BAREKENG: Journal of Mathematics and Its Applications
March 2024 Volume 18 Issue 1 Page 0589–0602
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol18iss1pp0589-0602

K-MEANS AND AGGLOMERATIVE HIERARCHY CLUSTERING ANALYSIS ON THE STAINLESS STEEL CORROSION PROBLEM

Yuli S. Afrianti [1*], Udjianna S. Pasaribu[2], Fadhil H. Sulaiman[3], Grace Angelia[4], Henry J. Wattimanela[5]

[1,2]Statistics Research Division, Faculty of Mathematics and Natural Science, Bandung Institute of Technology
[1]Doctoral Program in Mathematic, Faculty of Mathematics and Natural Science, Bandung Institute of Technology
[3,4]Undergraduate Program in Mathematics, Faculty of Mathematics and Natural Science,
Bandung Institute of Technology,
Ganesha Street No 10, Bandung, 40132, Indonesia.

[5] Statistics Study Program, Mathematics Department, Faculty of Mathematics and Natural Science,
Pattimura University

Ir. M. Putuhena Street, Poka, Ambon, Maluku, 97223, Indonesia.

*Corresponding author's e-mail: * **yuli.afrianti@itb.ac.id***

ABSTRACT

Article History:

Received: 3rd September 2023
Revised: 24th January 2024
Accepted: 6th February 2024

Keywords:

Clustering Analysis;
Corrosion;
Hierarchy;
K-Means;
Stainless Steel.

Stainless Steel (SS) is a material that is widely used in various fields because it is resistant to corrosion. However, if SS is exposed to heat at high temperatures for a long period of time, a sigma phase, namely the Fe-Cr compound, will form, which indicates that corrosion has begun. The appearance of this corrosion can be detected through color changes on the SS surface, ranging from light brown to dark blue. Corrosion events will be observed through the distribution of color on the sample surface at the location selected through the SS microstructure image. Cluster analysis will be used to group the colors on the surface of the SS sample through the images used. The results of cluster analysis can be used to identify SS color which indicates the appearance of corrosion in the sample. In this research, we will examine the determination of many clusters for K-Means and Agglomerative Hierarchy with Ward's Criterion, Single, Average, and Complete Linkages. In addition, the model quality measure was tested with Silhouette Coefficient. Single linkage gives the worst results because it gives the impression that only one dominant color appears so it can be said that it is unable to distribute each color to the specified cluster. Likewise with the average, because the number of clusters cannot be determined with certainty. On the other hand, the K-Means results are similar to Ward's results, this is reasonable because the basic idea of both is to find the minimum distance between each object and its center, in this case, the average is used as the measure of the center, while the results that are most similar to the original image are clustered using complete linkage. These results can be used as recommendations for academics and practitioners in the fields of Statistics, Mathematics and Materials Engineering in the subsequent analysis process to solve SS corrosion problems.

How to cite this article:

Y. S. Afrianti, U. S. Pasaribu1, F. H. Sulaiman, G. Angelia and H. J. Wattimanela., "K-MEANS AND AGGLOMERATIVE HIERARCHY CLUSTERING ANALYSIS ON THE STAINLESS STEELS CORROSSION PROBLEM," *BAREKENG: J. Math. & App.*, vol. 18, iss. 1, pp. 0589-0602, March, 2024.

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

Stainless Steel (SS) is a material that is widely used in various fields because it is resistant to stains, discoloration, or loss of mass due to rusting [1]. However, if SS is exposed to heat at high temperatures for a long time, a sigma phase, Fe-Cr compound, will form, which indicates that corrosion has begun to appear [2]. The appearance of this corrosion can be detected through changes in color on the surface of the SS, ranging from brown to blue. At first, all of the SS surfaces will be brown, then due to prolonged heating at high temperatures, a blue color will begin to appear at the edges of the area, until the entire surface turns all blue. The appearance of this blue color indicates that the SS is starting to corrode [3]. In this research, the SS sample that will be used are those that has been used for 24 years. After being subjected to a series of procedures in the materials laboratory, an image of the SS microstructure will be obtained. Corrosion events will be observed through the distribution of color on the surface of the sample at the selected location through the SS microstructure image. Next, cluster analysis will be used for the color grouping process on the surface of the SS sample through the image used. Then, the result can be used to identify the color of SS which indicates the appearance of corrosion.

Cluster analysis is a statistical method used to group a set of objects into a number of clusters that have the same characteristics. To be able to group the characteristics of the observation objects, the similarity of data in a cluster must be maximized or the similarity of data with other clusters will be minimized. There are various types of cluster analysis that can be used, some of which are K-Means, Affinity Propagation, Mean Shift, Spectral Clustering, Hierarchy, DBSCAN, OPTICS, and GMM. In previous research, the K-Means and GMM methods were used to identify corrosion [4], [5]. Identification using this method can replace manual identification, which takes more time, energy, and even money. However, in previous research, the segmentation was still in the form of two categories: whether or not there is corrosion. In this paper, apart from wanting to distinguish areas affected by corrosion, we also want to look at more categories of corrosion that have occurred. As already mentioned, corrosion is formed from several steps, which can be characterized by the appearance of different colors. It is hoped that in further research this category can be analyzed according to its corrosion development.

The cluster analyses used in this research are K-Means clustering and Hierarchy (Agglomerative Hierarchy) clustering. K-Means is a centroid model cluster analysis that groups n observations into k clusters based on their average. K-Means has been applied in various subjects, for example, in mapping the graduated or dropped-out students [6] and estimating the cost of hospitalization [7]. Next, the hierarchical clustering algorithm is a connectivity model that splits the data sets recursively into smaller clusters. Furthermore, agglomerative is a hierarchical clustering method that groups data in a bottom-up manner. Some applications of agglomerative clustering are film recommendation systems [8], clustering of hot spots to prevent forest fires [9], and product sales forecasting [10]. The K-Means method is the simplest method of segmentation. However, because of its simplicity, K-Means has many shortcomings, one of which is that the cluster sizes tend to be the same size. For this reason, the agglomerative hierarchy method was chosen, which can segment data into clusters with different sizes. The results of these two methods will be compared, and conclusions will be drawn regarding which method is better for segmenting corrosion. In the next chapter, we will discuss further how these two methods work.

2. RESEARCH METHODS

Cluster analysis is a technique to group similar observations into several clusters based on the observed values of several variables for each individual [11]. Cluster analysis aims to group objects based on the characteristics of each object. The general steps of cluster analysis are [12]: (1) choose a distance measure, (2) perform data standardization process (if needed), (3) choose a clustering procedure, and (4) perform interpretation of the clusters formed. Clustering can be categorized into seven groups that are Hierarchical clustering, Density-based clustering, partitioning clustering, Graph-based, Grid-based, Model-based, and Combinational clustering [13]. In this study, K-means and Agglomerative Hierarchy clustering procedures will be used. The difference approach between K-means and Agglomerative is displayed in **Figure 1**.

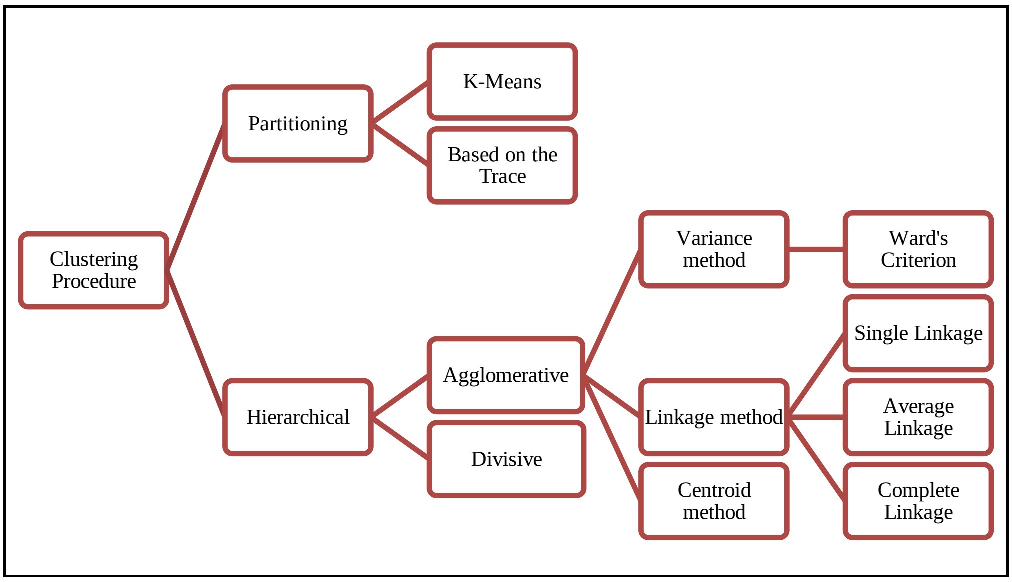


Figure 1. Clustering chart

2.1 K-Means

K-Means clustering algorithm was proposed independently by different researchers, including Steinhaus [1956], Lloyd [1982], MacQueen [1967], and Jancey [1966] from different disciplines during the 1950s and 1960s [14]. These researchers' various versions of the algorithms show four common processing steps with differences in each step [15]. The first time that sparked this clustering idea was Lloyd in 1957, but it was only published in 1982. In 1965, Forgey also published the same technique as Lloyd's, so this clustering method is known as Lloyd-Forgy in some sources.

K-Means is a partitional clustering algorithm that groups n data $(\vec{x}_i, i = 1, \dots, n)$ into k clusters $(C_r, r = 1, \dots, k)$. Each cluster is determined by a centroid which is the mean value of the cluster that is:

$$c_r = \frac{1}{N_{C_r}} \sum_{x_p \in C_j} \vec{x}_p \tag{1}$$

where N_{C_r} is a cardinality of cluster r. Suppose $L_2(\vec{x}_p, \vec{c}_q)$ is Euclidian distance between $\vec{x}_p$ dan $\vec{c}_q$, the K-Means aims to choose centroids that minimize the inertia, or within-cluster sum-of-squares criterion (SSE):

$$SSE = \sum_{q=1}^{k} \sum_{x_p \in C_q} \left[L_2(\vec{x}_p, \vec{c}_q)\right]^2 \tag{2}$$

In the K-Means algorithm, initially k centroids are initiated randomly. Then each data will be labeled according to the minimum distance from each data to each of the centroids. Distance determination can be calculated using various distance matrices, such as Euclidean, Manhattan, and Cosine Similarity. After that, a new centroid for each cluster will be created by calculating the average of all members of the cluster. This process guarantees to obtain a local minimum SSE value. The labeling process and centroid calculation will continue until the stopping criteria are met. Flowchart of K-Means can be seen in Figure 2.

Some advantages of using K-Means are relatively simple to implement, scales to large data sets, and guarantees convergence to a local minimum. However, there are disadvantages to this method which is manually selects the number of clusters, affected by initial centroid, and cannot group data with varying sizes and densities [16]. The results of K-Means cluster analysis tend to produce clusters of the same size.

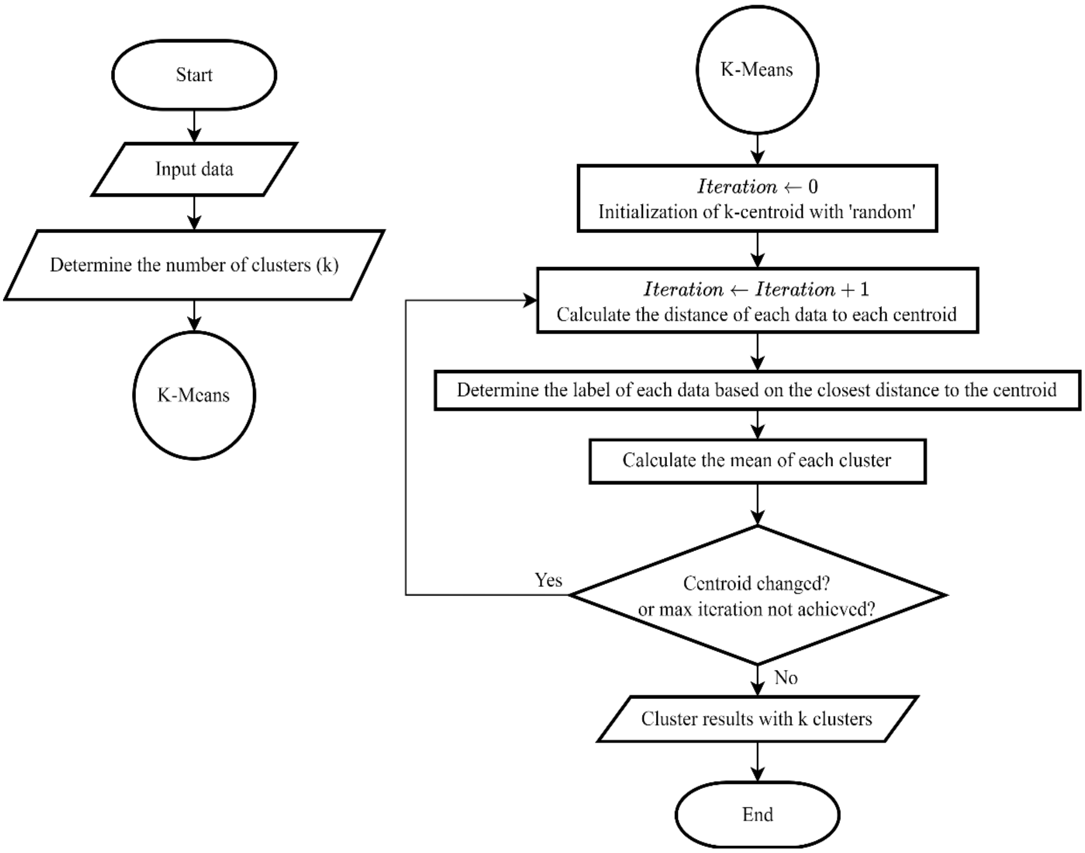

Start
Input data
Determine the number of clusters (k)
K-Means
K-Means
Iteration ← 0
Initialization of k-centroid with 'random'
Iteration ← Iteration + 1
Calculate the distance of each data to each centroid
Determine the label of each data based on the closest distance to the centroid
Calculate the mean of each cluster
Yes
Centroid changed?
or max iteration not achieved?
No
Cluster results with k clusters
End

Figure 2. Flowchart of K-Means

2.2 Agglomerative Hierarchy

Hierarchy clustering divides unlabeled data into different clusters according to their similarity. The data similarity is measured based on the linkage criterion [17]. Agglomerative is a hierarchical clustering method that groups data using a bottom-up manner. The way this cluster analysis works is by considering each data as a cluster that has only one member, and then two clusters that have similarities will be grouped into one new cluster. This process will continue to repeat until only one cluster remains [18], [19]. Agglomerative can be represented with a dendrogram, which shows the relationship between existing clusters [19]. Figure 3 describes the flowchart of Agglomerative. The advantages of using agglomerative are that it is easy to identify nested clusters, providing better results and ease of implementation, suitable for automation, reduces the effect of initial values of the cluster, and reduces the computing time and space complexity. On another side, the disadvantages are difficulty in handling different sizes, no direct minimization of objective function, and sometimes there is difficulty in identifying the exact number of clusters by the Dendrogram [20].

In Agglomerative Hierarchy, a similarity matrix will be determined based on the linkage criterion for each pair of data. This process takes up a lot of space in memory depending on the amount of data being processed. For solving this problem, the K-Nearest Neighbours (KNN) algorithm will be used, which is an algorithm used to identify the nearest neighbours of each data point in the dataset. This process involves calculating the distance between data points and selecting *k*-nearest neighbours for each point. The results of the KNN process are then used to form a connectivity matrix. This connectivity matrix represents the relationship between pairs of data points in the dataset. The matrix element is assigned the value 1 if the two data points are adjacent, and 0 if not. The connectivity matrix that has been created will be used as an initiation step in the Agglomerative Hierarchy process. This helps in defining initial clusters and forms the basis for subsequent merging. KNN aims to reduce the amount of data that will be processed so that less memory is needed. This process involves grouping several elements into a single unit, where the group reflects the merger of *k*-nearest neighbours.

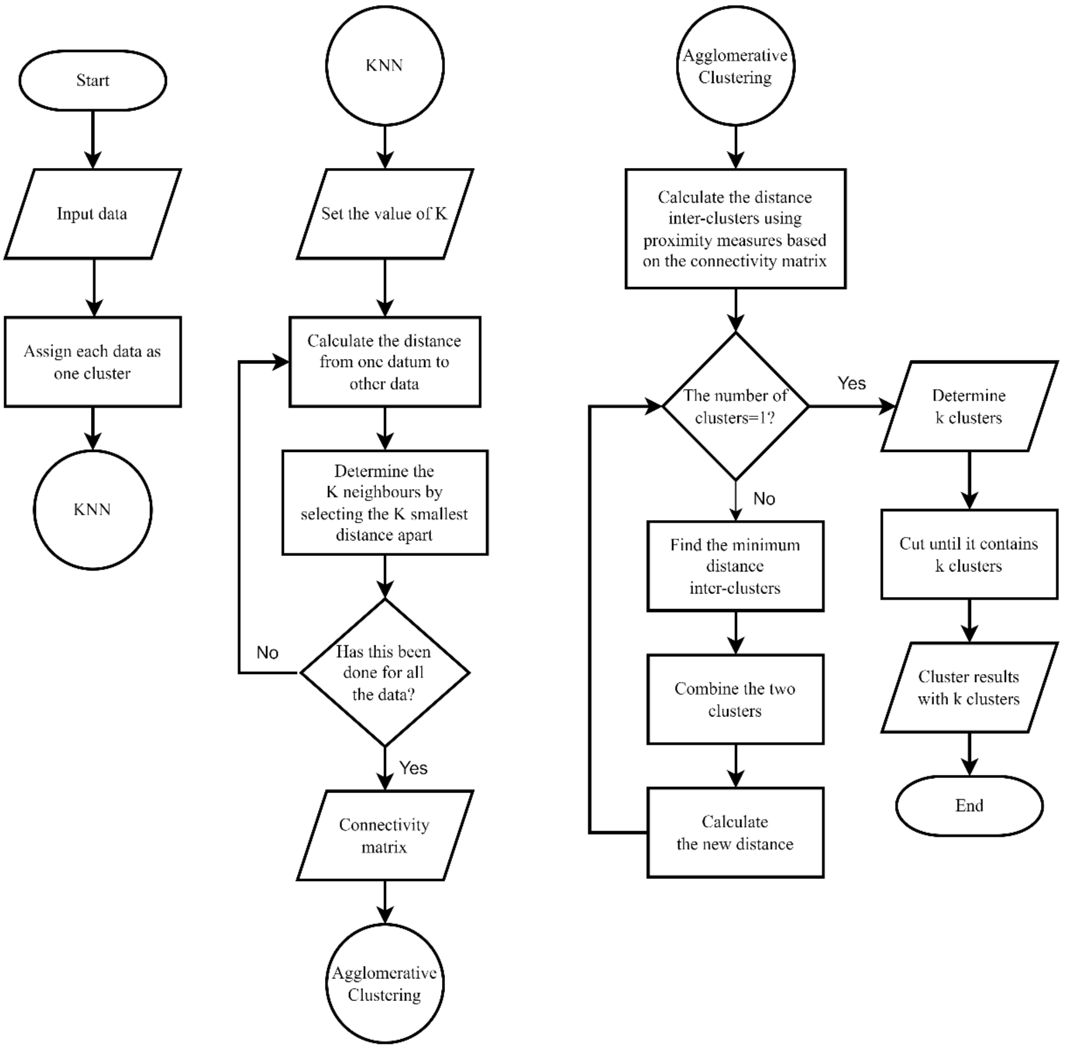

Start
Input data
Assign each data as one cluster
KNN
KNN
Set the value of K
Calculate the distance from one datum to other data
Determine the K neighbours by selecting the K smallest distance apart
Has this been done for all the data?
No
Yes
Connectivity matrix
Agglomerative Clustering
Agglomerative Clustering
Calculate the distance inter-clusters using proximity measures based on the connectivity matrix
The number of clusters=1?
Yes
No
Find the minimum distance inter-clusters
Combine the two clusters
Calculate the new distance
Determine k clusters
Cut until it contains k clusters
Cluster results with k clusters
End

Figure 3. Flowchart of Agglomerative Hierarchical

Suppose that n data $(\vec{x}_i, i = 1, \dots, n)$ will be clustered with agglomerative hierarchy into k clusters $(C_r, r = 1, \dots, k)$. Suppose also $d(\vec{x}_p, \vec{x}_q)$ is a distance metric. There are four types of proximity measures used for this research described in the next sub section that is Ward's Criterion, Single Linkage, Average Linkage, and Complete Linkage.

2.2.1 Ward's Criterion

A technique that combines clusters by minimizing the variance of the clusters being merged [23]. The variance between clusters defined by:

$$H_1(C_r, C_s) = \frac{\left(N_{C_s} - N_{C_r}\right)^2}{\left(N_{C_r} + N_{C_s}\right)\left(N_{C_r} N_{C_s}\right)} \left(\sum_{p=1}^{N_{C_r}} \vec{x}_p - \sum_{q=1}^{N_{C_s}} \vec{x}_q \right)^2 \tag{3}$$

where N_{C_r} and N_{C_s} is the cardinality of C_r and C_s.

2.2.2 Single Linkage

Single linkage clustering method can find arbitrarily shaped clusters in many applications such as image segmentation, spatial data mining, geological mapping, etc. [21]. The disadvantages of using this method are that it is sensitive to the presence of outliers and is not sensitive to the shape and size of clusters [22]. This technique combines clusters based on the minimum distance between members between two clusters. The distance between two clusters is the minimum distance between their members as [23], [24]:

$$H_2(C_r, C_s) = \min_{\vec{x}_p \in C_j, \vec{x}_q \in C_k} d(\vec{x}_p, \vec{x}_q) \tag{4}$$

2.2.3 Average Linkage

Average Linkage combines clusters based on the average distance of each pair of members in the set between the two clusters. The disadvantage to using this method is that it is sensitive to the shape and size of the cluster [22]. This method defines the distance between two clusters as the average distance between their members [23], [24]:

$$H_3(C_r, C_s) = \frac{1}{N_{C_r} + N_{C_s}} \sum_{p=1}^{N_{C_r}} \sum_{q=1}^{N_{C_s}} d(\vec{x}_p, \vec{x}_q) \tag{5}$$

2.2.4 Complete Linkage

This technique combines clusters based on the maximum distance between members between two clusters. The disadvantages of using this method is not affected by the presence of outliers and has problems with convex data shapes [22]. It defines the distance between clusters as the maximum distance between their members [22]:

$$H_4(C_j, C_k) = \max_{\vec{x}_p \in C_j, \vec{x}_q \in C_k} d(\vec{x}_p, \vec{x}_q) \tag{6}$$

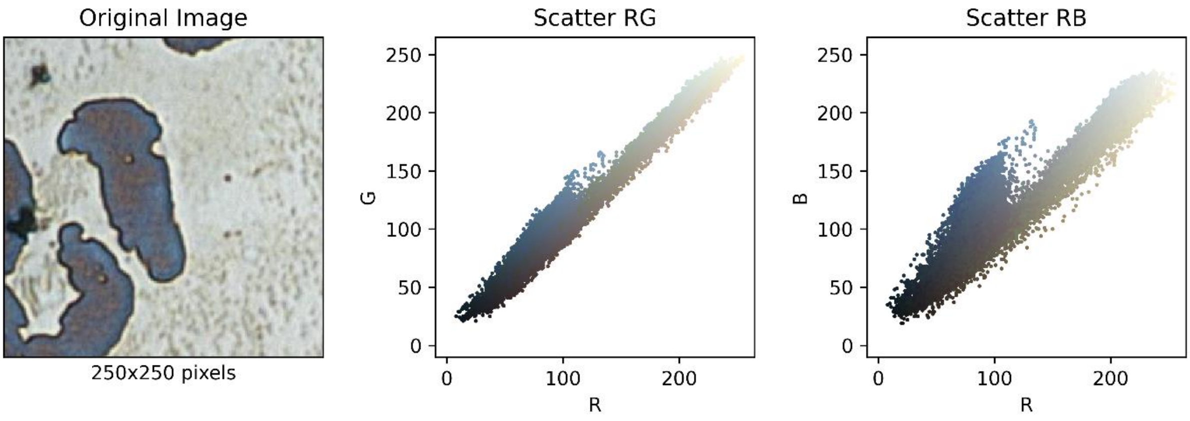
2.3 Silhouette Coefficient

The Silhouette Coefficient is a method for measuring clustering performance based on differences in distance of inter and intra clusters. The Silhouette Coefficient is in the range -1 and 1, where -1 represents the worst value and 1 represents the best value. Let $\boldsymbol{a_i}$ represent the average distance of $\vec{\boldsymbol{x}}_{\boldsymbol{i}}$ to all points in the intra-cluster and $\boldsymbol{b_i}$ is the average distance between $\vec{\boldsymbol{x}}_{\boldsymbol{i}}$ and the nearest cluster that the $\vec{\boldsymbol{x}}_{\boldsymbol{i}}$ is not a part of. The following is a calculation of the Silhouette Coefficient t.

$$\boldsymbol{SC} = \frac{\sum_{i=1}^{n} \frac{\boldsymbol{b_i - a_i}}{\mathbf{max}(\boldsymbol{a_i, b_i})}}{\boldsymbol{n}} \tag{7}$$

3. RESULTS AND DISCUSSION

A sample of the microstructure image of Stainless Steel that has been in use for 24 years was obtained after going through a series of procedures in the materials laboratory (See Figure 1 A). In this image, it can be clearly seen that the white part is a region of stainless steel that is not corroded. Then the light brown to dark brown color indicates that the corrosion process is just starting to occur. Meanwhile, the part that affected by corrosion is bluish brown, light blue, and dark blue, which indicates the formation of the sigma phase. This 250 × 250 pixels image will be used as input in the cluster analysis. The input of the image is the RGB value of each pixel, so there are $n = 625000$ RGB data with values ranging from 0 to 255 (Uint8 type). From this data the RG and RB scatterplots were created to describe their values in two dimensions (See Figure 1 B and Figure 1 C). It can be seen that its form extends from the bottom left which is dark blue to the top right which is white. This form can be divided into two areas such as the area affected by corrosion at the bottom left which is dark blue to light brown and also the background area at the top right which is white. Apart from that the density of the data is very close together making the data appear to be continuous.



(A) (B) (C)

Figure 4. **Microstructure image of Stainless Steel**

Notationally, each data in the image can be expressed by $\vec{x}_i = (R_i,\ G_i,\ B_i)$ with $R_i,\ G_i,\ B_i$ is the RGB value for each i-th data for $i = 1,\ 2,\ \dots,\ n$. Next, a distance measure is selected in the form of Euclidean distance that is:

$$L_2(\vec{x}_i,\ \vec{x}_j) = \sqrt{(R_i - R_j)^2 + (G_i - G_j)^2 + (B_i - B_j)^2} \tag{8}$$

Because the type of RGB data is Uint8 (0-255), it means that the range of the R, G, and B values are already the same size, therefore the standardization process is not needed. The cluster analysis will use two algorithms, the K-Means and Agglomerative Hierarchical with four proximity measure (Ward's Criterion, Single Linkage, Average Linkage, and Complete Linkage). For the K-Means algorithm, this research used several other parameters that can be seen in **Table 1**.

Table 1. **Other Parameters of K-Means in This Research**

Parameter	Value	Information
Initialization of centroid	Null	Random
Number of runs	10	The number of runs carried out with different random seeds. The final result is the best K-Means model with the lowest SSE
Maximum iteration	300	The algorithm is stopped when it reaches the maximum iteration
Tolerance	0.0001	The algorithm is stopped when the difference in distance from the centroid from the previous iteration is less than the tolerance

Apart from that, for the Agglomerative algorithm, the amount of data that will be processed is very large which results in large memory usage. However, due to memory limitations, before running the algorithm the connectivity matrix is first calculated using KNN (K-Nearest Neighbours). In the KNN algorithm, 10 nearest neighbours ($k = 10$) are selected as initial input.

The calculation of the algorithms uses the Python code with the sklearn.clustering library that is KMeans and AgglomerativeClustering. Meanwhile, the KNN algorithm is carried out using sklearn.neighbors library, namely kneighbors_graph.

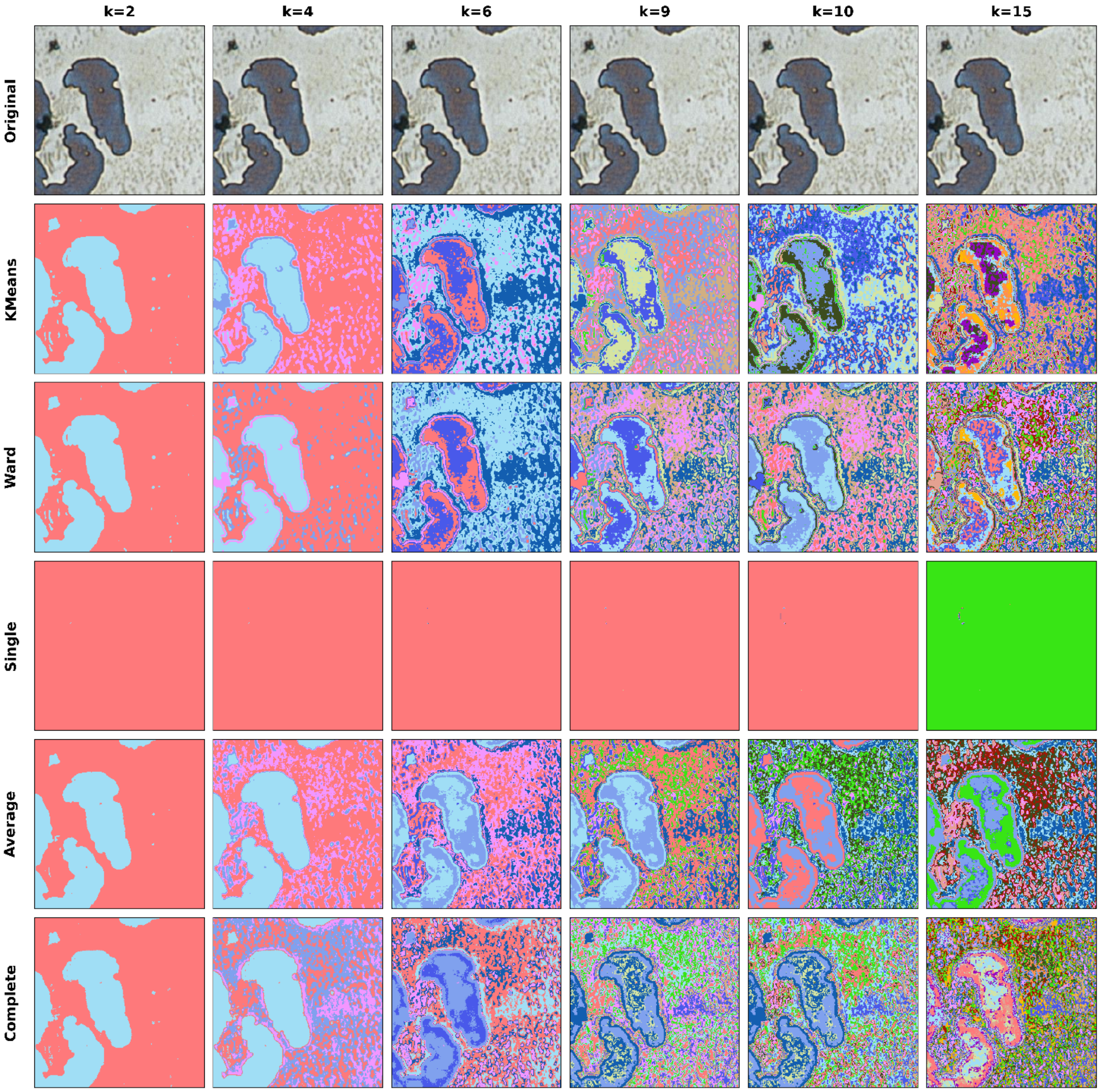


Figure 5. Overlay of Labels for Each Method With Various Number Of Clusters

3.1 Reducing Clustering Methods

First of all, each clustering method is run for number of clusters ranging from 1 to 15 and the results will be analysis one by one and some overlay results can be seen in Figure 5. One of the most striking results is Single Linkage method failed to classify colors at all number of clusters. There is one cluster that covers almost all the data which is depicted with one dominant color. The failure of this method is closely related to the way it works which considers the minimum distance between two clusters. Previously, as has been noticed in Figures 4 B and Figure 4 C that the values of the data were very close together so they seem continuous. This condition causes the Single Linkage to form one large cluster, not in accordance with the 5 color categories that have been determined in the corrosion process. Therefore, the Single Linkage will not be discussed further.

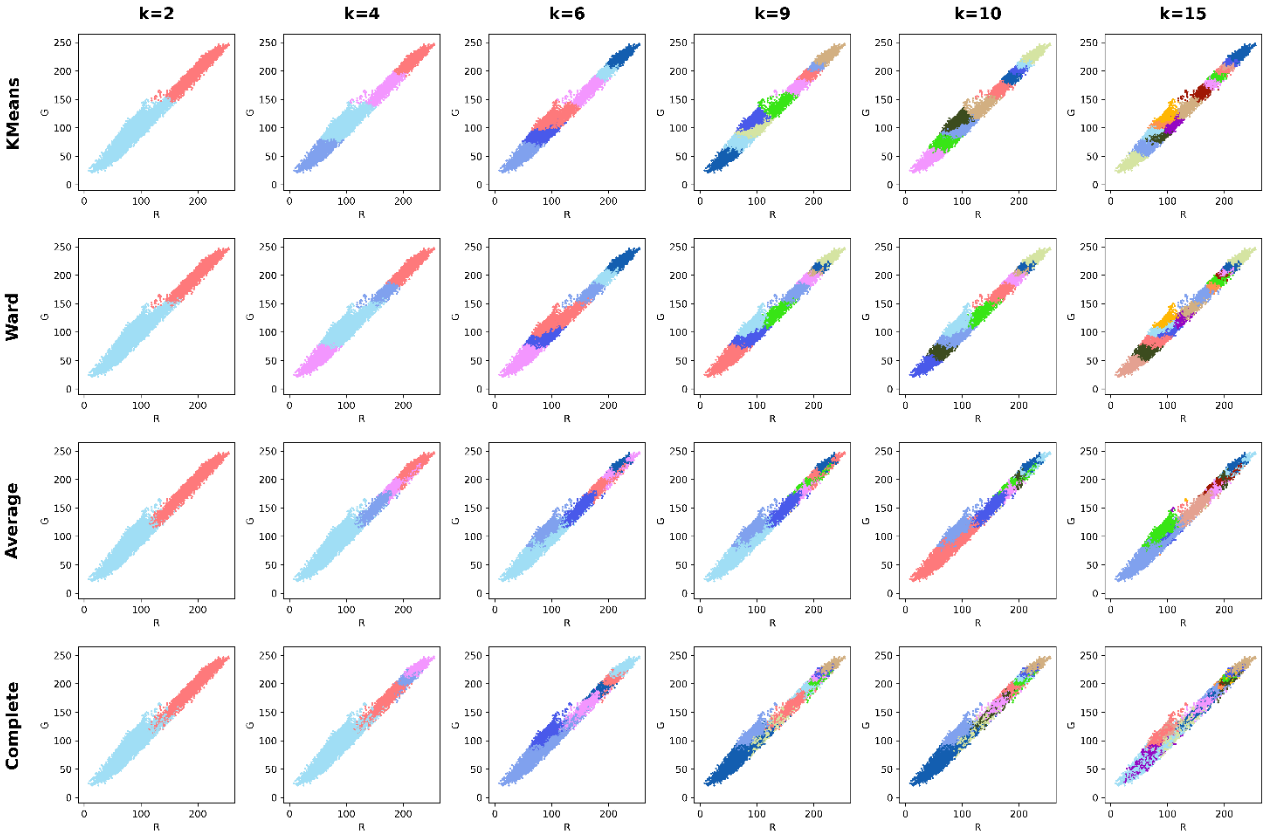


Figure 6. Scatter plot of labels for each method with various number of clusters

3.2 Optimal Number of Clusters

To analyze the optimal number of clusters, the steps taken are checking the overlay (See **Figure 5**) and compare it with the scatterplot (See **Figure 6**). However, we will only review RB because the analysis results via RG and RB are similar. In this analysis, each method must be able to divide the data into two areas, that is the foreground (the area affected by the corrosion process) and the background (the area not affected by the corrosion process). The foreground is mostly at the bottom left, while the background is at the top right of the scatter plot. Furthermore, to understand variations in the intensity of the corrosion process, the foreground area will be divided into five categories, starting from light brown to dark blue. Meanwhile, the background area is shown in white. Therefore, there must be five clusters at the bottom left of the scatter plot which are the foreground, while the other clusters above them must be the background. Lastly, to determine the clusters are included in the foreground and background can be analyzed in the overlay image.

Initially at $k = 2$, each method succeeded in separating areas affected by corrosion from areas not affected by corrosion. However, as the number of clusters increases, all methods tend to classify only the background part. At $k = 6$, these methods are starting to classify the color of the part affected by corrosion, but still only in 2 or 3 categories. Moreover, by increasing the number of clusters, the area not affected by the corrosion process is in the upper third of the scatterplot. Meanwhile, the area affected by the corrosion process which is must be divided into five clusters, is in the remaining two-thirds.

Next, for number of clusters $k = 9$, the K-Means algorithm and Average method succeeded in obtaining 5 clusters on the part affected by corrosion process. However, the classification results using the Average method are not optimal, because there are two clusters that only classify a small amount of data so it can be considered to only have 3 clusters. Meanwhile, the Complete and Ward methods are only able to classify corrosion color into 4 clusters. Later at $k = 10$, Ward methods succeeded in producing 5 clusters. After that, Complete can divide 5 clusters when $k = 15$. On the other hand, regarding the Average method, even though $k = 20$, it cannot divide 5 clusters well. This method fails to classify because it can only form small clusters at the later stage. So, it can be concluded that K-Means achieves the optimal number of clusters at $k^* = 9$, Ward at $k^* = 10$, and Complete at $k^* = 15$.

3.3 Efficiency of Clusters

The analysis results show that each method has different number of clusters for classifying 5 categories of corrosion process on SS. In terms of clustering efficiency, the K-Means method is proven to be more efficient than other methods. With the smallest k value, the K-Means method is able to classify the corrosion process into five parts. In contrast, the average method which requires larger number of clusters to achieve optimal corrosion color classification.

3.4 Characteristics of Clusters

On the other hand, we need to consider the characteristics of the clustering results. As explained, the K-Means method tends to divide cluster areas with relatively the same size. This effect can also be seen in the scatter plot that has been produced (See **Figure 6**). Similarly, Ward also has a similar effect. This similarity is due to the fact that both algorithms have the same goal, that is minimizing variance. However, after running Ward's method, it was found that this method required large memory size and longer computing time. This makes K-Means more efficient in terms of memory and computing time. On the other hand, other methods do not divide clusters uniformly. This can be seen from the size of the dark blue clusters is much larger than other clusters.

In real situations, the characteristics of the corrosion process on SS are definitely not equally distributed because over time the part that is exposed to continuous heat, which is brown, will turn blue. It can be concluded that Complete Linkage is a better method based on its characteristics.

3.5 Quality Aspect of Clustering Model

The quality measure of model will be assessed using the Silhouette Coefficient (See **Figure 7**). The figure shows the Silhouette Coefficient for each method for the number of clusters ranging from 2 to 15. From this figure, it can be seen that the K-Means and Ward methods have similar values, as do the Average and Complete methods. In contrast, the Single method stands separately. Previously, it was explained that the K-Means and Ward methods are similar in how they work. Interestingly, the Average and Complete Linkages, although they have differences in how they work, can produce similar Silhouette Coefficient. We suspect that this phenomenon may be caused by very high data density, which requires further research to understand it in depth.

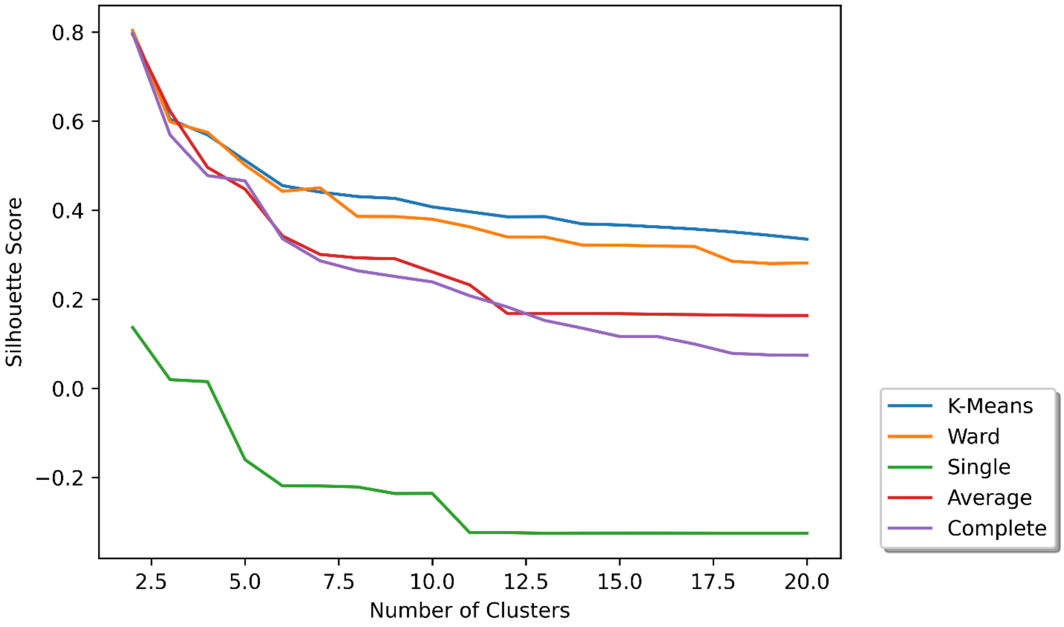


Figure 7. Silhouette Coefficient of the Clustering Methods

Based on Figure 7, the Silhouette Coefficient for the Single Linkage is far below of the other methods and most of them show negative values. These results are not surprising, considering that the Single Linkage has failed in classifying corrosion process. To analyze the remaining four models, the discussion will be divided into two groups. The first group, which consists of K-Means and Ward has a higher Silhouette Coefficient, indicating that these two methods are better than the second group which consists of Average and Complete. Initially the silhouette values of the all groups were similar, but at $k = 5$, there began to be differences between the two groups. This difference shows that the first group can produce better and more consistent groupings in optimal number of clusters.

By looking again at the optimal cluster results obtained, the K-Means results (with $k^* = 9$) have a silhouette coefficient of 0.4265, while Ward (with $k^* = 10$) is 0.3798 and Complete (with $k^* = 15$) is 0.1164. These results indicate that K-Means is a better method based on model quality.

Besides that, in the Average Linkage, the Silhouette Coefficient value tends to be constant starting from $k = 12$. It means that the quality of the model produced by this method does not change as the number of clusters increases. This is because the method produces cluster with a small number of members as the number of clusters increases.

3.6 Quality Aspect of Cluster

From the results of clustering with an optimal number of clusters, five clusters were determined as the foreground, this is adjusted to five colors as indicators of corrosion levels. Then the colors of each cluster are collected and presented in Figure 8. Then each cluster was assigned five color categories, namely light brown, dark brown, bluish brown, light blue and dark blue. In the K-Means and Ward method, the categories suitable for each cluster are dark brown, bluish brown, light blue, dark blue, and light brown. Meanwhile, the categories that are suitable for the Complete method for each cluster are dark brown, bluish brown, light blue, light brown, dark blue. From the color categories it can be seen that the Complete method better represents the corrosion process on SS. In addition, it is no longer surprising that the K-Means and Ward methods have similar results, because the formula for determining distance in both methods is almost the same. The percentage of points in each cluster also gives the same results for K-Means and Ward, namely dominated by light blue (around 12%).

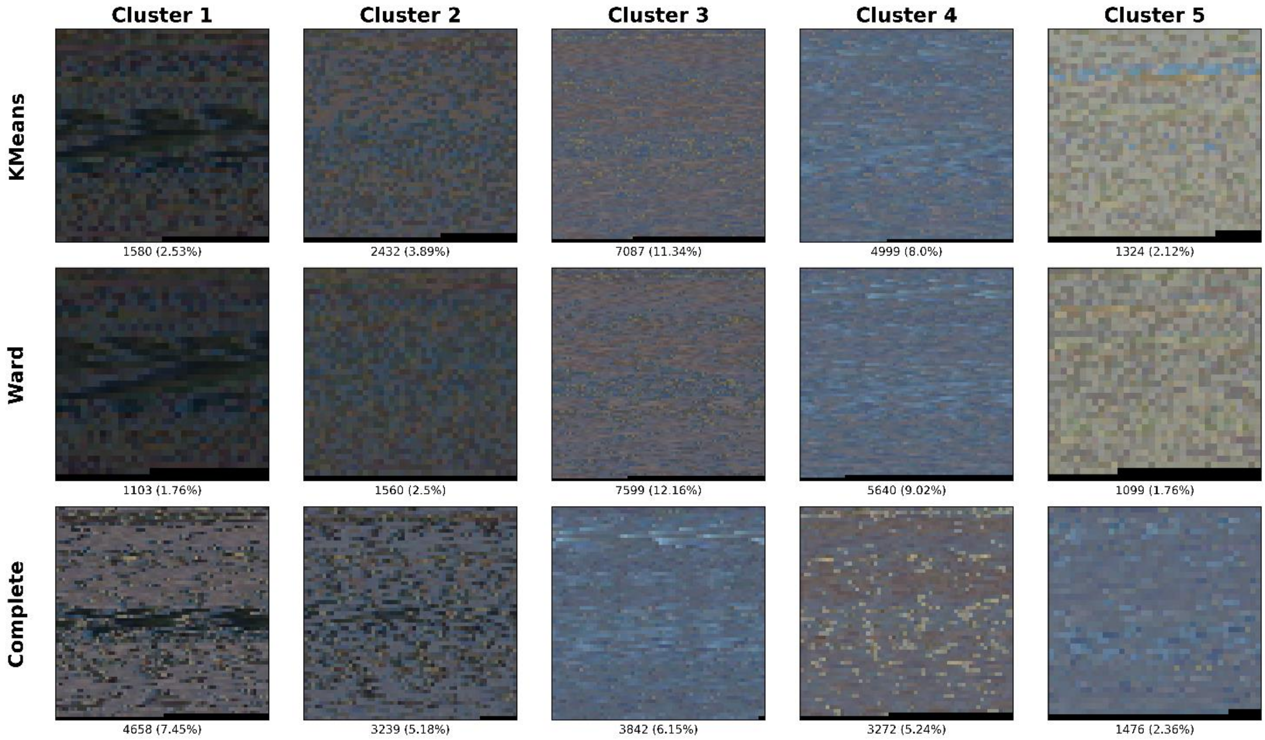


Figure 8. Collection of colors in each cluster with the most optimal method

3.7 Dendrogram of Agglomerative Result

Another result of Agglomerative can be displayed as a Dendrogram (see Figure 9). Dendrogram is a tree that shows how clusters are merged hierarchically [26]. Again, Single still gives the worst results, whereas Ward, Average, and Complete give slightly similar.

In the initial stage of the Agglomerative procedure, each object is considered as one cluster, so that many objects equal many clusters. Furthermore, each object will combine with other objects if they are similar based on the respective distances used in Ward, Single, Average, and Complete. This merging process will occur continuously until there is only one cluster. Figure 9 explains the Dendrogram for Agglomerative (Ward, Single, Average, and Complete). The Dendrogram describes clustering organizing a large amount of information into a small number of clusters that provide some meaningful information [27].

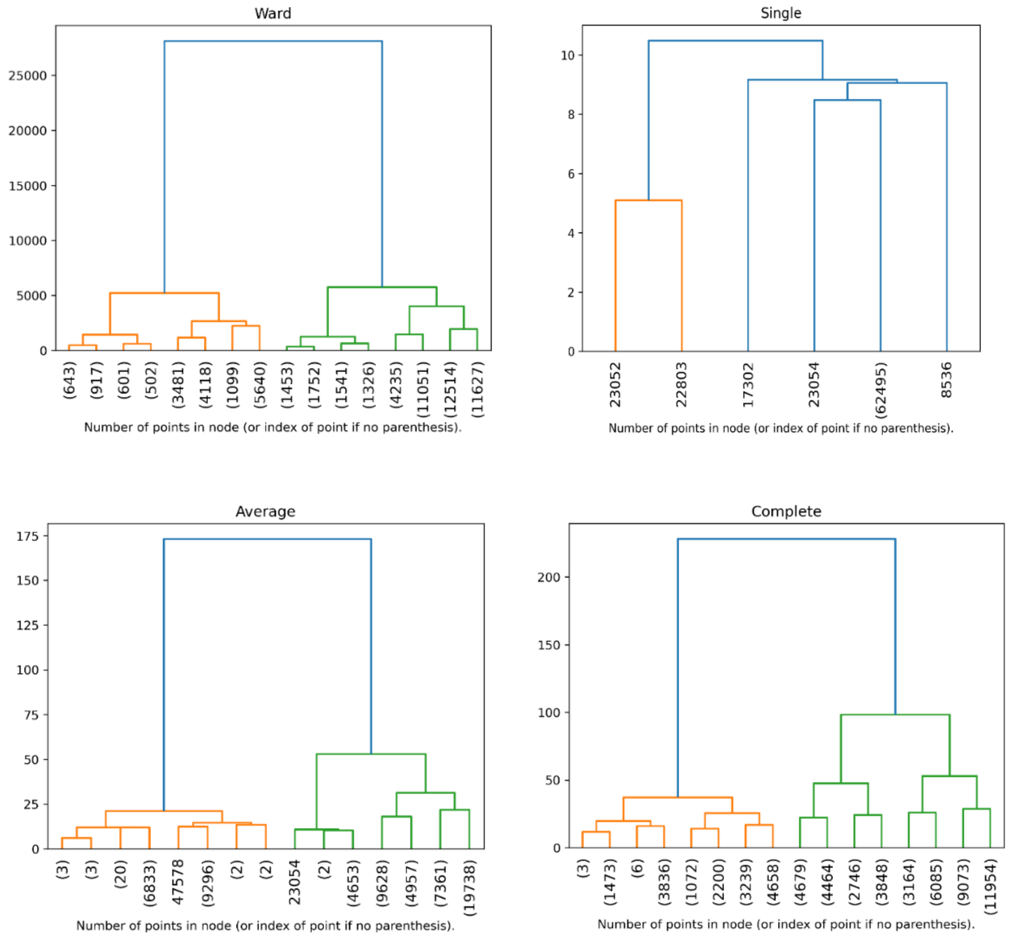


Figure 9. Dendrogram of Agglomerative with Ward, Single, Average, and Complete

4. CONCLUSIONS

Cluster analysis has been used to group surface colors in the microstructure of Stainless Steel images. Based on the results, the K-Means results are similar to Ward's results, this is reasonable because the basic idea of both is to find the minimum distance between each object and its center, in this case, the average is used as a measure of its center. Then, the similarities between these two methods can be seen from the results of the model quality aspect using the Silhouette Coefficient. The quality aspects in each cluster also provide the same order for these two models. Although quite different from K-means and Ward, the results of the Complete method on the quality aspect of each cluster provide a color sequence that corresponds to the level of corrosion in the samples used. Thus, it can be said that this result is similar to the original image. Based on the results of cluster analysis, determining the best method can be divided into several things, such as, K-

Means is the best method (based on the number of clusters produced and model quality), whereas Complete is the best (based on its characteristics and cluster quality).

REFERENCES

[1] H. K. D. H. Bhadeshia, “Preface to the fourth edition,” *Elsevier EBooks*, 2017, doi: 10.1016/b978-0-08-100270-4.00023-8.

[2] G. S. Da Fonseca, P. M. De Oliveira, M. G. Diniz, D. V. Bubnoff, and J. A. De Castro, “SIGMA phase in Superduplex Stainless Steel: formation, kinetics and microstructural path,” *Mater. Res.-Ibero-Am. J. Mater.*, vol. 20, no. 1, pp. 249–255, Jan. 2017, doi: 10.1590/1980-5373-mr-2016-0436.

[3] Y. S. Afrianti, H. Ardy, U. S. Pasaribu, and F. D. E. Latief, “Identification of Inhomogeneous Temperature on Stainless Steel using Statistical Analysis,” *J. Phys.*, vol. 2084, no. 1, p. 012008, Nov. 2021, doi: 10.1088/1742-6596/2084/1/012008.

[4] N. N. Almanza-Ortega et al., “Corrosion analysis through an adaptive preprocessing strategy using the K-Means algorithm,” *Procedia Comput. Sci.*, vol. 219, pp. 586–595, Jan. 2023, doi: doi: 10.1016/j.procs.2023.01.327.

[5] T. Gibbons, G. Pierce, K. Worden, and I. Antoniadou, “A Gaussian mixture model for automated corrosion detection in remanufacturing,” in *A Gaussian mixture model for automated corrosion detection in remanufacturing*, in Advances in Transdisciplinary Engineering, vol. 8. University of Skövde, Sweden: IOP Press, Sep. 2018, pp. 63–68. doi: https://doi.org/10.3233/978-1-61499-902-7-63.

[6] M. Darwis, L. H. Hasibuan, M. B. Firmansyah, N. Ahady, and R. Tiaharyadini, “Implementation of K-Means clustering algorithm in mapping the groups of graduated or dropped-out students in the Management Department of the National University,” *JISA J. Inform. Dan Sains*, vol. 4, no. 1, pp. 1–9, Jun. 2021, doi: 10.31326/jisa.v4i1.848.

[7] I. B. G. Sarasvananda, R. Wardoyo, and A. K. Sari, “The K-Means clustering algorithm with semantic similarity to estimate the cost of hospitalization,” *IJCCS*, vol. 13, no. 4, p. 313, Oct. 2019, doi: 10.22146/ijccs.45093.

[8] V. Nellie, V. C. Mawardi, and N. J. Perdana, “IMPLEMENTASI METODE AGGLOMERATIVE HIERARCHICAL CLUSTERING UNTUK SISTEM REKOMENDASI FILM,” *E-J. Ilmu Komput. Dan Sist. Inf.*, vol. 11, no. 1, Jun. 2023, doi: 10.24912/jiksi.v11i1.24070.

[9] R. Kusumastuti, E. Bayunanda, A. M. Rifa’i, M. R. G. Asgar, F. I. Ilmawati, and K. Kusrini, “Clustering titik panas menggunakan algoritma Agglomerative Hierarchical Clustering (AHC),” *Cogito Smart J.*, vol. 8, no. 2, pp. 501–513, Dec. 2022, doi: 10.31154/cogito.v8i2.438.501-513.

[10] R. E. Van Ruitenbeek, G. Koole, and S. Bhulai, “A hierarchical agglomerative clustering for product sales forecasting,” *Decis. Anal. J.*, vol. 8, p. 100318, Sep. 2023, doi: 10.1016/j.dajour.2023.100318.

[11] S. Sinharay, “An overview of statistics in education,” *Elsevier EBooks*, pp. 1–11, 2010, doi: 10.1016/b978-0-08-044894-7.01719-x.

[12] “Analisis Cluster,” Universitas pendidikan Ganesha. Accessed: Aug. 24, 2023. [Online]. Available: https://cdn.undiksha.ac.id/wp-content/uploads/sites/10/2019/06/19222821/analisis-kluster.pdf

[13] K. M. A. Patel and P. Thakral, “The best clustering algorithms in data mining,” presented at the 2016 International Conference on Communication and Signal Processing (ICCSP), Melmaruvathur, India, pp. 2042–2046. doi: 10.1109/ICCSP.2016.7754534.

[14] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and H. Jia, “K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data,” *Inf. Sci.*, vol. 622, pp. 178–210, Apr. 2023, doi: 10.1016/j.ins.2022.11.139.

[15] J. Pérez-Ortega, N. N. Almanza-Ortega, A. Vega-Villalobos, R. Pazos-Rangel, C. Zavala-Díaz, and A. M. Rebollar, “The K-Means algorithm evolution,” *IntechOpen EBooks*, 2020, doi: 10.5772/intechopen.85447.

[16] Google Machine Learning Courses, “k-Means Advantages and Disadvantages,” Google Machine Learning.

[17] N. Xu, R. B. Finkelman, S. Dai, C. Xu, and M. Peng, “Average Linkage Hierarchical Clustering Algorithm for Determining the Relationships between Elements in Coal,” *ACS Omega*, vol. 6, no. 9, pp. 6206–6217, Feb. 2021, doi: 10.1021/acsomega.0c05758.

[18] P. Praveen, M. Kumar, M. A. Shaik, R. Ravikumar, and R. Kiran, “The comparative study on agglomerative hierarchical clustering using numerical data,” *IOP Conf. Ser.*, vol. 981, no. 2, p. 022071, Dec. 2020, doi: 10.1088/1757-899x/981/2/022071.

[19] L. R. Emmendorfer and A. M. P. Canuto, “A generalized average linkage criterion for Hierarchical Agglomerative Clustering,” *Appl. Soft Comput.*, vol. 100, p. 106990, Mar. 2021, doi: 10.1016/j.asoc.2020.106990.

[20] S. T. Ahmed, S. S. Kumar, B. Anusha, P. Bhumika, M. Gunashree, and B. Ishwarya, “A Generalized Study on Data Mining and Clustering Algorithms,” *Springer EBooks*, pp. 1121–1129, Jan. 2020, doi: 10.1007/978-3-030-41862-5_114.

[21] K. K. Mohbey and G. S. Thakur, “An experimental survey on single linkage clustering,” *Int. J. Comput. Appl.*, vol. 76, no. 17, pp. 6–11, Aug. 2013, doi: 10.5120/13337-0327.

[22] Y. Rani and H. Rohil, “A Study of Hierarchical Clustering Algorithm,” *Int. J. Inf. Comput. Technol.*, vol. 3, no. 11, pp. 1225–1232, 2013.

[23] C. Shalizi, “Distances between Clustering, Hierarchical Clustering.” Carnegie Mellon University, Sep. 14, 2009. [Online]. Available: https://www.stat.cmu.edu/~cshalizi/350/lectures/08/lecture-08.pdf

[24] T. Li, A. Rezaeipanah, and E. M. T. E. Din, “An ensemble agglomerative hierarchical clustering algorithm based on clusters clustering technique and the novel similarity measurement,” *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 6, pp. 3828–3842, Jun. 2022, doi: doi: 10.1016/j.jksuci.2022.04.010.

[25] Vijaya, S. Sharma, and N. Batra, “Comparative Study of Single Linkage, Complete Linkage, and Ward Method of Agglomerative Clustering,” presented at the 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, 2019, pp. 568–573. doi: 10.1109/COMITCon.2019.8862232.

[26] J. Gao, “Clustering Lecture 3: Hierarchical Methods.”

[27] S. Patel, S. Sihmar, and A. Jatain, “A study of hierarchical clustering algorithms,” presented at the International Conference on Computing for Sustainable Global Development, Mar. 2015, pp. 537–541. [Online]. Available: http://ieeexplore.ieee.org/document/7100308.