Anime Segmentation Based on User Preferences: Applying Clustering to Identify Groups of Anime with Similar Genres, Themes, and Popularity

Riswan E. Tarigan[1,*], Yoana Sonia Wijaya[2]

[1,2]*Department of Information Systems, Pelita Harapan University, Jalan M.H. Thamrin Boulevard No.1100 Lippo Karawaci, Tangerang, 15811, Indonesia*

(Received: September 15, 2024; Revised: November 10, 2024; Accepted: January 5, 2025; Available online: April 1, 2025)

Abstract

The anime industry has experienced significant growth, with an increasing focus on user preferences for content discovery and engagement. This study applies clustering techniques, specifically K-means, to segment anime based on user preferences, genres, themes, and popularity. By analyzing a comprehensive dataset containing attributes such as user ratings, popularity, genres, and themes, the research identifies distinct groups of anime that align with varying viewer tastes. The clustering results reveal that anime can be categorized into several groups, including highly popular but critically less-acclaimed titles, well-regarded but moderately popular anime, and niche, critically acclaimed series that appeal to smaller but dedicated audiences. This segmentation allows streaming platforms to offer more personalized recommendations, enhancing user experience and engagement by matching viewers with content that best fits their preferences. Although clustering techniques provide valuable insights into anime content, the study acknowledges certain limitations, such as overlap between clusters, indicating that some anime may not fit perfectly into a single category. This highlights the need for further improvements in segmentation accuracy. The study suggests exploring hybrid clustering methods, combining K-means with other techniques, and integrating demographic data, such as age, gender, and geographic location, to refine recommendations. Overall, the application of clustering algorithms to better understand user preferences in anime offers a promising approach to developing more effective and personalized recommendation systems. This can ultimately improve user satisfaction and engagement in the rapidly growing and competitive anime streaming market.

Keywords: Anime segmentation, User preferences, Clustering, K-means, Personalization

1. Introduction

The anime industry has experienced remarkable growth in recent years, fueled by the increasing importance of user preferences in content discovery and audience engagement. Audiences are no longer passive consumers of content; they are actively shaping the creation of anime by influencing key decisions about the themes, genres, and storylines that are prioritized in production. This dynamic is especially evident in emerging markets like China and Nigeria, where the anime industry is evolving to cater to local tastes and cultural sensibilities [1], [2]. The growing influence of audiences is further amplified by the rise of video streaming platforms, which have democratized access to anime content. These platforms not only provide access to a diverse range of anime but also offer personalized experiences, enabling viewers to interact with the content they consume in ways that were previously not possible [3].

Streaming services have revolutionized the way anime is consumed and discovered, offering an unprecedented level of convenience, accessibility, and customization. The shift from traditional broadcasting to digital streaming platforms has expanded the reach of anime to a global audience, leading to a more diverse and dynamic viewer base. This transformation has made understanding user preferences more important than ever, as creators and producers are now tasked with tailoring content to the specific tastes and viewing habits of their audiences. Anime, with its broad appeal, plays a crucial role in intercultural communication, serving as a medium through which diverse cultural narratives are expressed and shared across national and ethnic boundaries. The global nature of anime has turned it into an important cultural export, and understanding the preferences of international audiences is essential for maintaining its relevance

***Corresponding author: Riswan E. Tarigan (riswan.tarigan@uph.edu)**
DOI: https://doi.org/10.47738/ijaim.v5i1.99

and impact [4]. As the anime industry expands, segmenting anime content by attributes such as genres, themes, and popularity becomes increasingly essential for improving user recommendations. The sheer volume of anime content available today makes it challenging for viewers to easily discover new titles that align with their specific interests. This content overload often leads to decision fatigue, where users become overwhelmed by the abundance of choices and struggle to identify content that matches their preferences. By using advanced data analytics techniques, particularly big data and machine learning, it is possible to segment anime audiences with greater precision, ensuring that users are presented with recommendations that are more likely to resonate with them. Big data analytics provides a way to process and analyze massive datasets, allowing platforms to create tailored experiences that enhance viewer satisfaction and foster long-term engagement [5], [6].

Machine learning, specifically clustering techniques, offers a powerful solution to the challenge of audience segmentation. Clustering algorithms, such as K-means, hierarchical clustering, and density-based clustering, enable anime content to be grouped into categories based on shared characteristics. This process allows anime to be organized by factors like genre, theme, score, and popularity, making it easier for viewers to find content that aligns with their tastes. For example, anime classified under genres like action, drama, or sci-fi can be grouped together, making it easier for users who prefer these genres to discover relevant titles. This segmentation enhances the accuracy of content recommendations and improves the user experience by providing more personalized suggestions [7], [8].

Furthermore, the application of clustering algorithms can help producers and streaming platforms identify distinct groups within the audience based on their viewing habits, preferences, and demographic characteristics. By analyzing clusters of users with similar preferences, platforms can create more effective marketing strategies, offer more targeted recommendations, and even influence the direction of future anime production. For example, if a cluster of viewers shows a strong preference for anime with psychological themes or fantasy genres, producers may prioritize creating content that caters to this audience segment. This tailored approach to content creation not only improves the user experience but also fosters a deeper connection between viewers and the anime they consume [5], [9].

Clustering techniques can also be used to enhance the discovery process for users, helping them find anime that suits their preferences more efficiently. Traditional recommendation systems, which rely on user ratings or genre tags, may fail to capture the full range of user preferences, especially when dealing with more nuanced tastes. Advanced clustering methods, such as K-means clustering, address this by grouping similar anime based on multiple attributes, allowing for a more comprehensive understanding of what different viewers enjoy. This is particularly important in a dynamic and fast-evolving industry like anime, where trends and viewer behaviors can change rapidly. By continuously adapting to these shifts, clustering methods help maintain the relevance and engagement of recommendations, ensuring that users are always presented with content that matches their current interests [10], [11].

The importance of accurate audience segmentation cannot be overstated. Producers and streaming platforms face significant challenges in identifying the specific needs and preferences of diverse viewer segments. Given the complexity and diversity of user behavior, it is essential for platforms to develop robust methodologies that can categorize users based on both psychographic and demographic characteristics. This requires a deep understanding of not just what viewers watch but also why they watch it. Clustering techniques can reveal the underlying patterns in consumer behavior, helping platforms understand the motivations and preferences that drive viewership. By accurately segmenting the audience, producers can better align their content offerings with the demands of specific groups, ensuring that the right content reaches the right viewers [12].

Big data analytics and machine learning also play an important role in identifying and understanding audience motivations. By analyzing large datasets of user interactions, viewing histories, and preferences, platforms can gain insights into what drives viewer engagement and satisfaction. These insights can be used to tailor content recommendations and marketing strategies that are more likely to resonate with different audience segments. However, achieving high accuracy in segmentation remains a challenge due to the complexity and dynamic nature of anime consumption. With new genres, themes, and viewing trends constantly emerging, segmentation models must be flexible and adaptable in order to remain effective [5], [13].

The application of clustering algorithms such as K-means provides a systematic approach to segmenting anime content based on shared attributes, such as genre, theme, score, and popularity. This segmentation allows streaming platforms

to curate content recommendations that are more aligned with the specific preferences of different audience segments. By grouping similar anime together, platforms can offer more relevant suggestions, enhance the viewer experience and increase user satisfaction. Research has shown that clustering algorithms like K-means can identify distinct patterns in user preferences, leading to more accurate and personalized recommendations [14]. This process ultimately contributes to the growth and sustainability of the anime industry by ensuring that content production and marketing strategies align with audience demands.

In conclusion, the ability to segment anime content and audience preferences using clustering algorithms is crucial for improving the viewer experience and driving engagement in the anime industry. As streaming platforms continue to expand and evolve, understanding the preferences and behaviors of their audiences will be essential for maintaining relevance in an increasingly competitive market. Clustering techniques like K-means offer a practical and effective way to categorize anime based on shared characteristics, leading to more personalized content recommendations and better-targeted marketing strategies. By leveraging the power of machine learning and big data analytics, the anime industry can continue to thrive by meeting the diverse and dynamic preferences of its global audience.

2. Literature Review

One significant study by Ferreira et al. [9] addressed the automatic classification of anime by proposing a unique clustering methodology that uses Robotic Process Automation (RPA). This approach focuses on categorizing cultural products, like anime, based on audience feedback and public data, tailoring content to specific demographic segments. The study emphasizes how advanced clustering methods can bridge the gap between anime creators and viewers, ensuring that content is better aligned with the preferences of different audience groups. By using such methods, the anime industry can enhance user engagement by offering more personalized and relevant content.

Another relevant contribution comes from González et al. [15], who explored clustering techniques for genre classification in media. Using K-means clustering and other unsupervised learning methods, they analyzed user interactions and preferences, extracting notable features from viewer data. This study highlighted how accurate genre identification can significantly improve recommendation systems, making content discovery easier and more satisfying for users. Similarly, Yin et al. [16] applied demographic clustering to optimize movie recommendations, grouping users by age, gender, and viewing habits. Their findings suggest that demographic clustering can also benefit anime by providing insights into viewer preferences across different groups, ultimately enabling more tailored content experiences. Together, these studies underscore the importance of clustering algorithms in refining recommendation systems, personalizing content delivery, and enhancing overall user satisfaction in the anime industry.

K-means Clustering is one of the most widely used clustering algorithms due to its simplicity and efficiency, particularly when dealing with large datasets. The algorithm works by partitioning data into KKK distinct clusters based on feature similarity, where each data point belongs to the cluster with the nearest mean value. K-means has been successfully employed in various studies to categorize anime based on viewer preferences, genres, and ratings, making it an excellent tool for personalized recommendations [17], [18], [19]. Its ability to quickly segment user behavior patterns, especially in streaming platforms, makes it a valuable tool for content creators and streaming services that aim to enhance user interaction. For example, Liu [17] demonstrated that K-means could effectively segment anime streaming data to improve content discoverability, leading to better recommendations for viewers by grouping similar anime titles together. This clustering method allows for improved content discovery, making it easier for users to find titles that match their interests.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is another powerful clustering algorithm that is particularly useful for identifying clusters of varying shapes and sizes. Unlike K-means, which requires the number of clusters to be predefined, DBSCAN defines clusters based on the density of points within the feature space. This feature makes DBSCAN highly effective in handling noisy data and outliers, which are common in user feedback and viewing history data related to anime [20]. The algorithm's flexibility allows it to detect clusters in datasets where viewer preferences may not follow a simple, uniform pattern, which is often the case in the anime industry. For example, anime viewers may have highly specific preferences that do not neatly fit into predefined clusters. DBSCAN addresses this issue by clustering data based on density, which can capture more complex patterns in user behavior.

Moreover, adaptations of DBSCAN, such as HDBSCAN and OPTICS, have been developed to improve the algorithm's performance when dealing with varying density structures [21]. These improvements make DBSCAN a robust tool for analyzing the diverse preferences of anime viewers, ensuring that more accurate and meaningful clusters are formed.

Another useful clustering technique is Hierarchical Clustering, which builds a hierarchy of clusters that can be visualized in a dendrogram. This method can follow either a bottom-up (agglomerative) or top-down (divisive) approach, offering flexibility in how clusters are formed. Hierarchical clustering is particularly useful in anime segmentation because it allows for the exploration of relationships between anime titles at different levels of granularity. For instance, hierarchical clustering can identify both broad genre clusters (such as action or fantasy) and more specific sub-genres or themes (e.g., supernatural action or magical realism) [22]. This approach enables streaming platforms to provide nuanced recommendations based on the user's engagement with specific themes or genres, thus improving the personalization of content. Hierarchical clustering is also valuable for understanding how different demographic groups engage with various types of anime. By grouping users based on factors like age or gender, hierarchical clustering can reveal insights into the preferences of specific audience segments, facilitating more targeted recommendations [22].

K-means, DBSCAN, and hierarchical clustering, each offer distinct advantages depending on the type of data and the goals of the clustering task. K-means is efficient for large, well-structured datasets where the number of clusters is predefined, making it suitable for genre and preference-based segmentation. DBSCAN is ideal for datasets with noise and outliers, as it can identify arbitrary-shaped clusters and adapt to varying densities, making it effective in handling complex viewing patterns. Hierarchical clustering, with its ability to create a hierarchy of clusters, offers valuable insights at different levels of granularity, enabling deeper analysis of user preferences and demographic segmentation.

The combination of these methods can be particularly beneficial for the anime industry. K-means offers computational efficiency and ease of implementation, while DBSCAN and hierarchical clustering provide more flexibility in handling noisy, complex, or varied data. These methods can be combined or adapted into hybrid models to enhance the precision of anime segmentation and improve recommendation accuracy. For instance, using DBSCAN for initial segmentation and K-means for refinement, or applying hierarchical clustering to explore user preferences at multiple levels, can result in a more comprehensive and effective clustering process.

In conclusion, clustering algorithms like K-means, DBSCAN, and hierarchical clustering are invaluable tools in the anime industry, offering different strengths for segmenting content and enhancing recommendation systems. These methods allow streaming platforms to offer more personalized content by grouping anime based on shared characteristics such as genre, theme, and viewer preferences. As the anime industry continues to grow and diversify, these clustering techniques can be further refined and integrated to meet the evolving demands of viewers. Future research should explore the potential for hybrid clustering models that combine the strengths of each method, providing even more accurate and relevant content recommendations. By leveraging advanced clustering techniques, the anime industry can enhance user satisfaction, improve engagement, and ultimately ensure that content is more closely aligned with the preferences of diverse audience segments.

3. Methodology

The following diagram illustrates the steps involved in the data analysis process using the K-Means algorithm, as shown in figure 1.

 ISSN 2776-8007

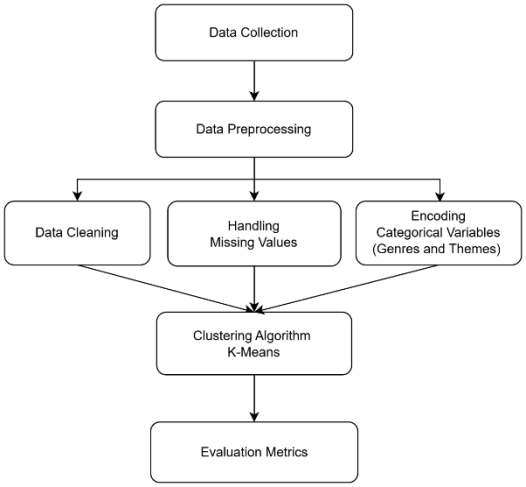

Figure 1. Research Methodology

3.1. Data Collection

The dataset used for clustering consists of detailed information about various anime titles. Each entry in the dataset represents a single anime and includes a variety of features that are relevant for the clustering analysis. These features are used to identify patterns and group similar anime based on user preferences and attributes. The key attributes in the dataset include those listed in Table 1. These attributes provide both quantitative data (e.g., score, popularity, favorites) and qualitative data (e.g., genres, themes, demographics), which are crucial for performing clustering and segmenting anime based on user preferences and content characteristics.

Table 1. Key Anime Attributes

Attribute	Description
anime_id	Unique identifier for each anime.
anime_url	URL pointing to the anime's page on MyAnimeList or a similar platform.
image_url	URL of the image representing the anime.
name	The official title of the anime.
english_name	The English title of the anime (if available).
japanese_names	The original Japanese title of the anime.
score	The average user rating of the anime, indicating its popularity and general quality.
genres	List of genres associated with the anime (e.g., Action, Drama, Fantasy).
themes	List of themes or sub-genres related to the anime (e.g., Psychological, Time Travel).
demographics	Target demographic for the anime (e.g., Shounen, Seinen, Shoujo).
producers	The production companies behind the anime.
studios	The animation studio that created the anime.
source	The origin of the anime (e.g., Manga, Light Novel, Visual Novel).
duration	Duration of each episode or the total episode time (e.g., 24 min per episode).
rating	The content rating (e.g., PG-13, R-17+).
rank	The ranking of the anime based on user ratings and popularity.
popularity	The popularity ranking based on the number of views and followers.
favorites	Number of users who have added the anime to their favorites list.
scored_by	The number of users who have rated the anime.
members	The total number of users who have added the anime to their list or followed it.

These features provide a rich set of information that is essential for performing clustering analysis. The primary goal of clustering is to group anime into categories that share similar characteristics, such as similar genres, themes, scores, or popularity. By analyzing these attributes, the clustering algorithm can uncover hidden patterns and relationships that can be used to better understand anime preferences and recommend similar titles to users.

3.2.Data Preprocessing

Data preprocessing is a crucial step in preparing the dataset for clustering analysis. The first step in preprocessing involves cleaning the data by removing irrelevant columns that do not contribute significantly to the clustering process. For example, columns such as anime_id, anime_url, and image_url are discarded, as they do not contain useful information for clustering. The focus is placed on relevant features such as score, popularity, favorites, members, genres, and themes, which are essential for identifying patterns and grouping anime based on user preferences.

Next, we address the issue of missing values. For quantitative features like score, popularity, favorites, and members, any rows containing missing data are removed. This ensures that the clustering analysis is based on complete and accurate data. For categorical variables such as genres and themes, missing values are replaced with empty strings ('). This step ensures that the encoding process proceeds smoothly without any issues caused by null values.

To prepare the categorical variables for clustering, we use CountVectorizer to encode the genres and themes columns, which are text-based features. The CountVectorizer transforms these textual descriptions into numerical vectors, where each genre or theme is represented as a separate feature. For example, the genres "Action", "Drama", and "Fantasy" would be represented by separate binary columns, with a 1 indicating the presence of that genre in the anime. This encoding method enables the clustering algorithm to work with categorical data effectively.

After encoding the categorical variables, the dataset is normalized to ensure that all features have comparable scales. Since quantitative features like score and popularity may vary significantly in their range, we apply StandardScaler to standardize the dataset. This scaling technique adjusts the features to have a mean of 0 and a standard deviation of 1, ensuring that each feature contributes equally to the clustering process.

3.3. Clustering Algorithm

K-means clustering works by minimizing the variance within clusters, ensuring that the items within a cluster are as similar as possible while being different from those in other clusters. The algorithm works iteratively by assigning data points to the nearest cluster centroid and recalculating the centroids until convergence. The formula used to calculate the distance between a data point x_i and a cluster centroid C_k is typically the Euclidean distance, given by:

$$d(x_i, C_k) = \sqrt{\sum_{j=1}^{n} (x_{ij} - c_{kj})^2} \quad (1)$$

x_i is the data point, C_k is the centroid of the cluster, n is the number of features, $x_{ij}$and C_{kj} are the j -th features of the data point and centroid, respectively. K-means was selected for this task due to its simplicity, efficiency, and widespread use in clustering applications. It is computationally inexpensive compared to other clustering algorithms, making it suitable for large datasets like the one used in this study. Additionally, K-means is a well-understood algorithm, and its ability to cluster mixed data types combining both numerical (e.g., score, popularity) and categorical (e.g., genres, themes) features makes it a versatile choice for this analysis.

One of the key steps in applying K-means clustering is determining the optimal number of clusters, denoted as k. Several methods can be used to select the best value for k, two of the most common being the Elbow Method and the Silhouette Score. The Elbow Method involves running the K-means algorithm for different values of k and calculating the within-cluster sum of squares (WCSS) for each value. WCSS measures the compactness of the clusters, and the formula for WCSS is:

$$WCSS = \sum_{k=1}^{K} \sum_{x_i \in C_k} (x_i - c_k)^2 \quad (2)$$

C_k is the set of data points in cluster k, c_k is the centroid of cluster k, x_i is a data point in cluster k. The Elbow Method involves plotting WCSS against the number of clusters k. The optimal k is typically chosen where the WCSS curve shows a sharp "elbow," signifying that adding more clusters does not result in a significant reduction in WCSS.

The Silhouette Score provides a measure of how well each data point fits into its assigned cluster, compared to how well it fits into other clusters. The silhouette score $s(x_i)$ for a point x_i is given by:

$$s(x_i) = \frac{b(x_i) - a(x_i)}{\max(a(x_i), b(x_i))} \quad (3)$$

$a(x_i)$ is the average distance between x_i and all other points within the same cluster; $b(x_i)$ is the minimum average distance from x_i to all points in any other cluster. A higher silhouette score indicates that the data points are well-clustered. By calculating the silhouette score for different values of k, the optimal number of clusters can be selected as the one that maximizes the silhouette score.

Once the optimal value of k is determined, K-means is applied to the dataset. The algorithm begins by randomly selecting k centroids, which serve as the initial centers of the clusters. In the assignment step, each data point x_i is assigned to the nearest centroid based on the Euclidean distance formula, meaning the algorithm calculates the distance from each point to all centroids and assigns the point to the closest centroid. After all data points have been assigned, the update step is performed, where the centroids are recalculated by taking the mean of all points assigned to each cluster. This process of initialization, assignment, and update is repeated iteratively until the centroids converge, meaning that the centroids no longer change significantly, indicating that the algorithm has successfully partitioned the dataset into stable clusters.

$$c_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i \quad (4)$$

C_k is the set of data points in cluster k, c_k is the new centroid of cluster k. The result is a set of k clusters, each containing data points that are similar to one another based on the features provided.

4. Results and Discussion

4.1. Result

The clustering analysis identifies distinct groups of anime based on shared characteristics such as popularity, user ratings, and thematic elements. Each cluster represents a set of anime that exhibit similar patterns in these attributes, revealing how viewers' preferences differ across various types of anime. By segmenting the dataset into multiple clusters, we can understand the distribution of anime titles and their alignment with audience tastes. Table 2 show the breakdown of each cluster, which provides insights into how anime titles are distributed across these characteristics.

Table 2. Cluster Composition of Anime Based on User Preferences

Cluster	Key Characteristics	Common Genres/Themes
Cluster 0	High popularity, low to moderate scores. This anime is widely popular but have varying scores, indicating broad appeal with less critical acclaim.	Genres like shounen, action, mainstream comedy, slice-of-life. This anime typically has a wide reach but are not necessarily praised for their quality or storytelling depth.
Cluster 1	Moderate popularity, high scores. This anime is well-regarded for their quality and storytelling but do not achieve the massive popularity of Cluster 0.	Genres like psychological thriller, fantasy, drama. This anime is appreciated for their depth and production quality, appealing to audiences who prioritize quality over mass appeal.
Cluster 2	Low popularity, low scores. This anime has low popularity and lower ratings, possibly due to niche or polarizing themes.	Genres like experimental anime, avant-garde, niche genres like horror or mecha. These titles appeal to a specific audience but lack broader popularity.
Cluster 3	Low popularity, high scores. This anime is critically acclaimed but have not gained widespread recognition or commercial success.	Genres like historical drama, slice-of-life, or niche sci-fi/fantasy. This anime may be "hidden gems," well-regarded for their storytelling and quality, but not widely recognized.

 ISSN 2776-8007

Cluster 4	Moderate popularity, moderate scores. This anime strikes a balance between popularity and ratings, appealing to a general audience without standing out dramatically in either direction.	Genres like romance, comedy, slice-of-life. This anime is loved by specific demographic groups but do not dominate in terms of overall popularity.

The distribution of anime across the clusters, as described in table 2, is visually represented in figure 2, which demonstrates how the different clusters are distinctively separated based on key characteristics such as popularity and ratings. Figure 2 shows a clear separation of the clusters, with Cluster 0 displaying a broad range of anime titles, reflecting their high popularity and varying scores. This visual representation reinforces the notion that Cluster 0 contains widely popular anime that may lack critical acclaim or depth, aligning with the description in table 2. Similarly, Cluster 2 is visually distinct as it groups anime with lower popularity and ratings, indicating its niche or underperforming nature, which corresponds to the analysis in the table. The visualization highlights the clustering patterns driven by the underlying characteristics of popularity and user ratings.

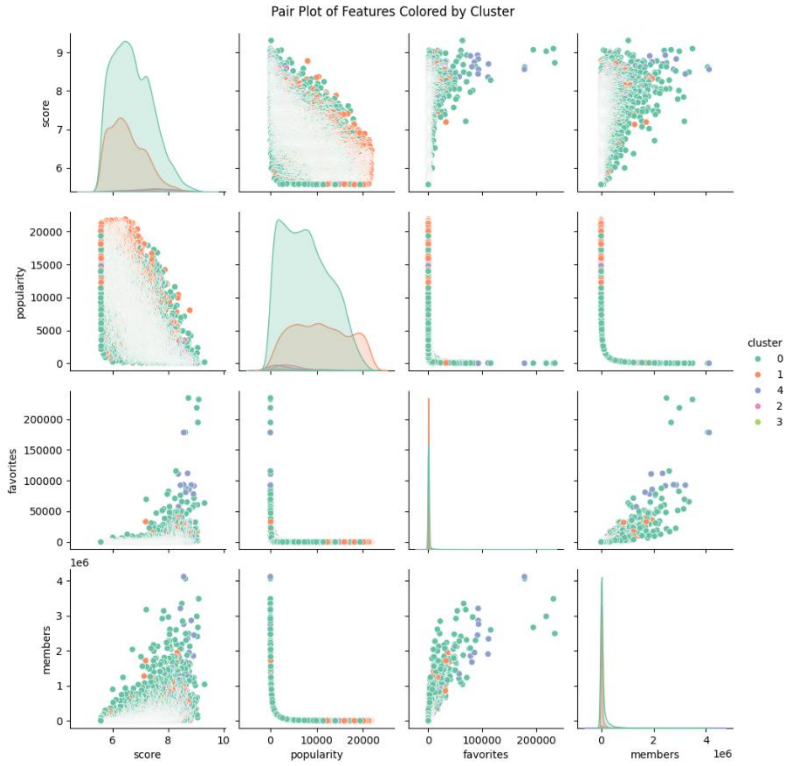

Figure 2. Pair Plot of Features Colored by Cluster

Furthermore, figure 3 further supports the cluster breakdown in table 2 by illustrating the cluster distribution. This figure shows the number of anime titles within each cluster, highlighting the uneven distribution across the groups. As shown in the figure, Cluster 0 contains the largest number of anime, which aligns with its characterization in table 2 as the cluster with broad appeal and the highest number of popular titles. On the other hand, clusters like Cluster 2 and Cluster 3 contain fewer anime, reflecting their more niche or critically acclaimed yet less commercially successful nature. Figure 3 reinforces these findings, providing a clear visual confirmation of how anime is distributed across the different clusters based on popularity and ratings.

 ISSN 2776-8007

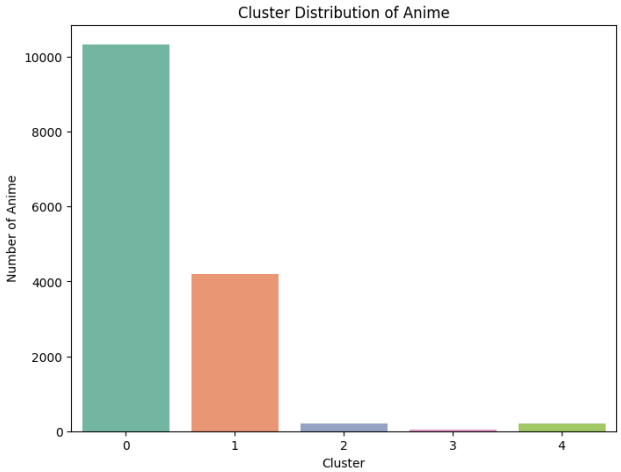

Figure 3. Cluster Distribution of Anime

Figure 4 shows the separation of anime into distinct clusters based on their features, such as score, popularity, and other attributes. In this visualization, the black X markers represent the centroids of the clusters, which indicate the central point of each group. The clusters are clearly separated, with minimal overlap between them. This confirms that the clustering algorithm has successfully differentiated between various types of anime. The separation suggests that the features used for clustering, such as popularity and ratings, are effective in distinguishing between the different groups. The centroids provide a clear view of the main characteristics of each cluster, offering insight into how the anime are grouped based on their similarities.

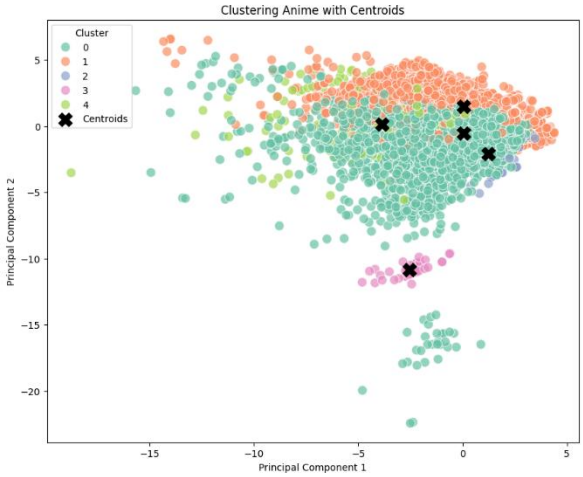

Figure 4. Clustering Anime with Centroids

Next, after observing how the clusters are separated based on key features like score and popularity in figure 4, figure 5 expands this perspective by focusing on how user preferences, such as genre, theme, and rating, serve as the primary differentiating factors in the clustering process. Figure 5 provides a visualization of how anime titles are grouped based on characteristics or attributes that reflect user preferences. This plot is similar to figure 4, but here, the focus is specifically on how these user-driven factors shape the clustering process.

The spread of the clusters across the principal components in the plot suggests that there are significant differences in how various user groups engage with anime. These principal components likely represent key features in the dataset, such as score, popularity, and genre. The fact that the clusters are spread across these components indicates that user preferences are diverse, with different groups of viewers gravitating toward anime based on their personal tastes. For instance, one cluster may be dominated by users who prefer high-action or shounen anime, while another cluster might reflect viewers more inclined toward niche genres like psychological thrillers or fantasy.

 ISSN 2776-8007

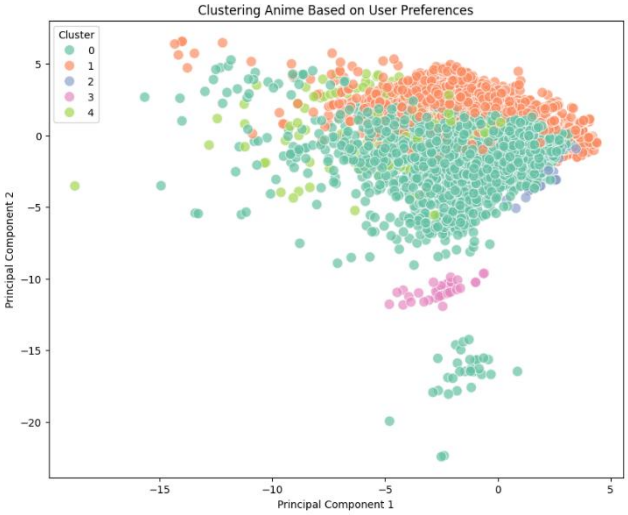

Figure 5. Clustering Anime Based on User Preferences

This visualization reinforces the concept of personalized recommendations, where the clusters represent groups of anime that are likely to appeal to different subsets of viewers. By grouping similar anime together based on the shared preferences of their audiences, streaming platforms or recommendation systems can offer more relevant suggestions to users, improving the overall viewing experience. The clear separation between clusters suggests that the algorithm has successfully segmented the anime in a way that mirrors actual user behavior, allowing for a deeper understanding of the specific preferences that drive engagement with different types of anime. This ability to create clusters based on user preferences is a valuable tool for enhancing content discovery and ensuring that viewers are presented with anime that closely match their tastes.

However, while the clustering method has successfully differentiated anime based on preferences, figure 6 provides further insights into the quality of the clustering by evaluating the silhouette score, which measures how well each anime fits within its assigned cluster compared to other clusters. The average silhouette score of 0.02 indicates that while the clusters are generally distinct, there is still room for improvement. Some anime appears to be positioned near the borders between clusters, suggesting that the separation is not perfect. Although the score is low, the silhouette plot still provides a valuable visual representation of how well the clustering method is working overall. It indicates that there is some overlap between clusters, which may need to be addressed for better accuracy and more refined clustering.

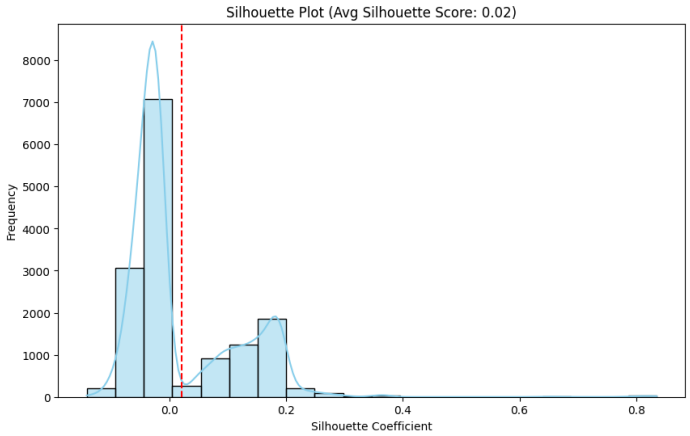

Figure 6. Silhouette Plot

Building on the insights from figure 6, figure 7 provides a deeper understanding of the relationships between key features like score, popularity, favorites, and members. The heatmap reveals that score and popularity are negatively correlated, meaning that as the popularity of an anime increases, its score tends to decrease. This suggests that anime with large audiences might not always receive the highest ratings, likely due to the broad appeal of mainstream genres like action or shounen that attract large followings but may not be as critically acclaimed. In other words, popular

anime might have less depth or storytelling quality, which results in lower user ratings despite their widespread popularity.

The heatmap also shows a strong positive correlation between favorites and members, indicating that anime with a high number of users who have added them to their favorites tend to also have a larger total number of members. This suggests that anime attracting a dedicated fanbase are also able to appeal to a broader audience, likely due to qualities such as strong character development, engaging plotlines, or high-quality animation. The large number of favorites reflects a high level of viewer engagement, where this anime resonates strongly with viewers who are more likely to recommend or return to them repeatedly.

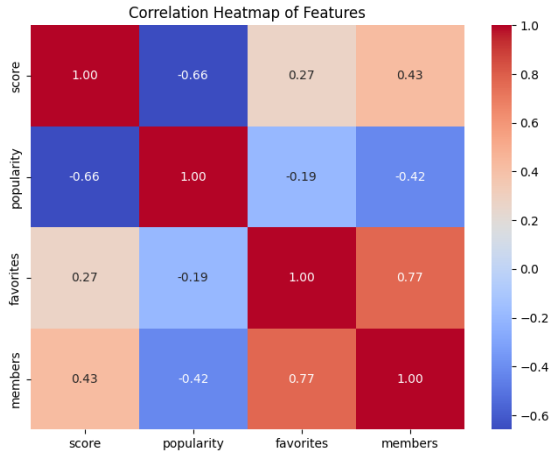

Figure 7. Correlation Heatmap of Features

These correlation insights are crucial for understanding the composition of the clusters identified in the clustering analysis. For instance, Cluster 0, which contains highly popular but lower-scoring anime, aligns with the negative correlation between score and popularity. On the other hand, anime in Cluster 1, which have high scores and moderate popularity, likely benefit from the positive correlation between favorites and members, as these titles appeal to more dedicated and engaged audiences. Figure 7 thus helps explain the clustering patterns by providing a deeper understanding of how different features interact with each other, shedding light on the reasons behind the grouping of anime into distinct clusters.

4.2. Discussion

The clustering analysis identifies distinct groups of anime based on shared characteristics such as popularity, user ratings, and thematic elements. The anime is effectively segmented into groups, with titles that have similar ratings and popularity clustering together. This suggests that popularity and user ratings are crucial factors in distinguishing between different types of anime. The ability to group anime based on these characteristics aligns well with previous research by Ferreira et al. [9], who highlighted the importance of clustering content based on audience preferences to create personalized recommendations. Additionally, González et al. [15] demonstrated that clustering methods like K-means are instrumental in improving genre classification, supporting the notion that clustering is a powerful tool for identifying meaningful patterns in content features.

Further analysis reveals that the clusters also reflect user preferences, with anime grouped according to preferences for genre, theme, and rating. This personalized approach allows for a more refined recommendation system that takes into account individual tastes. The distribution of clusters indicates a significant variation in how different user groups engage with anime, with some clusters dominated by high-action or shounen genres, while others gravitate toward niche genres like psychological thrillers or fantasy. This focus on user preferences extends the work of Yin et al. [16], who applied demographic clustering to optimize movie recommendations by considering specific user tastes. González et al. [15] further reinforced this concept, showing how user interaction and preferences play a pivotal role in clustering and enhancing content classification. Our study builds on these ideas by incorporating detailed user-driven factors, such as genre and theme, which are essential for personalized recommendations in the anime industry.

While user preferences play a significant role in distinguishing clusters, evaluating the effectiveness of the clustering method itself is equally important. Figure 6, the silhouette plot, helps assess the quality of the clustering process by measuring how well each anime fits within its assigned cluster compared to others. The average silhouette score of 0.02 suggests that the clustering is effective, but some overlap between clusters remains. This indicates that while the algorithm is successful in differentiating between groups, there is room for improvement, particularly at the boundaries between clusters. The low silhouette score points to the potential need for more refined clustering methods or adjustments to the algorithm.

This observation is consistent with the work of Liu [17], who pointed out that while K-means is efficient for large datasets, it struggles with complex data, especially when clusters exhibit overlapping features. Similarly, Zhou et al. [20] noted that clustering methods like K-means can sometimes lead to suboptimal separations, particularly when the data is noisy or the number of clusters is not well-defined, underscoring the potential for further algorithmic refinement.

Additionally, the correlation between key features like score, popularity, favorites, and members provide valuable insights. The analysis shows a negative correlation between score and popularity, indicating that highly popular anime tends to have lower ratings. This can be attributed to the broad appeal of mainstream genres such as action or shounen, which attract a large audience but may not always receive critical acclaim. The positive correlation between favorites and members suggests that anime with more favorites also have a larger following, reflecting higher levels of viewer engagement. These findings echo Ferreira et al. [9], who explored how audience feedback and engagement metrics like favorites influence content recommendations. González et al. [15] also emphasized the importance of understanding the relationship between user ratings and engagement for refining recommendation systems, which is clearly demonstrated in our findings. The positive correlation between favorites and members highlights the role of engagement in shaping content delivery and understanding viewer behavior, which is crucial for creating personalized content curation strategies.

This study introduces a novel approach by combining K-means clustering with user preferences, such as genre, theme, and rating, to segment anime into distinct groups. While previous research has primarily focused on clustering based on broad content features like genre or demographic data, this study goes further by incorporating user-driven factors, which provide a more personalized perspective on how anime should be grouped. By clustering anime according to viewer preferences, this research adds an important layer of personalization that previous methods may not have fully addressed. The ability to analyze and group anime based on individual tastes allows for more accurate and effective personalized recommendation systems, offering deeper insights into how user behavior shapes content discovery.

The implications of this research are significant for both the anime industry and streaming platforms. The ability to group anime based on user preferences enables platforms to offer more tailored recommendations, enhancing user satisfaction by suggesting content that aligns closely with individual tastes. Personalized recommendations are crucial in today's streaming environment, where viewers face an overwhelming selection of content. This approach helps prioritize recommendations based on factors such as genre, theme, and rating, rather than relying solely on popularity. As a result, this personalized system can improve content discoverability, increase user engagement, and ultimately drive higher retention rates for platforms.

However, there are limitations to this study that must be addressed. First, the clustering method is heavily reliant on the selected features, such as popularity, rating, and genre, which may not capture the full complexity of viewer preferences. Additionally, the use of K-means clustering introduces challenges, particularly regarding the need to predefine the number of clusters, which may not always result in the most optimal groupings. The silhouette score of 0.02 suggests that some overlap remains between clusters, indicating that the separation between groups is not perfect. Furthermore, this analysis does not account for external factors influencing anime ratings, such as seasonal trends, cultural differences, or regional preferences, all of which could affect the clustering outcomes.

Future research could explore several avenues to refine and expand upon these findings. One promising direction is the use of hybrid clustering models that combine the strengths of different clustering techniques, such as DBSCAN and hierarchical clustering, to better handle data with varying densities and complex relationships. Integrating demographic information such as age, gender, and geographic location could provide even more personalized insights into user preferences, leading to more accurate content recommendations. Additionally, implementing dynamic clustering

methods that adapt over time to account for changing viewer behaviors and seasonal trends could further enhance clustering accuracy, providing real-time adaptation to shifting audience preferences. These advancements would help build more sophisticated recommendation systems capable of delivering highly personalized content that aligns with evolving user tastes.

5. Conclusion

This study explores the importance of anime segmentation based on user preferences, with a focus on clustering anime using algorithms, particularly K-means. By analyzing attributes such as genre, theme, user ratings, and popularity, the research identifies groups of anime with similar characteristics. The clustering results reveal that anime can be categorized into distinct groups, such as Cluster 0, which contains popular anime with lower critical ratings, while Cluster 1 groups anime that are highly regarded for their quality but with lower popularity. The study highlights how understanding user preferences can enhance the user experience by providing more personalized and relevant recommendations.

The use of clustering techniques allows streaming platforms to better understand the diverse preferences of their audience and match them with the types of content they enjoy. For instance, anime with high popularity often has lower ratings, while critically acclaimed but less well-known anime tends to have a smaller but more dedicated following. This study demonstrates that clustering based on genre, theme, and ratings provides deeper insights into user preferences, which can be translated into more accurate recommendations. This is especially relevant in the context of streaming platforms that face the challenge of suggesting content that aligns with the diverse tastes of their users.

However, while clustering algorithms have provided valuable insights, the study also acknowledges certain limitations, particularly the overlap between clusters as indicated by the low silhouette score. This suggests that while clustering is effective in grouping anime based on certain characteristics, there is still room for improvement in the precision of segmenting user preferences. The study suggests exploring hybrid clustering models or combining other techniques to handle more complex and dynamic data, and considering demographic factors like age, gender, and geographic location to create even more personalized recommendations.

6. Declarations

6.1. Author Contributions

Conceptualization: R.E.T., Y.S.W.; Methodology: R.E.T., Y.S.W.; Software: R.E.T.; Validation: Y.S.W.; Formal Analysis: R.E.T.; Investigation: R.E.T.; Resources: Y.S.W.; Data Curation: R.E.T.; Writing – Original Draft Preparation: R.E.T.; Writing – Review and Editing: Y.S.W.; Visualization: R.E.T.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] W. Jiang, "Key Selection Factors Influencing Animation Films From the Perspective of the Audience," *Mathematics*, vol. 12, no. 10, pp. 1–21, 2024, doi: 10.3390/math12101547.

[2] D. O. Fayenuwo and J. Iwuh, "Bracing the Opportunities in the Nigerian Animation Industry: Unlocking the Challenging Phase," *Animation*, vol. 20, no. 1, pp. 41–58, 2025, doi: 10.1177/17468477251324057.

[3] X. An and Y. Li, "Unraveling the Positive Impact of Video Streaming Platforms on Cinematic Storytelling and Their Profound Influence on Individual Audiences," *Commun. Humanit. Res.*, vol. 29, no. 1, pp. 145–150, 2024, doi: 10.54254/2753-7064/29/20230628.

[4] L. Wang and M. A. M. Desa, "Dissemination and Interpretation of Animated Films in Intercultural Communication in China," *Int. J. Relig.*, vol. 5, no. 10, pp. 5419–5428, 2024, doi: 10.61707/mr2bkd32.

[5] L. Yan and L. Zheng, "Using Big Data to Tailor Marketing and Improve User Engagement in the Animation Industry," *J. Cases Inf. Technol.*, vol. 27, no. 1, pp. 1–28, 2025, doi: 10.4018/jcit.374029.

[6] H. Zhao, X. Luo, R. Ma, and L. Xi, "An Extended Regularized K-Means Clustering Approach for High-Dimensional Customer Segmentation With Correlated Variables," *IEEE Access*, vol. 9, no. March, pp. 48405–48412, 2021, doi: 10.1109/access.2021.3067499.

[7] B. Türkmen, "Customer Segmentation With Machine Learning for Online Retail Industry," *Eur. J. Soc. Behav. Sci.*, vol. 31, no. 2, pp. 111–136, 2022, doi: 10.15405/ejsbs.316.

[8] B. J. Jansen, J. Salminen, and S. Jung, "Making Meaningful User Segments From Datasets Using Product Dissemination and Product Impact," *Data Inf. Manag.*, vol. 4, no. 4, pp. 237–249, 2020, doi: 10.2478/dim-2020-0048.

[9] J. C. V Ferreira, T. R. Furtado, R. D. D. Regis, G. R. Diniz, P. Gonçalves, and V. P. da S. C. Tavares, "Anime Clustering for Automatic Classification and Configuration of Demographics," *Cuadernos.Info*, vol. 2023, no. 54, pp. 67–94, 2023, doi: 10.7764/cdi.54.53193.

[10] V. Sampath, S. Sureshkumar, and I. H. Kim, "The Efficacy of Yeast Supplementation on Monogastric Animal Performance—A Short Review," *Life*, vol. 13, no. 10, pp. 1–13, 2023, doi: 10.3390/life13102037.

[11] L. Lv, J. Wang, R. Wu, H. Wang, and I. Lee, "Density Peaks Clustering Based on Geodetic Distance and Dynamic Neighbourhood," *Int. J. Bio-Inspired Comput.*, vol. 17, no. 1, pp. 24–33, 2021, doi: 10.1504/ijbic.2021.113363.

[12] P. S. Durga, J. A. Paulson, and M. Srinivasareddy, "Customer Segmentation Analysis for Improving Sales Using Clustering," *Int. J. Sci. Res. Arch.*, vol. 9, no. 2, pp. 708–715, 2023, doi: 10.30574/ijsra.2023.9.2.0663.

[13] S. M. Miraftabzadeh, C. G. Colombo, M. Longo, and F. Foiadelli, "K-Means and Alternative Clustering Methods in Modern Power Systems," *IEEE Access*, vol. 11, no. October, pp. 119596–119633, 2023, doi: 10.1109/access.2023.3327640.

[14] S. Premkumar, D. Jebaseelan, and K. Annamalai, "Feature Recognition and Machine Learning in Finite Element Models Through a Clustering Algorithm," *Int. J. Simul. Multidiscip. Des. Optim.*, vol. 13, no. 26, pp. 1–9, 2022, doi: 10.1051/smdo/2022011.

[15] F. R. González, M. Torres-Ruiz, G. Rivera-Torruco, L. Chonona-Hernández, and R. Quintero, "A Natural-Language-Processing-Based Method for the Clustering and Analysis of Movie Reviews and Classification by Genre," *Mathematics*, vol. 11, no. 23, pp. 4735, 2023, doi: 10.3390/math11234735.

[16] L. J. Yin, N. Z. M. Safar, H. Kamaludin, N. Abdullah, M. A. M. Yusof, and C. Supriyanto, "Adopting Machine Learning in Demographic Filtering for Movie Recommendation System," *J. Soft Comput. Data Min.*, vol. 4, no. 1, pp. 1–12, 2023, doi: 10.30880/jscdm.2023.04.01.001.

[17] Y. Liu, "Customer Segmentation in User Behavior Analysis: A Comparative Study of Clustering Algorithms," *Highlights Bus. Econ. Manag.*, vol. 21, no. December, pp. 758–764, 2023, doi: 10.54097/hbem.v21i.14758.

[18] A. D. Buchdadi and A. S. M. Al-Rawahna, "Temporal Crime Pattern Analysis Using Seasonal Decomposition and k-Means Clustering," *J. Cyber Law*, vol. 1, no. 1, pp. 65–87, 2025.

[19] I. F. Ashari, R. Banjarnahor, D. R. Farida, S. P. Aisyah, A. P. Dewi, and N. Humaya, "Application of Data Mining With the K-Means Clustering Method and Davies Bouldin Index for Grouping IMDB Movies," *J. Appl. Informatics Comput.*, vol. 6, no. 1, pp. 7–15, 2022, doi: 10.30871/jaic.v6i1.3485.

[20] W. Zhou, L. Wang, X. Han, Y. Wang, Y. Zhang, and Z. Jia, "Adaptive Density Spatial Clustering Method Fusing Chameleon Swarm Algorithm," *Entropy*, vol. 25, no. 5, pp. 1–28, 2023, doi: 10.3390/e25050782.

[21] L. Küchenhoff, P. Lukas, C. M-Zumaran, P. Rothhaar, A. Ruggieri, V. Lohmann, T. Hofer, M. L. Stanifer, S. Boulant, S. R. Talemi, and F. Graw, "Extended Methods for Spatial Cell Classification With DBSCAN-CellX," *Sci. Rep.*, vol. 13, no. 1, pp. 1–11, 2023, doi: 10.1038/s41598-023-45190-4.

[22] V. Ratnasari and A. T. R. Dani, "Mapping the Provincial Food Security Conditions in Indonesia Using Cluster Ensemble-Based Mixed Data Clustering-Robust Clustering With Links (CEBMDC-ROCK)," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 13, no. 2, pp. 611–617, 2023, doi: 10.18517/ijaseit.13.2.16457.