008 to

December 31, 2022 from the historical stock data provider site, namely www.finance.yahoo.com. Some of the information obtained includes stock code, opening price, highest price, lowest price, closing price, trading volume, and so on. The following is given the stock code listed on the LQ45 index at the end of 2022.

TABLE 1. List of LQ-45 index shares on December 2022

No.	**Code**	**No.**	**Code**	**No.**	**Code**	**No.**	**Code**
1.	ACES	14.	BRPT	27.	INTP	40.	TINS
2.	ADRO	15.	BUKA	28.	ITMG	41.	TLKM
3.	AKRA	16.	CPIN	29.	JPFA	42.	TOWR
4.	AMRT	17.	EMTK	30.	KLBF	43.	TPIA
5.	ANTM	18.	ESSA	31.	MDKA	44.	UNTR
6.	ARTO	19.	EXCL	32.	MEDC	45.	UNVR
7.	ASII	20.	GOTO	33.	PGAS		
8.	BBCA	21.	HRUM	34.	PTBA		
9.	BBNI	22.	ICBP	35.	SCMA		
10.	BBRI	23.	INCO	36.	SIDO		
11.	BBTN	24.	INDF	37.	SMGR		
12.	BMRI	25.	INDY	38.	SRTG		
13.	BRIS	26.	INKP	39.	TBIG		

2.3. **Data Pre-processing.** Daily stock transaction data listed on the LQ45 index consisting of 45 stocks were collected and data pre-processing was carried out. Data pre-processing aims to transform the data to suit the research. Data will be removed null or missing values at the closing price of the stock. The goal is to get closing prices for all stocks so that there are no null or missing values. The main step is to select a list of stocks in the LQ-45 Index in 2022 that have Initial Public Offering (IPO) before 2008.

2.4. **Labelling Method.** In previous research, a data labeling method based on the continuous trend features of stock data has been presented. Based on the features of continuous stock trends, stock data will be labeled with two classes, namely uptrend and downtrend or no change [32]. This data labeling method is used to mark the trend of stock data and then the data set will be trained and tested to train the model and predict the corresponding stock trend. First, historical closing price data (c) for each stock in the LQ45 index in the Indonesian capital market will be taken as follows:

$$\text{Closing price data } (c) = \begin{bmatrix} c_1 \\ c_2 \\ \dots \\ c_{N-1} \\ c_N \end{bmatrix}$$

where c_i is the closing price of the stock on day i.

Then, the above closing price dimension is expanded by the parameter length λ utilizing the sliding window method so that the data set includes historical price data of length λ. The parameter λ can be determined based on the investor's experience. In this discussion, λ is set to a value of 11 which is consistent with previous research [32], [33]. The steps of dimension expansion with sliding window are given as follows:

$$c = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_{N-1} \\ c_N \end{bmatrix} \rightarrow C = \begin{bmatrix} c_\lambda & c_{\lambda-1} & c_{\lambda-2} & \cdots & c_1 \\ c_{\lambda+1} & c_\lambda & c_{\lambda-1} & \cdots & c_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{N-1} & c_{N-2} & c_{N-3} & \cdots & c_{N-\lambda} \\ c_N & c_{N-1} & c_{N-2} & \cdots & c_{N-\lambda+1} \end{bmatrix}$$

or can be written as

$$\begin{aligned} C &= \begin{bmatrix} c_\lambda & c_{\lambda-1} & c_{\lambda-2} & \cdots & c_1 \\ c_{\lambda+1} & c_\lambda & c_{\lambda-1} & \cdots & c_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{N-1} & c_{N-2} & c_{N-3} & \cdots & c_{N-\lambda} \\ c_N & c_{N-1} & c_{N-2} & \cdots & c_{N-\lambda+1} \end{bmatrix} \\ &= (d_{ij})_{(N-\lambda+1)\times(\lambda)} \\ &= \begin{bmatrix} \mathbf{D}_1 \\ \mathbf{D}_2 \\ \vdots \\ \mathbf{D}_{N-\lambda} \\ \mathbf{D}_{N-\lambda+1} \end{bmatrix} \end{aligned}$$

where C represents the matrix after dimension expansion, D_i denotes the ith row of C matrix, and d_{ij} represents the value of the ith row and jth column of C matrix.

Next, the definitions of continuous uptrend and continuous downtrend are given. The starting and ending points of the sliding window historical data in a certain period of time are denoted by the vectors C_{Aw} and C_{Ak}, where the number of starting and ending points is equal to the number of sliding window method applications of $N - \lambda + 1$ shown in Equation (4) as follows:

$$C_{Aw} = \begin{bmatrix} C_{Aw_1} \\ C_{Aw_2} \\ \vdots \\ C_{Aw_{N-\lambda}} \\ C_{Aw_{N-\lambda+1}} \end{bmatrix}, C_{Ak} = \begin{bmatrix} C_{Ak_1} \\ C_{Ak_2} \\ \vdots \\ C_{Ak_{N-\lambda}} \\ C_{Ak_{N-\lambda+1}} \end{bmatrix}. \tag{3}$$

Then, from the data of the start and end points on the obtained sliding windows can be labeled with the following formula:

$$label_i = \begin{cases} 1, & D_i \in \left\{D_i \mid C_{Aw_k} \leq d_{ij} \leq C_{Ak_k}, C_{Aw_k} < C_{Ak_k}, k = 1, 2, \ldots, C_{Aw_{N-\lambda+1}}\right\} \\ 0, & D_i \in \left\{D_i \mid C_{Aw_k} \leq d_{ij} \leq C_{Ak_k}, C_{Aw_k} \geq C_{Ak_k}, k = 1, 2, \ldots, C_{Ak_{N-\lambda+1}}\right\} \end{cases}. \tag{4}$$

The data labels obtained from Equation (4) above, can be written into a label vector for each stock as follows:

$$\mathbf{LA} = \begin{bmatrix} label_1 \\ label_2 \\ \vdots \\ label_{N-\lambda} \\ label_{N-\lambda+1} \end{bmatrix}.$$

The frequency of investment is different for each investor and the market movement exhibits fluctuation characteristics. Labeling 0 means that in one sliding window method for each stock is decreasing or fixed, while labeling 1 means that in one sliding window method for each stock is increasing which refers to the continuous trend feature.

2.5. **Statistical Calculation Metrics.** Appropriate statistical metrics were selected to measure the quality of the portfolios formed from a risk and return perspective. Return measurement metrics include cumulative return at the end of the period and annualized return converted from the average over the study period. Standard deviation is commonly used to measure portfolio risk. For each stock, the amount of data in the study period is $L + 1$ and the rate of return is formulated as follows:

$$r_t = \frac{p_t - p_{t-1}}{p_t - 1} = \frac{p_t}{p_{t-1}} - 1$$

where r_t represents the daily return.

Then, the yield to maturity and average return can be determined with the following formulation:

$$YTM = \sum_{t=1}^{L} r_t$$

and

$$\bar{r_t} = \frac{\sum_{t=1}^{L} r_t}{L} = \frac{YTM}{L}.$$

In general, there are 252 selling days for one year so the annualized return can be formulated as follows:

$$ANR = 252 \cdot \bar{r_t} = 252 \cdot \frac{\sum_{t=1}^{L} r_t}{L}.$$

In the same way, the level of risk or standard deviation (STD) of an investment can be formulated as follows:

$$STD = \sqrt{\frac{\sum_{t=1}^{L}(r_t - \bar{r_t})^2}{L-1}}.$$

After obtaining ANR and STD, the sharpe ratio value can be formulated with the following formula:

$$SHR = \frac{\bar{r_t} - r_f}{STD}$$

where r_f is the risk-free interest rate issued by Bank Indonesia depending on the month and time of issuance.

The above equations are valid for only one stock. In portfolio formation, information about returns and weights among different stocks must be considered because each stock has different characteristics. The following table shows the calculation metrics for a single stock and a portfolio.

Table 2. Calculation metrics for a single stock and portfolio

Metrics	**Single Stock**	**Portofolio**
Return	$r_t = \frac{p_{t+1}}{p_t} - 1$	$R_t = (r_t^1, r_t^2, \ldots, r_t^N)$
YTM	$\sum_{t=1}^{L} r_t$	$w^T \cdot \sum_{t=1}^{L} R_t$
Mean return	$\bar{r_t} = \frac{\sum_{t=1}^{L} r_t}{L}$	$\bar{r}_t^p = \frac{\sum_{t=1}^{L} R_t}{L}$
ANR	$252 \cdot \bar{r_t}$	$252 \cdot w^T \cdot \bar{r}_t^p$
STD	$\sqrt{\frac{\sum_{t=1}^{L}(r_t - \bar{r_t})^2}{L-1}}$	$\sqrt{w^T \cdot cov(R) \cdot w}, R = (R_1, R_2, \ldots, R_L)^T$
SHR	$\frac{\bar{r_t} - r_f}{STD}$	$\frac{\bar{r_t} - r_f}{STD}$

2.6. **Stock Portfolio Formation.**

2.6.1. *Portfolio Mean-Variance.* Determining the weight or proposal of each asset in a portfolio is generally done using a mathematical approach. This mathematical approach was first proposed by Markowitz known as mean-variance theory. Mean-variance portfolio is defined as a portfolio that has a minimum variance among all possible assets that can be selected at the same expected return. Before forming a portfolio, the following definitions are given about the return covariance matrix and the portfolio return variance matrix. The return covariance matrix is a matrix that has entries of the covariance of returns between all assets selected to form a portfolio. According to [21], the return covariance matrix is formulated as follows:

$$\Sigma = cov(R_p) = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} & \cdots & \sigma_{1,n} \\ \sigma_{2,1} & \sigma_{2,2} & \cdots & \sigma_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n,1} & \sigma_{n,2} & \cdots & \sigma_{n,n} \end{bmatrix}$$

where $\sigma_{i,j}$ is covariance between asset i and j. The metric return portfolio is a variation of all the aset returns that are selected to build a portfolio. According to [21], the return portfolio's matrix variants are transformed as follows:

$$\begin{aligned} Var(R_p) &= Var(w_1R_{1,t} + w_2R_{2,t} + \cdots + w_nR_{n,t}) \\ &= w^T\Sigma w \end{aligned}$$

with

$$r = \begin{bmatrix} R_{1,t} \\ R_{2,t} \\ \vdots \\ R_{n,t} \end{bmatrix} \text{dan} w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}.$$

According to [21], optimizing the mean-variance portfolio means minimizing variation while maintaining the bobot $\mathbf{w}= (w_1, w_2, \dots, w_n)^T$. The model Mean-Variance portfolio is provided below as follows:

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\Sigma\mathbf{w} \tag{5}$$

with constraints

$$\mathbf{w}^T\mathbf{1}_n = 1 \tag{6}$$

where

$$\mathbf{1}_n = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{n\times 1}.$$

By solving the optimization constraints in Equation (5) and Equation (6), we will obtain

$$\mathbf{w} = \frac{\Sigma^{-1}\mathbf{1}_n}{\mathbf{1}_n^T\Sigma^{-1}\mathbf{1}_n}. \tag{7}$$

Next, consider the second derivative of L with respect to $\mathbf{w}$,

$$\begin{aligned} \frac{\partial^2 L}{\partial \mathbf{w}^T \partial \mathbf{w}} &= \frac{\partial}{\partial \mathbf{w}^T}\left(\Sigma\mathbf{w} - \lambda\mathbf{1}_n\right) \\ &= \Sigma > 0. \end{aligned}$$

Thus, Equation (7) is the global minimum of L. Consequently, Equation (7) is the solution to the optimization problem in Equation (5).

2.6.2. *Portfolio Inverse Volatility.* The Naïve Risk Parity technique, also known as the inverse volatility portfolio, is a method of allocating weights to a portfolio that is proportionate to the inverse of the volatility, as determined by standard deviation. Inversely weighing each asset according to its volatility will not yield useful weights. As a result, all securities were scaled to one in order to achieve consistency. The primary drawback of the inverse volatility portfolio is that the correlation coefficient's benefits of diversification are eliminated since it ignores the variance-covariance matrix. Therefore, pairwise correlations between assets will be ignored for assets with a higher standard deviation. Then, using the homogeneous correlation assumption, the optimal portfolio weight vector is computed using the same formula as the ERC technique.

$$w_i = \frac{\sigma_i^{-1}}{\sum_{i=1}^{N} \sigma_i^{-1}}$$

where w_i is the weight of each asset i and σ_i is the volatility of each asset i for $i = 1, 2, \ldots, n$.

2.6.3. *Portfolio Equally Weighted.* To create a robust and well-balanced portfolio, one approach is to use the Equally Weighted portfolio strategy. The investor using this method must distribute their funds equally among a variety of assets. The goal is to provide each asset that is taken into consideration for the portfolio the same amount of weight. The naïve portfolio or $1/N$ portfolio is another name for this [19].

$$w_i = \frac{1}{N}, \forall i = 1, 2, \ldots, N.$$

One drawback of this method is that Equally Weighted portfolios do not include active distributions on certain assets. Furthermore, as the majority of the portfolio's overall risk would come from a small number of highly volatile assets, this strategy would also suggest poor diversification [19]. Additionally, if all assets have the same mean, variance, and correlation, the equally weighted portfolio is an ideal mean-variance investment. To be more precise, an Equally Weighted portfolio is a special portfolio on the efficient frontier that also happens to be the least volatile portfolio [15].

2.6.4. *Portfolio Risk Parity.* By allocating the same risk weights to each asset, the Risk Parity (RP) or Equal Risk Contribution (ERC) portfolio model is a specific type of Risk Budgeting (RB) portfolio that seeks to prevent changes in the volatility of the entire portfolio when the volatility of any one asset changes [23]. This entails creating a portfolio in which each asset has a certain amount of risk. The return from the portfolio can be defined as follows if there are N assets, the return on each asset is r_i, and the weight is w_i for $i = 1, 2, \ldots, n$.

$$R = \sum_{i=1}^{N} r_i w_i$$

and risk from portfolio ($\sigma(w)$) is

$$\sigma(w) = \sqrt{w^T \Sigma w} = \sqrt{\sum_{i=1}^{N} w_i^2 \sigma_i^2 + \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} 2 w_i w_j \sigma_{i,j}}$$

Determining the marginal risk contribution parity and the specific contributions of each asset to the overall risk of the portfolio is a crucial first step in building a risk parity portfolio. The partial derivative of the overall portfolio risk to the weight of asset i is known as the Marginal Risk Contribution (MRC) of asset i. This definition can be expressed as follows:

$$MRC_i = \frac{\Sigma w}{\sqrt{w^T \Sigma w}}$$

The definition of the Total Risk Contribution (TRC) of asset i is the product of each asset's weight and the Marginal Risk Contribution (MRC), which can be expressed as follows:

$$TRC_i = w_i \cdot MRC_i = w_i \cdot \frac{\Sigma w}{\sqrt{w^T \Sigma w}}$$

The total risk of a portfolio can be calculated using the definitions of total risk contribution (TRC) and marginal risk contribution (MRC) as a basis:

$$\partial(w) = \sum_{i=1}^{N} TRC_i = \sum_{i=1}^{N} w_i \frac{\partial\sigma(w)}{\partial w_i} = \frac{w^T \Sigma w}{\sqrt{w^T \Sigma w}} = \sqrt{w^T \Sigma w}.$$

The goal of risk parity is to allocate an equal risk budget to each portfolio asset. Therefore, the marginal contribution of each asset to risk is the same. In other words, a risk parity portfolio must meet the following characteristics:

$$TRC_i = TRC_j$$

with $i = 1, 2, \ldots, N$, $j = 1, 2, \ldots, N$.

The goal of the risk parity portfolio approach is to diversify the capital risk across all securities and determine the best allocation. The portfolio problem with $N-$assets can be phrased as follows by preventing short sells and setting a limit of one on the overall weight of each share ([5]):

$$TRC_i = TRC_j, \forall i, j$$

with the constraint $\sum_{i=1}^{N} w_i = 1, 0 \leq w_i \leq 1$ where $i = 1, 2, \ldots, N, j = 1, 2, \ldots, N$. The problem above can be solved using the Newton Raphson method to get the weight of each share.

2.7. **Formation of Stock Portfolio Based on Continuous Trend Features.** The method for forming a stock portfolio discussed consists of two parts. The first method is to improve the calculation of returns based on continuous trend features according to the RRCT algorithm which significantly improves relevant evaluation metrics as well as calculation metrics. In the second method, shares are grouped by clustering based on sustainable trend features and portfolio formation is carried out from the clustering data. Below are the Algorithm 1 and Algorithm 2 which explain the two methods.

Algorithm 1 RRCT, Algorithm for Calculating Revised Returns Based on Continuous Trend Features

```
Require: Raw stock data $\mathbf{X} = \{x_1, x_2, \ldots, x_N\}^T$ where $x$ is a data vector. Vector labels $\mathbf{Y}$:
  $\mathbf{Y} = \{label_1, label_2, \ldots, label_N\}$ where $label_i \in \{0.1\}$.
  Output: Stock return matrix, $\mathbf{RT} = [RT_1, RT_2, \ldots, RT_L]^T$. Stock return revision matrix
  based on continuous feature trends, $\mathbf{RC} = [RC_1, RC_2, \ldots, RC_L]^T$.
  Initialization: $Longhold$=0, $\mathbf{RT}$=[], $\mathbf{RC}$=[].
  Run this process for each stock $S_j$:
  temp_vec_rt=[]
  temp_vec_rc=[]
  for $L[i] \in \boldsymbol{Y}:$ do
    $return \leftarrow$ (closing price[i]-closing price[i-1])/closing price[i-1]
    temp_vec_rt[i] $\leftarrow$ $return$
    if $Longhold > 0$: then
      $return \leftarrow$ (closing price[i]-closing price[i-1])/closing price[i-1]
    end
    if $Longhold = 0$: then
      $return \leftarrow 0$
    end
    if $Longhold = 0 and L[i] > 0$: then
      $longhold \leftarrow 1$
      $return \leftarrow 0$
    end
    if $Longhold = 0 and (L[i] > 0 or L[i] last\ data)$: then
      $longhold \leftarrow 0$
      $return \leftarrow$ (closing price[i]-closing price[i-1])/closing price[i-1]
    end
    $temp\_vec\_rc[i] \leftarrow return$
    $\mathbf{RT}[j] \leftarrow$ temp_vec_rt
    $\mathbf{RC}[j] \leftarrow$ temp_vec_rc
  end
```

Based on the Algorithm 1 above, the algorithm provides an overview of the steps for calculating return revisions based on continuous trend features.

Algorithm 2 PCTM, Stock Portfolio Formation Algorithm Based on Continuous Trend Feature Clustering.

```
Require: Prediction label vector $\mathbf{Y}$: $\mathbf{Y} = \{label_1, label_2, \ldots, label_N\}$ where $label_i \in \{0.1\}$.
  The return matrix of shares in the traditional way, $\mathbf{RT} = [RT_1, RT_2, \ldots, RT_N]^T$. Revision
  matrix return based on continuous trend features, $\mathbf{RC} = [RC_1, RC_2, \ldots, RC_L]^T$. $K$
  parameters of the k-means clustering algorithm.
  MS ← Clustering matrix
  MSF ← Standardized clustering matrix
  Cluster $K$ ← algorithm k-means with parameters $k$
  Stock Pool 1 ← Select one stock with the highest SHR in each cluster
  Stock Pool 2 ← Select one stock with the highest ANR in each cluster
  Stock Pool 3 ← Select one stock with the lowest STD in each cluster
  Stock Pool 4 ← Randomly select one stock in each cluster
  for each stock pool do
      if each stock pool = Stock Pool 4 then
          Portfolios are formed with equal weights with RC and RT.
          The portfolio is formed with inverse volatility with RC and RT.
          The portfolio is formed with risk parity with RC and RT.
          The portfolio is formed with the lowest Markowitz STD with RC and RT.
      end
      else
          Portfolios are formed with equal weights with RC.
          The portfolio is formed with inverse volatility with RC.
          The portfolio is formed with risk parity with RC.
          The portfolio is formed with the lowest Markowitz STD with RC.
      end
      Metrics for each portfolio are formed.
  end
```

Next, the Algorithm 2 provides an overview of the steps for forming a stock portfolio based on sustainable trend features. Several methods for forming stock portfolios include equal weights, inverse volatility, risk parity, and Markowitz.

In the second method, the stocks are grouped based on continuous trend features. The shares consisting of 27 shares are not grouped based on correlation between shares or share price information, but the shares are grouped based on continuous trend features, namely label information as shown in the following Equation (8):

$$MS = [V_1, V_2, \ldots, V_{S_n}]^T = \left[\begin{array}{c|cccc} \text{kode}_1 & \text{label}_1^1 & \text{label}_2^1 & \cdots & \text{label}_{TL}^1 \\ \text{kode}_2 & \text{label}_1^2 & \text{label}_2^2 & \cdots & \text{label}_{TL}^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \text{kode}_{S_n} & \text{label}_1^{S_n} & \text{label}_2^{S_n} & \cdots & \text{label}_{TL}^{S_n} \end{array}\right] \tag{8}$$

So that the data grouping process is more effective, the MS data above is standardized with the standardization process presented in Equation (9) and Equation (10) as follows:

$$sf_i^j = \frac{x - \mu}{\sigma} \tag{9}$$

with

$$x = \text{label}_i^j, \mu = \frac{\sum_i^{S_n} \text{label}_i^j}{S_n}, \sigma = \sqrt{\frac{\sum_{j=1}^{S_n} \left(\text{label}_i^j - \mu\right)^2}{S_n}}. \tag{10}$$

Furthermore, the data features obtained after carrying out the standardization process are as follows:

$$MSF = [VF_1, VF_2, \ldots, VF_{S_n}]^T = \left[\begin{array}{c|cccc} \text{kode}_1 & sf_1^1 & sf_2^1 & \cdots & sf_{TL}^1 \\ \text{kode}_2 & sf_1^2 & sf_2^2 & \cdots & sf_{TL}^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \text{kode}_{S_n} & sf_1^{S_n} & sf_2^{S_n} & \cdots & sf_{TL}^{S_n} \end{array}\right].$$

The k-means clustering technique is used to arrange the MSF matrix above. The selection of parameter k seeks to take into account the performance of various stocks as well as the impact of parameter k on portfolio construction. In order to compare the parameter k based on the number of clusters created by the k-means method, which makes use of the t-SNE method, an experiment will be conducted. The parameter value k derived from the clustering technique meets the requirements of some investors as well as the study's experiments. Following k-means clustering, k distinct groupings of stocks will be found. To create a portfolio, four approaches are taken into consideration, as Table 3. The weight calculation method used in forming the portfolio is shown in Table 4. The calculation results of ANR, STD, and SHR from each portfolio will be compared and analyzed for each portfolio formation method.

Table 3. Portfolio construction method

Experiment	Explanation
Minimum STD	Selecting stocks with minimum standard deviation in each cluster with k shares
Maximum SHR	Selecting shares with maximum sharpe ratio in each cluster with k shares
Maximum ANR	Selecting shares with maximum annual return in each cluster with k shares
Random	Randomly selects k shares that have done previous preprocessing data

Table 4. Portfolio weight calculation method

Portfolio Method	Explanation
Equal weights	Each share has the same weight
Inverse volatility	Calculates the weight of shares based on the inverse volatility algorithm
Risk parity	Calculating the weight of shares based on the risk parity algorithm
MVMSTD	Calculating weights based on Markowitz theory with minimum standard deviation

3. Results and Discussion

In this section, the results of preliminary data, RC, and RT are compared and analyzed. The results of each metric are analyzed and the conclusion is accordingly given.

3.1. **Preliminary Data Results.** In the previous chapter, data preprocessing was carried out from the initial data of stocks listed in the LQ-45 Index so that 27 stocks were obtained. The following are the initial data results which include average return, standard deviation, annualized return, and sharpe ratio. In Table 5, it can be noted that EXCL.JK stocks have the highest mean return and annualized return values with an average return of 0.002707 or 0.2% and an annualized return of 0.682062635 or 68.20%. Then, for stocks that have the lowest risk or standard deviation, BBCA.JK stock has a value of 0.019028361 or 1.90%. Furthermore, in Table 5 which has the highest sharpe ratio value, CPIN.JK is 0.036183 or 3.61%.

Table 5. Summary data for initial data

Code	Mean Return	STD	ANR	SHR
ACES.JK	0.00079	0.025053031	0.199053448	0.021594
AKRA.JK	0.000785	0.026266357	0.197786148	0.020405
ANTM.JK	0.000322	0.031676378	0.081253182	0.002322
ASII.JK	0.000497	0.024209975	0.125289127	0.010255
BBCA.JK	0.000846	0.019028361	0.213250472	0.031392
BBNI.JK	0.000738	0.024669553	0.186095936	0.019846
BBRI.JK	0.000814	0.024585093	0.205037088	0.022971
BMRI.JK	0.000755	0.023867253	0.190358137	0.021221
BRPT.JK	0.000822	0.033867979	0.207243066	0.016933
CPIN.JK	0.001338	0.03009243	0.337108488	0.036183
EXCL.JK	0.002707	0.155788423	0.682062635	0.015776
INCO.JK	0.000452	0.032911186	0.113851818	0.006165
INDF.JK	0.000509	0.022289577	0.128327216	0.01168
INKP.JK	0.001235	0.03531389	0.311234334	0.027926
INTP.JK	0.000406	0.02644872	0.102186056	0.005921
ITMG.JK	0.000668	0.030514235	0.168298461	0.01373
JPFA.JK	0.00105	0.031096732	0.264479295	0.025746
KLBF.JK	0.000845	0.023388309	0.212818099	0.025467
MEDC.JK	0.00055	0.033022319	0.138529431	0.00911
PGAS.JK	0.000265	0.028801853	0.066831208	0.000566
PTBA.JK	0.000544	0.029231432	0.136993675	0.010083
SCMA.JK	0.000905	0.03006133	0.228016399	0.02182
SMGR.JK	0.000358	0.02529322	0.090289663	0.004325
TINS.JK	0.000381	0.03226096	0.095973787	0.00409
TLKM.JK	0.000363	0.019702101	0.091446361	0.005786
UNTR.JK	0.000643	0.027913594	0.16205816	0.014122
UNVR.JK	0.000538	0.020274407	0.135597573	0.014264

3.2. **Analysis of RT and RC.** After the labeling information is harvested based on the continuous trend features through the labeling method, the calculation of the return rate is revised according to Algorithm 1, and the results of RT and RC are obtained. Then, summary data for RC data will be provided. From the data in Table 6 and Table 7, the difference in values between data RT and data RC is given.

Table 6. Summary data for data RT

Code	Mean Return	STD	ANR	SHR
ACES.JK	0.000755	0.025128	0.190305	0.020148
AKRA.JK	0.000773	0.026197	0.194693	0.019990
ANTM.JK	0.000358	0.031632	0.090325	0.003463
ASII.JK	0.000490	0.024177	0.123581	0.009989
BBCA.JK	0.000862	0.019034	0.217107	0.032186
BBNI.JK	0.000768	0.024670	0.193421	0.021023
BBRI.JK	0.000822	0.024591	0.207061	0.023292
BMRI.JK	0.000794	0.023853	0.199975	0.022834
BRPT.JK	0.000888	0.033806	0.223710	0.018898
CPIN.JK	0.001325	0.030089	0.333882	0.035762
EXCL.JK	0.002777	0.156338	0.699921	0.016174
INCO.JK	0.000440	0.032908	0.110812	0.005799
INDF.JK	0.000417	0.022015	0.105038	0.007627

Code	Mean Return	STD	ANR	SHR
INKP.JK	0.001055	0.034738	0.265824	0.023201
INTP.JK	0.000437	0.026419	0.110215	0.007134
ITMG.JK	0.000586	0.030446	0.147663	0.011071
JPFA.JK	0.001034	0.031085	0.260603	0.025261
KLBF.JK	0.000859	0.023380	0.216539	0.026108
MEDC.JK	0.000586	0.032993	0.147662	0.010216
PGAS.JK	0.000238	0.028587	0.060032	-0.000373
PTBA.JK	0.000562	0.029245	0.141566	0.010698
SCMA.JK	0.000836	0.029736	0.210619	0.019737
SMGR.JK	0.000392	0.025299	0.098761	0.005653
TINS.JK	0.000333	0.032165	0.083880	0.002610
TLKM.JK	0.000382	0.019686	0.096168	0.006742
UNTR.JK	0.000617	0.027866	0.155505	0.013213
UNVR.JK	0.000544	0.020226	0.137095	0.014592

Table 7. Summary data for data RC

Code	Mean Return	STD	ANR	SHR
ACES.JK	0.000762	0.025131	0.192046	0.020420
AKRA.JK	0.000791	0.025987	0.199351	0.020864
ANTM.JK	0.000375	0.031435	0.094495	0.004011
ASII.JK	0.000511	0.024149	0.128687	0.010840
BBCA.JK	0.000874	0.019023	0.220131	0.032836
BBNI.JK	0.000777	0.024667	0.195811	0.021411
BBRI.JK	0.000835	0.024581	0.210480	0.023854
BMRI.JK	0.000812	0.023831	0.204500	0.023608
BRPT.JK	0.000900	0.033802	0.226835	0.019267
CPIN.JK	0.001333	0.030089	0.335885	0.036026
EXCL.JK	0.002803	0.156417	0.706287	0.016327
INCO.JK	0.000464	0.032879	0.117010	0.006552
INDF.JK	0.000426	0.022011	0.107341	0.008044
INKP.JK	0.001061	0.034741	0.267304	0.023368
INTP.JK	0.000456	0.026388	0.115033	0.007867
ITMG.JK	0.000616	0.030426	0.155195	0.012060
JPFA.JK	0.001064	0.031036	0.268155	0.026267
KLBF.JK	0.000855	0.023361	0.215562	0.025962
MEDC.JK	0.000608	0.032971	0.153126	0.010881
PGAS.JK	0.000255	0.028573	0.064298	0.000219
PTBA.JK	0.000581	0.029225	0.146521	0.011379
SCMA.JK	0.000886	0.029633	0.223274	0.021500
SMGR.JK	0.000363	0.025162	0.091482	0.004536
TINS.JK	0.000350	0.032017	0.088233	0.003162
TLKM.JK	0.000391	0.019645	0.098650	0.007257
UNTR.JK	0.000642	0.027849	0.161722	0.014107
UNVR.JK	0.000548	0.020227	0.138158	0.014799

Table 8 shows the difference in the results of the evaluation calculation metrics for RT data and RC data. It can be seen that the mean return, annualized return, and sharpe ratio show relatively negative results, meaning that the value of RC data significantly exceeds the value of RT data. Meanwhile, the standard deviation value shows a relatively positive result, meaning that the standard deviation value of RC data is significantly smaller than the standard deviation value of RT data.

Table 8. Differences between data RT and data RC

Code	Mean Return	STD	ANR	SHR
ACES.JK	-0.000007	-3,575657e-06	-0.001740	-0.000272
AKRA.JK	-0.000018	2,108474e-04	-0.004658	-0.000874
ANTM.JK	-0.000017	1,968089e-04	-0.004170	-0.000548
ASII.JK	-0.000020	2,811727e-05	-0.005105	-0.000851
BBCA.JK	-0.000012	1,140834e-05	-0.003023	-0.000650
BBNI.JK	-0.000009	3,403448e-06	-0.002390	-0.000387
BBRI.JK	-0.000014	1,050546e-05	-0.003419	-0.000562
BMRI.JK	-0.000018	2,177636e-05	-0.004525	-0.000774
BRPT.JK	-0.000012	3,836088e-06	-0.003125	-0.000369
CPIN.JK	-0.000008	-1,932218e-07	-0.002003	-0.000264
EXCL.JK	-0.000025	-7,891368e-05	-0.006367	-0.000153
INCO.JK	-0.000025	2,953594e-05	-0.006199	-0.000753
INDF.JK	-0.000009	4,028937e-06	-0.002304	-0.000417
INKP.JK	-0.000006	-2,873307e-06	-0.001479	-0.000167
INTP.JK	-0.000019	3,031286e-05	-0.004818	-0.000733
ITMG.JK	-0.000030	1,991448e-05	-0.007532	-0.000990
JPFA.JK	-0.000030	4,918563e-05	-0.007553	-0.001006
KLBF.JK	0.000004	1,859774e-05	0.000977	0.000145
MEDC.JK	-0.000022	2,188510e-05	-0.005464	-0.000664
PGAS.JK	-0.000017	1,464342e-05	-0.004266	-0.000592
PTBA.JK	-0.000020	2,047819e-05	-0.004955	-0.000680
SCMA.JK	-0.000050	1,026815e-04	-0.012655	-0.001763
SMGR.JK	0.000029	1,371165e-04	0.007279	0.001117
TINS.JK	-0.000017	1,477687e-04	-0.004353	-0.000552
TLKM.JK	-0.000010	4,101677e-05	-0.002482	-0.000515
UNTR.JK	-0.000025	1,641618e-05	-0.006216	-0.000894
UNVR.JK	-0.000004	-1,112150e-06	-0.001062	-0.000208

3.3. **Visualization of The Results of Cluster Analysis.** In this part, the t-SNE algorithm is used to visualize the results of the cluster analysis. For different values of parameter k, a histogram is used to show the number of each category after clustering in Figure 2, which can be more intuitive to detect the difference in the number of each category, indicating that there is a distance-based difference between each stock.

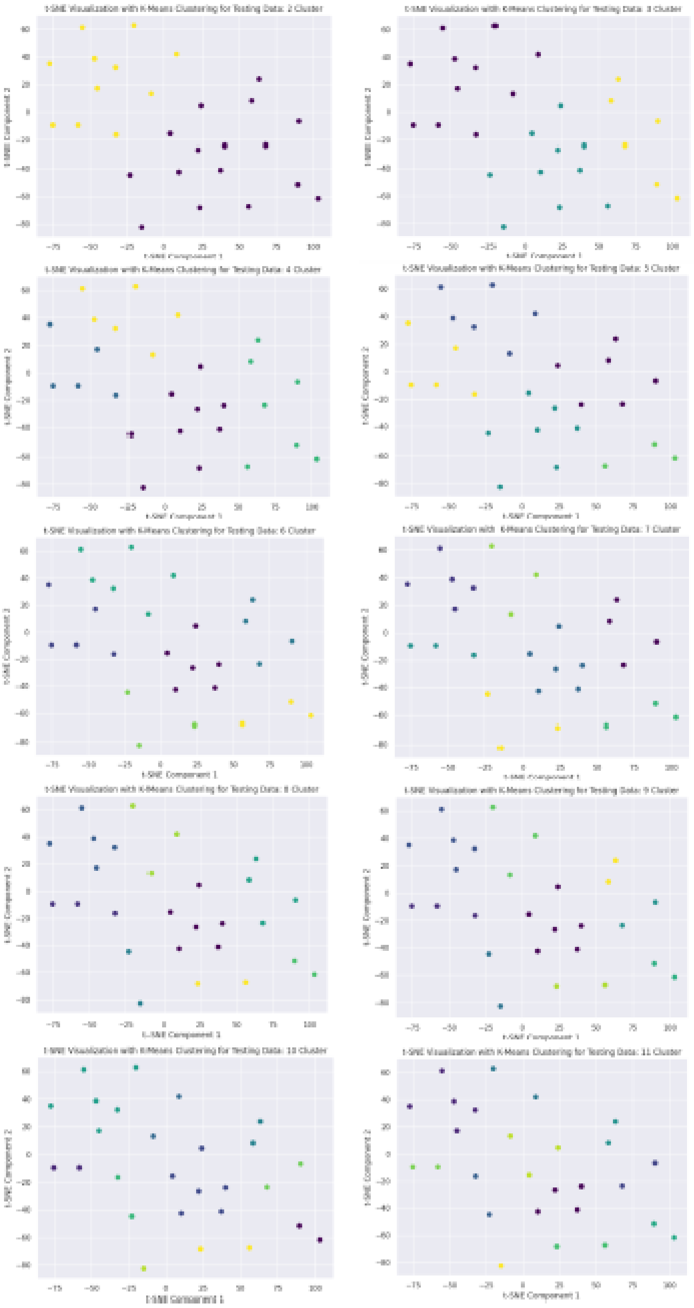

Figure 2. Visualization based on the t-SNE for the clustering results

3.4. **Portfolio Construction.** In this section, we will provide the results of the investment portfolio based on the Algorithm 2, from which further results will be concluded. An illustration of the results of portfolio return calculations for $k = 2$ to $k = 11$ will be given.

In Table 9, it illustrates that the more clusters formed, the smaller the portfolio return will be. On the other hand, for portfolio formation based on random stock selection will give unfavorable results because there are differences in portfolio return values between each cluster formed and experience fluctuating patterns. For portfolio formation based on SHR max, the portfolio return value at the beginning of the smallest number of clusters provides a fairly

Table 9. Portfolio return value from portfolio construction

Cluster	Metrics	Equally Weights	Inverse Volatility	Risk Parity	Mean Variance
2	SHR max	0.0011968	0.0012065	0.0012065	0.00122265
	ANR max	0.0020610	0.0015679	0.0015679	0.00137731
	SD min	0.0007576	0.0007794	0.0007794	0.0008025
	Random	0.0003878	0.0003884	0.0003884	0.00038904
3	SHR max	0.0010890	0.0010537	0.0010537	0.00096893
	ANR max	0.0016693	0.0012543	0.0012543	0.00113408
	SD min	0.0006878	0.0006965	0.0006965	0.00070113
	Random	0.0007967	0.0007485	0.0007485	0.00065359
4	SHR max	0.0010418	0.0010222	0.0010222	0.00095396
	ANR max	0.0015172	0.0011997	0.0011997	0.00110081
	SD min	0.0005796	0.0006072	0.0006072	0.00065901
	Random	0.0006971	0.0007058	0.0007058	0.0007639
5	SHR max	0.0010071	0.0009871	0.0009871	0.0009174
	ANR max	0.0013875	0.0011020	0.0011020	0.0009955
	SD min	0.0006161	0.0006362	0.0006362	0.00068173
	Random	0.0008196	0.0008063	0.0008063	0.00075403
6	SHR max	0.0009869	0.0009716	0.0009716	0.00091042
	ANR max	0.0013018	0.0010412	0.0010412	0.00092677
	SD min	0.0005785	0.0005888	0.0005888	0.00062671
	Random	0.0010750	0.0007861	0.0007861	0.00072974
7	SHR max	0.0009376	0.0009254	0.0009254	0.00086067
	ANR max	0.0012038	0.0009805	0.0009805	0.00087733
	SD min	0.0005509	0.0005669	0.0005669	0.00059463
	Random	0.0006735	0.0006842	0.0006842	0.00068117
8	SHR max	0.0009157	0.0009035	0.0009035	0.000845
	ANR max	0.0011486	0.0009483	0.0009483	0.00085477
	SD min	0.0006514	0.0006463	0.0006463	0.00060441
	Random	0.0006851	0.0006574	0.0006574	0.00060155
9	SHR max	0.0008621	0.0008503	0.0008503	0.00082134
	ANR max	0.0010691	0.0008851	0.0008851	0.00082694
	SD min	0.0006235	0.0006198	0.0006198	0.00059543
	Random	0.0009523	0.0007702	0.0007702	0.00073237
10	SHR max	0.0010406	0.0008739	0.0008739	0.00082075
	ANR max	0.0010406	0.0008739	0.0008739	0.00082075
	SD min	0.0006437	0.0006370	0.0006370	0.000627
	Random	0.0009326	0.0007548	0.0007548	0.00067925
11	SHR max	0.0008130	0.0007972	0.0007972	0.00077251
	ANR max	0.0009848	0.0008220	0.0008220	0.00077604
	SD min	0.0006825	0.0006634	0.0006634	0.0006203
	Random	0.0009052	0.0007503	0.0007503	0.00071078

high value which then experiences a relatively small decline. However, when a certain k value will give a return increase to a relatively high portfolio return value. Meanwhile, based on ANR max and SD min, the portfolio return value decreases as the number of clusters formed increases.

Table 10 illustrates that the more clusters formed, the smaller the portfolio risk will be. On the other hand, for portfolio formation based on random stock selection will provide unfavorable results because there are differences in the risk value of the portfolio between each cluster formed and experiencing a fluctuating pattern. For portfolio formation based on SHR max, ANR max, SD min, and Random, the portfolio risk value at the beginning of the smallest

Table 10. Portfolio risk value from portfolio construction

Cluster	Metrics	Equally Weights	Inverse Volatility	Risk Parity	Mean Variance
2	SHR max	0.0256750	0.0255035	0.0255035	0.02539458
	ANR max	0.0800529	0.0362132	0.0362132	0.02969833
	SD min	0.0173638	0.0164932	0.0164932	0.01617013
	Random	0.0184426	0.0170123	0.0170123	0.01658959
3	SHR max	0.0202358	0.0187750	0.0187750	0.01755652
	ANR max	0.0540925	0.0237404	0.0237404	0.0210238
	SD min	0.0144775	0.0141875	0.0141875	0.01417148
	Random	0.0205390	0.0197978	0.0197978	0.01922215
4	SHR max	0.0198962	0.0185842	0.0185842	0.01715464
	ANR max	0.0420430	0.0208840	0.0208840	0.01913939
	SD min	0.0147131	0.0143236	0.0143236	0.01397759
	Random	0.0184252	0.0176346	0.0176346	0.01686778
5	SHR max	0.0181693	0.0171150	0.0171150	0.01599089
	ANR max	0.0344844	0.0178331	0.0178331	0.01652987
	SD min	0.0132809	0.0129750	0.0129750	0.01270696
	Random	0.0157322	0.0156416	0.0156416	0.01527354
6	SHR max	0.0159253	0.0151994	0.0151994	0.01410706
	ANR max	0.0295300	0.0156279	0.0156279	0.01425571
	SD min	0.0127032	0.0125162	0.0125162	0.01221137
	Random	0.0300211	0.0160168	0.0160168	0.01496661
7	SHR max	0.0143703	0.0138023	0.0138023	0.01281417
	ANR max	0.0257939	0.0143069	0.0143069	0.01319508
	SD min	0.0117945	0.0116731	0.0116731	0.01138349
	Random	0.0144856	0.0145402	0.0145402	0.0140586
8	SHR max	0.0132007	0.0126723	0.0126723	0.01181101
	ANR max	0.0229145	0.0129699	0.0129699	0.01198362
	SD min	0.0127847	0.0126356	0.0126356	0.01190178
	Random	0.0130980	0.0129911	0.0129911	0.01210378
9	SHR max	0.0132173	0.0128007	0.0128007	0.01176439
	ANR max	0.0212265	0.0130210	0.0130210	0.01192205
	SD min	0.0129462	0.0128404	0.0128404	0.01188124
	Random	0.0207370	0.0124098	0.0124098	0.011688
10	SHR max	0.0197885	0.0128077	0.01280771	0.01168689
	ANR max	0.0197885	0.0128077	0.01280771	0.01168689
	SD min	0.0130728	0.0129657	0.0129657	0.01171814
	Random	0.0194019	0.0123767	0.0123767	0.01146142
11	SHR max	0.0130493	0.0127849	0.01278496	0.01151635
	ANR max	0.0187133	0.0127700	0.0127700	0.01153247
	SD min	0.0129559	0.0128384	0.0128384	0.01160022
	Random	0.0185777	0.0127633	0.0127633	0.01153527

number of clusters gives a fairly high value which then experiences a relatively small decrease. However, when a certain k value will give a return to a relatively high portfolio risk value. In addition, portfolio formation based on SD min will also provide the smallest portfolio risk results compared to others.

Table 11 illustrates that the more clusters formed, the smaller the annualized return of the portfolio, especially those formed based on SHR max, ANR max, and SD min. On the other hand, for portfolio formation based on random stock selection, the results are not good because there are differences in the annualized return value of the portfolio between each cluster formed and experiencing a fluctuating pattern. For portfolio formation based on SHR

Table 11. Portfolio annualized return value from portfolio construction

Cluster	Metrics	Equally Weights	Inverse Volatility	Risk Parity	Mean Variance
2	SHR max	0.3015942	0.3040547	0.3040547	0.3081077
	ANR max	0,5193957	0.3951146	0.3951146	0.3470828
	SD min	0.1909261	0.1964255	0.1964255	0.2022298
	Random	0.0977469	0.0979013	0.0979013	0.0980370
3	SHR max	0.2744397	0.2655537	0.2655537	0.2441714
	ANR max	0,4206885	0.3160850	0.3160850	0.2857869
	SD min	0.1733366	0.1755381	0.1755381	0.1766845
	Random	0.2007772	0.1886263	0.1886263	0.1647038
4	SHR max	0.2625385	0.2576086	0.2576086	0.2403980
	ANR max	0.3823423	0.3023405	0.3023405	0.2774040
	SD min	0.1460769	0.1530226	0.1530226	0.1660695
	Random	0.1756909	0.1778788	0.1778788	0.1925033
5	SHR max	0.2538126	0.2487532	0.2487532	0.2311858
	ANR max	0.3496556	0.2777157	0.2777157	0.2508659
	SD min	0.1552707	0.1603228	0.1603228	0.1717957
	Random	0.2065630	0.2032113	0.2032113	0.1900153
6	SHR max	0.2487228	0.2448573	0.2448573	0.2294245
	ANR max	0.3280681	0.2623890	0.2623890	0.2335469
	SD min	0.1457953	0.1483866	0.1483866	0.1579315
	Random	0.2709217	0.1981136	0.1981136	0.1838947
7	SHR max	0.2362941	0.2332238	0.2332238	0.2168877
	ANR max	0.3033719	0.2470962	0.2470962	0.2210873
	SD min	0.1388353	0.1428598	0.1428598	0.1498477
	Random	0.1697467	0.1724377	0.1724377	0.1716547
8	SHR max	0.2307631	0.2276957	0.2276957	0.2129411
	ANR max	0.2894561	0.2389882	0.2389882	0.2154012
	SD min	0.1641641	0.1628907	0.1628907	0.1523117
	Random	0.1726466	0.1656833	0.1656833	0.1515905
9	SHR max	0.2172646	0.2142866	0.2142866	0.2069769
	ANR max	0.2694362	0.2230496	0.2230496	0.2083896
	SD min	0.1571252	0.1561949	0.1561949	0.1500475
	Random	0.2400013	0.1940929	0.1940929	0.1845564
10	SHR max	0.2622553	0.2202458	0.2202458	0.2068282
	ANR max	0.2676537	0.2202458	0.2202458	0.2068282
	SD min	0.1622200	0.1605462	0.1605462	0.1580034
	Random	0.2350234	0.1902347	0.1902347	0.1711720
11	SHR max	0.2048929	0.2009086	0.2009086	0.1946724
	ANR max	0.2481722	0.2071481	0.2071481	0.1955610
	SD min	0.1720071	0.1671873	0.1671873	0.1563165
	Random	0.2281341	0.1890943	0.1890943	0.1791166

max, ANR max, and SD min, the annualized return value of the portfolio at the beginning of the smallest number of clusters provides a fairly high value which then experiences a relatively small decline. However, when the value of k is certain, it will give an increase back to the relatively high annualized return value of the portfolio. In addition, portfolio formation based on ANR max will provide the highest annualized return portfolio compared to others.

Table 12 illustrates that the more clusters formed, the smaller the portfolio sharpe ratio results, especially those formed based on SHR max, ANR max, and SD min. On the other hand, for the formation of portfolios based on random stock selection, the results are less favorable in the calculation of the portfolio's sharpe ratio. For portfolio formation based on SHR max,

Table 12. Portfolio sharpe ratio value from portfolio construction

Cluster	Metrics	Equally Weights	Inverse Volatility	Risk Parity	Mean Variance
2	SHR max	0.0369194	0.0375505	0.0375505	0.0383449
	ANR max	0.0226375	0.0364236	0.0364236	0.0379959
	SD min	0.0292992	0.0321689	0.0321689	0.0342362
	Random	0.0075363	0.0082059	0.0082059	0.0084475
3	SHR max	0.0415180	0.0428702	0.0428702	0.0410126
	ANR max	0.0262606	0.0423501	0.0423501	0.0421037
	SD min	0.0303193	0.0315547	0.0315547	0.0319115
	Random	0.0266730	0.0252362	0.0252362	0.0210533
4	SHR max	0.0398529	0.0416138	0.0416138	0.0411005
	ANR max	0.0301675	0.0455308	0.0455308	0.0445110
	SD min	0.0224816	0.0250171	0.0250171	0.0293406
	Random	0.0243302	0.0259134	0.0259134	0.0305319
5	SHR max	0.0417350	0.0431327	0.0431327	0.0418056
	ANR max	0.0330185	0.0478406	0.0478406	0.0451669
	SD min	0.0276531	0.0298501	0.0298501	0.0340627
	Random	0.0362821	0.0356420	0.0356420	0.0330724
6	SHR max	0.0463476	0.0475516	0.0475516	0.0468928
	ANR max	0.0356573	0.0506996	0.0506996	0.0475513
	SD min	0.0259507	0.0271599	0.0271599	0.0309397
	Random	0.0275202	0.0335441	0.0335441	0.0321279
7	SHR max	0.0479305	0.0490204	0.0490204	0.0477417
	ANR max	0.0370226	0.0511391	0.0511391	0.0476265
	SD min	0.0256084	0.0272427	0.0272427	0.0303719
	Random	0.0293189	0.0299433	0.0299433	0.0307480
8	SHR max	0.0505147	0.0516604	0.0516604	0.0504706
	ANR max	0.0392649	0.0539300	0.0539300	0.0505583
	SD min	0.0314867	0.0314584	0.0314584	0.0298708
	Random	0.0333034	0.0314504	0.0314504	0.0291359
9	SHR max	0.0463984	0.0469854	0.0469854	0.0486589
	ANR max	0.0386448	0.0488611	0.0488611	0.0484856
	SD min	0.0289364	0.0288873	0.0288873	0.0291662
	Random	0.0339244	0.0420081	0.0420081	0.0413647
10	SHR max	0.0400131	0.0488062	0.0488062	0.0489310
	ANR max	0.0400131	0.0488062	0.0488062	0.0489310
	SD min	0.0302027	0.0299400	0.0299400	0.0322664
	Random	0.0352406	0.0408833	0.0408833	0.0375484
11	SHR max	0.0432338	0.0428911	0.0428911	0.0454670
	ANR max	0.0393256	0.0448800	0.0448800	0.0457092
	SD min	0.0334729	0.0322895	0.0322895	0.0320173
	Random	0.0353325	0.0392906	0.0392906	0.0400411

ANR max, and SD min, the sharpe ratio value of the portfolio at the beginning of the smallest number of clusters gives a fairly high value which then experiences a relatively small decrease. However, at a certain k value, it will give a return to a relatively high portfolio sharpe ratio value. In addition, portfolio formation based on SHR max will provide the highest annualized return compared to ANR max, SD min, and Random.

3.5. **Optimal Number of Clusters.** In this study, the optimal number of clusters was determined using the Gini Index method. The value of the Gini Index obtained for each number of clusters is then plotted as shown in Figure 3.

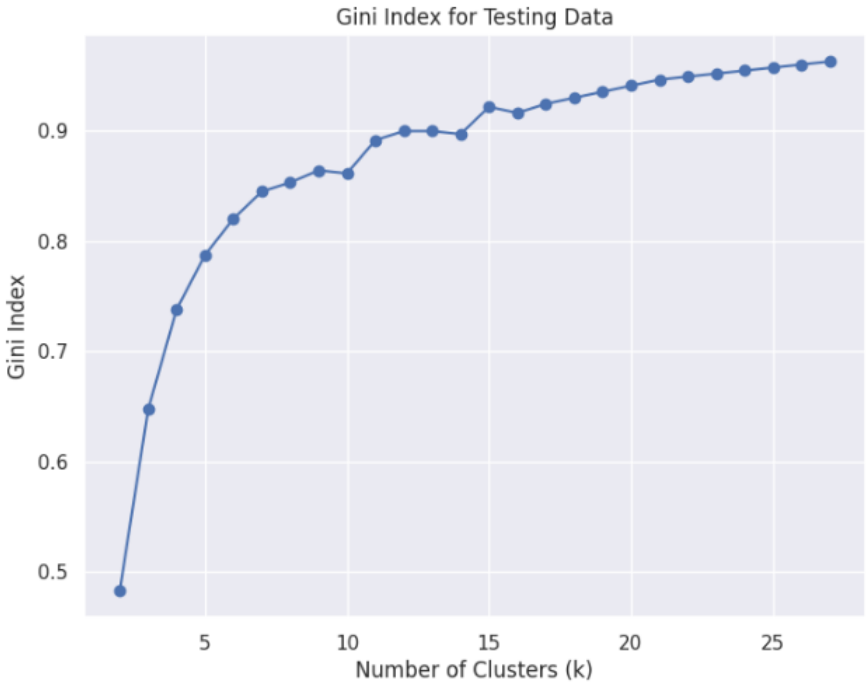

Figure 3. Gini index plot for each cluster

Determining the optimal number of clusters can be chosen from the location of the "elbow" of the Gini Index graph. This concept is based on the idea that the optimal number of clusters reflects the point where adding clusters no longer provides a significant decrease or increase in variance or Gini Index. In Figure 3 above, it can be seen that the Gini Index value increases as the number of clusters increases. The plot also provides the location of the "elbow" on the graph which is the number of clusters to be used for the next step. It can be seen that the "elbow" location is located at a Gini Index value of around 0.85, which means that the optimal number of clusters that can be used is 10 clusters.

4. Conclusions

The conclusions we found from the analysis conducted in this paper are as follows.

(1) This portfolio formation based on the sustainable trend feature improves on previous methods of calculating portfolio returns without considering losses from market declines. The labeling of the continuous trend feature also shows that the portfolio formed based on the RC return has quite good results and has a lower investment risk. It also depends on the price movement of each stock in a situation where it will experience an uptrend or downtrend within a certain time parameter.

(2) The k-means clustering is a distance-based clustering method where the grouping of data objects is based on the distance between different stocks. This provides a difference from the traditional mean-variance portfolio theory that only considers the angular information between stocks to reduce portfolio risk and increase portfolio return.

(3) The collection of stocks that have the maximum return, maximum ANR, minimum STD, and maximum SHR values of each cluster for portfolio formation according to

different weight calculation methods is useful in the selection of the stock pool, that is, which assets should be included in the stock pool for portfolio formation.

(4) The selection of the optimal number of clusters in portfolio formation can use the Gini Index method where the results show that the optimal number of clusters that can be formed is 10 clusters. In addition, the determination of the optimal number of clusters formed can be observed based on the results of experiments that have been carried out which show the same results as the results of the Gini Index method with the optimal number of clusters formed is 10 clusters. This is also supported by the results of the calculation of the sharpe ratio which has a high enough value when 10 clusters are formed which refers to the performance of a company in the formation of the portfolio.

References

[1] Alvarez, M., Luo, Y., Cahan, R., Jussa, J., dan Chen, Z., 2011, *Risk Parity and Risk-Based Allocation*, Portfolios under construction.

[2] Asness, C.S., Frazzini, A., dan Pedersen, L.H., 2012, *Leverage Aversion and Risk Parity*, Financ. Anal. J. 68, 47–59.

[3] Awaysheh, A., Wilcke, J., Elvinger, F., Rees, L., Fan, W., dan Zimmerman, K. L., 2016, *Evaluation of supervised machine-learning algorithms to distinguish between inflammatory bowel disease and alimentary lymphoma in cats*, Journal of Veterinary Diagnostic Investigation, 28(6), 679–687.

[4] Awaysheh, A., Wilcke, J., Elvinger, F., Rees, L., Fan, W., dan Zimmerman, K. L., 2019, *Review of Medical Decision Support and Machine-Learning Methods*, Veterinary Pathology, 56(4), 512–525.

[5] Bechis, L., 2020, Machine Learning Portfolio Optimization: Hierarchical Risk Parity and Modern Portfolio Theory *Thesis*, Luiss Guido Carli, Italy.

[6] Benesty, J., Chen, J., Huang, Y., dan Cohen, I., 2009, *Pearson Correlation Coefficient. In Noise Reduction in Speech Processing*, Springer: Berlin, Germany, pp. 1–4.

[7] Breiman, L., 2001, *Classification and Regression Tree.*

[8] Choueifaty, Y., dan Coignard, Y., 2008, *Toward Maximum Diversification*, J. Portfolio Manag. 35, 40–51.

[9] Cieslak, M.C., Castelfranco, A.M., Roncalli, V., Lenz, P.H., dan Hartline, D.K., 2020, *T-Distributed Stochastic Neighbor Embedding (T-SNE): a Tool for Eco-Physiological Transcriptomic Analysis*, Mar. Genom. 51, 100723.

[10] Dowd, K., 2002, *An Introduction to Market Risk Measurement*, John Wiley and Sons, Ltd., Chichester, England.

[11] Fahim, A., Salem, A., Torkey, F.A., dan Ramadan, M., 2006, *An Efficient Enhanced K-means Clustering Algorithm*, J. Zhejiang Univ.-Sci. A 7 ,1626–1633.

[12] Hartono, J., 2017, *Teori Portofolio Dan Analisis Investasi*, Yogyakarta: BPFE-Yogyakarta.

[13] Hinton, G., dan Roweis, S.T., 2002, *Stochastic Neighbor Embedding. In Advances in Neural Information Processing Systems*, volume 15, pages 833–840.

[14] Hull, J., 2021, *Machine Learning in Business: An Introduction to the World of Data Science.* Independently Published.

[15] Hult, H., Lindskog, F., Hammarlid, O., and Rehn, C.J., 2010, *Risk and Portfolio Analysis: Principles and Methods*, Springer.

[16] Husnan, S., dan Pudjiastuti, 2015, *Dasar-Dasar Manajemen Keuangan Edisi Ketujuh*, Yogyakarta: UPP STIM YKPN.

[17] Lohre, H., Opfer, H., dan Orszag, G., 2014, *Diversifying risk parity*, J. Risk 16, 53–79.

[18] MacQueen, J., 1967, *Some Methods for Classification and Analysis of Multivariate Observations, in: Proceedings of The Proceedings of The Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297.

[19] Maillard, S., Roncalli, T., dan Teïletche, J., 2010, *The Properties of Equally Weighted Risk Contribution Portfolios*, J. Portfolio Manag. 36 (2010) 60–70.

[20] Markowitz, H., 1952, Portfolio Selection, *Journal of Finance*, Vol. 7, 77-91.

[21] Megha, A., 2015, *Developments in Mean-Variance Efficient Portfolio Selection*, The Palgrave Macmillan, New York.

[22] Mumpuni, M., dan Darmawan, H., 2017, *Panduan Berinvestasi Saham Untuk Pemula*, PT. Solusi Finansialku Indonesia.

[23] Pantchev, V., dan Kahra, H., 2017, Factor Investing with Risk Parity Portfolios, *Thesis*, University of Oulu, Finland.

[24] Perdametra, H., 2016, *Laba Akuntansi Dan Arus Kas Operasi Terhadap Perubahan Harga Saham Pada Perusahaan Otomotif Yang Terdaftar Di Bursa Efek Indonesia (BEI)*, Jurnal Akuntansi dan Bisnis.

[25] Sapp, C. E., 2017, *Preparing and Architecting for Machine Learning*, Gartner Technical Professional Advice, 1–38.

[26] Seffens, W., Evans, C., Taylor, H. A., Quarells, R. C., Arnett, D. K., Gibbons, G. H., dan Wilson, J. G., 2015, *Machine Learning Data Imputation and Classification in a Multicohort Hypertension Clinical Study*, Bioinformatics and Biology Insights, 9s3, 43–54.

[27] Sharpe, W.F., *Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk*, J. Finance 19 (1964) 425–442.

[28] Smrithy, G., Balakrishnan, R., dan Sivakumar, N., 2019, *Anomaly Detection Using Dynamic Sliding Window in Wireless Body Area Networks*, In Data Science and Big Data Analytics, Springer: Berlin, Germany, pp. 99–108.

[29] Tandelilin, 2001, *Analisis Investasi dan Manajemen Portofolio Edisi Pertama*, Yogyakarta : BPFE.

[30] Ünlü, R., dan Xanthopoulos, P., 2021, *A Reduced Variance Unsupervised Ensemble Learning Algorithm Based on Modern Portfolio Theory*, Expert Syst. Appl. 180.

[31] Van der Maaten, L., dan Hinton, G., 2008, *Visualizing data using T-sne*, J. Mach. Learn, Res. 9.

[32] Wu, D., Wang, X., Su, J., Tang, B., dan Wu, S., 2020, *A Labeling Method for Financial Time Series Prediction Based on Trends*, Entropy 22, 1162.

[33] Wu, D., Wang, X., dan Wu, S., 2021, *A Hybrid Method Based on Extreme Learning Machine and Wavelet Transform Denoising for Stock Prediction*, Entropy 23, 440.

[34] Wu, D., Wang, X., dan Wu, S., 2022, *Construction of Stock Portfolios Based on K-Means Clustering of Continuous Trend Features*, Knowledge-Based Systems, Vol. 252.