Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14332
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Evaluation of Clustering Algorithms for Identifying Shoe Characteristics Patterns at XYZ Footwear

William Watasendjaja[1)*], Billy Chandra[2)], Ito Wasito[3)]
[1,2,3)]Pradita University, Indonesia
[1)]william.watasendjaja.s2@student.pradita.ac.id, [2)]billy.chandra.putra@student.pradita.ac.id,
[3)]ito.wasito@pradita.ac.id

Submitted :Dec 12, 2024 | **Accepted** : Dec 29, 2024 | **Published** : Jan 12, 2025

Abstract: As the third-largest shoe-exporting country in the world, Indonesia faced a 25% decline in shoe exports in 2023 compared to the year before, both in terms of net weight and sales value. This decline in shoe exports occurred due to the increase of complexity and variety in customer orders to shoe manufacturers. These reasons require shoe manufacturers to enhance their production planning systems to become more efficient and competitive. To address this problem, this study explores the application of clustering algorithms to optimize the production planning process in shoe manufacturing companies. Using a case study at XYZ Footwear, clustering algorithms such as K-Means, Support Vector Clustering (SVC), and Deep Autoencoder were evaluated and compared to find the most effective algorithms in identifying patterns in shoe characteristics, thereby improving shoe manufacturers' production planning process. The datasets consist of the 2024 production season data, categorized into shoe categories, models, and variants, and purchase orders. The result shows that the combination of Deep Autoencoder and K-Means has better performance than just K-Means or Support Vector Clustering (SVC), achieving a silhouette score of 0.4822 and a Davies-Bouldin Index (DBI) of 0.6741. These findings highlight the effectiveness of combining deep learning (Deep Autoencoder) with clustering algorithms (K-Means) in identifying patterns in shoe characteristics.

Keywords: deep autoencoder; k-means clustering; patterns identification; shoe characteristics; support vector clustering

INTRODUCTION

Based on World Footwear data, in 2023 Indonesia ranked as the third-largest shoe-exporting country in the world, after China and Vietnam, with shoe exports exceeding 445 million pairs (3.2% of the global total) (World Footwear, 2024). However, this figure is lower compared to shoe exports in 2022, which exceeded 535 million pairs (3.5% of the global total) (World Footwear, 2023). This decline is also supported by data from Badan Pusat Statistik (BPS) Indonesia, which shows that sports shoe exports in 2023 decreased by about 25% compared to 2022, both in net weight (272.6 thousand tons in 2022 and 203.9 thousand tons in 2023) and in sales value (US$ 5.79 billion in 2022 and US$ 4.30 billion in 2023) (Badan Pusat Statistik Indonesia, 2024).

One of the reasons for the decline in shoe export figures is that customer orders to shoe manufacturing companies have become significantly more complex and varied compared to previous years. This increase in product complexity and variety requires shoe manufacturing companies to enhance their production planning systems to become more efficient, ensuring they can still meet customer demands that are far more complex and diverse (Firtikiadis, Manavis, Kyratsis, & Efkolidis, 2024).

Clustering algorithms are methods used to group data based on specific similarities. K-Means, Support Vector Clustering (SVC), and Deep Autoencoders are among the most widely used algorithms in their respective categories. The application of clustering algorithms can assist shoe manufacturing companies in grouping diverse products, thereby simplifying production planning (Feng, Lin, & Wang, 2022) (Ezugwu, et al., 2022) (Drid, Abdelhamid, & taleb-ahmed, 2022) (Chen & Guo, 2023).

Based on the background above, this study is expected to provide a significant contribution to optimizing the production planning process in shoe manufacturing companies by laying the foundation for developing production

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14332
e-ISSN : 2541-2019
p-ISSN : 2541-044X

planning systems that support data-driven decision-making, thereby improving efficiency in shoe manufacturing companies.

The aim of this study is to analyze and compare the K-Means, Support Vector Clustering (SVC), and Deep Autoencoder algorithms to identify the most effective algorithm for optimizing the production planning process in shoe manufacturing companies through a case study at XYZ Footwear.

LITERATURE REVIEW

A literature review was conducted for this study to find the foundation of the study and to ensure that the study has a clear and structured methodology. The literature review in this study focuses on the algorithms that will be used, such as K-Means, Support Vector Clustering, and Deep Autoencoder.

K-Means

K-Means is a clustering algorithm that groups data based on the distance between centroids (the average distance of all points within a cluster). This algorithm is relatively simple and easy to implement, and it is very efficient for large datasets. However, it requires the number of clusters to be determined in advance and is highly sensitive to outliers and the initial selection of centroids (Feng, Lin, & Wang, 2022) (Ezugwu, et al., 2022).

Support Vector Clustering (SVC)

Support Vector Clustering (SVC) is a kernel-based algorithm that can generate complex cluster boundaries, handle outliers, and does not require predefined cluster numbers. However, it can be computationally expensive when dealing with large datasets and is highly sensitive to the selection of certain parameters, especially the kernel selection (Drid, Abdelhamid, & taleb-ahmed, 2022).

Deep Autoencoder

Deep autoencoder is a technique that combines the strengths of deep learning (specifically autoencoders) with clustering algorithms to group data into clusters. Autoencoders are a type of neural network designed to learn efficient coding of input data then map the input data to a latent space, while clustering algorithms (such as k-means) are used to group data points into distinct clusters based on some similarity measure (Chen & Guo, 2023).

METHOD

Based on the literature study conducted, the models used in this study are K-Means, Support Vector Clustering (SVC), and Deep Autoencoder. These models were selected because they are capable of handling large datasets and perform very well in extracting information from the provided data (Feng, Lin, & Wang, 2022) (Ezugwu, et al., 2022) (Drid, Abdelhamid, & taleb-ahmed, 2022) (Chen & Guo, 2023). The research methodology used in this study was shown in Fig. 1.

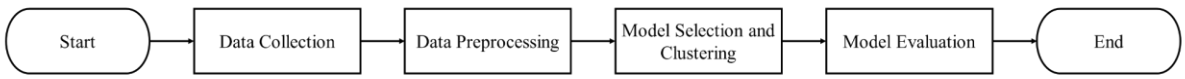

Fig. 1 Research methodology

Data Collection

The data collection process is carried out by directly retrieving data from the XYZ Footwear database. The data collected pertains to the 2024 production season, consisting of two production seasons: SS24 (September 2023 - February 2024) and FW24 (March 2024 - August 2024). The dataset consists of 96.098 rows, categorized into 174 shoe models, 333 shoe variants, and 7.132 purchase orders (POs). With such a large and diverse dataset, it is expected to provide a comprehensive depiction of the complexity involved in the shoe production planning process.

Data Preprocessing

The data preprocessing process is performed to make sure that the dataset is ready to be applied in the models by removing duplicates, removing empty data, encoding categorical features, standardizing the features, and applying dimensional reduction (Ndung'u, 2022) (Chaudhary, 2023) (Mutinda & Langat, 2024).

Model Selection and Clustering

The model selection process is performed to find the best parameters for each selected algorithm. For K-Means, an elbow method is used to find the optimal number of clusters for the dataset. For Support Vector Clustering (SVC), different kernel functions, such as linear, polynomial, sigmoid, and radial based function (RBF)

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14332
e-ISSN : 2541-2019
p-ISSN : 2541-044X

were tested to find the optimal kernel function. For Deep Autoencoder, different activations, optimizers, learning rates, and dropouts were tested to find the optimal hyperparameters. The selected models then trained on the dataset and clustering was performed to partition the data into groups (Coraggio & Coretto, 2023).

Model Evaluation

The model evaluation used in this study are silhouette score and davies-bouldin index (DBI). Silhouette score is an evaluation metric that measures how similar each point is to its own cluster compared to the other clusters. The value ranges from -1 (bad clustering) to +1 (well-defined clustering), while scores close to 0 indicate overlapping clusters (Žagar & Demšar, 2022) (Oyewole & Thopil, 2022) (Yin, et al., 2024). The silhouette score for a data point (i) is calculated as:

$$S(i) = \frac{b(i) - a(i)}{max[a(i),b(i)]} \quad (1)$$

where S = silhouette score; a = the average distance from a data point to other data points within the same cluster; and b = the average distance from a data point to the data points in the nearest cluster.

Davies-bouldin index (DBI) is an evaluation metric that measures the average similarity ratio of each cluster with the cluster that is most similar to it. The lowest possible value for DBI is 0, and the lower the score indicates better clustering (Žagar & Demšar, 2022) (Oyewole & Thopil, 2022) (Yin, et al., 2024). The davies-bouldin index (DBI) is calculated as:

$$DBI = \frac{1}{n_c}\sum_{i=1}^{n_c} \quad max_{i \neq j}\left(\frac{r_i + r_j}{d(c_i, c_j)}\right) \quad (2)$$

where DBI = davies-bouldin index; n_c = the number of clusters; r_i, r_j = the average distance from the centroid to the data points in cluster i and j; and $d(c_i, c_j)$ = the distance between the centroids of cluster i and j.

RESULT

The results based on the selected algorithms: K-Means, Support Vector Clustering (SVC), and Deep Autoencoder are as described below.

K-Means

The pseudocode used in the K-Means algorithm is shown in Fig. 2.

1. Load the dataset:
   - read the dataset from CSV file
2. Feature selection:
   - select relevant features from the dataset: ‘po_number’, ‘category_id’, ‘model_id’, ‘article_id’, ‘main_size’
3. Data preprocessing:
   - remove duplicate rows
   - drop rows with missing values
   - apply label encoding to convert categorical features to numeric
   - standardize features to scale the value into a similar range
   - apply Principal Component Analysis (PCA) to reduce the dataset into 2 dimensions
4. Elbow method for optimal clusters:
   - perform K-Means clustering for cluster between 1 and 10
   - calculate the Sum of Squared Errors (SSE) for each number of clusters
   - plot the SSE values against the number of clusters
   - identify the “elbow” point to find the optimal number of clusters
5. K-Means clustering:
   - perform K-Means clustering using the optimal number of clusters
   - assign cluster labels to each data point
   - plot the data points colored by their assigned cluster
6. Cluster evaluation:
   - extract the number of points in each cluster
   - compute the Silhouette Score
   - compute the Davies-Bouldin Score
7. Cluster identification:
   - use heatmap to visualize the distribution of each feature across clusters

Fig. 2 K-Means Pseudocode

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14332
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Using the elbow method to find the optimal number of clusters as visualized in Fig. 3, the optimal number of clusters found using K-Means clustering are 4 clusters.

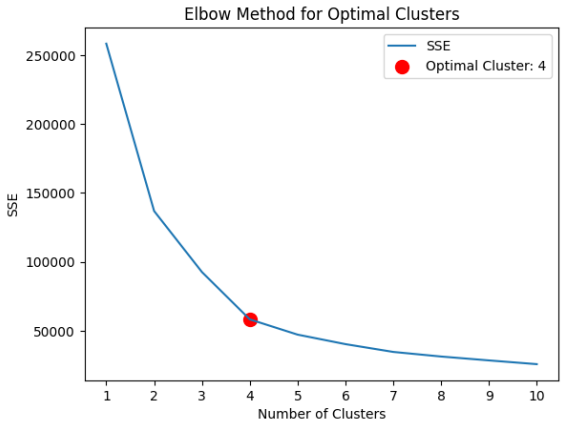

Fig. 3 Elbow method for optimal clusters

The dataset is then applied to the K-Means clustering using the optimal number of clusters found. The clustering results and the identification of each cluster are visualized in Fig. 4. For the cluster identification, the darker the heatmap indicates more data points in those side of the cluster, while the brighter the heatmap indicates less data points in those side of the cluster. The x-axis in the model identification was arranged from upper to lower class. The upper-class shoe models were models used by semi-professional to professional, while the lower-class shoe models were models used by amateurs to semi-professional. The x-axis in the main size identification was arranged from biggest to smallest shoe size.

Fig. 4 K-Means clustering visualization and identification of clusters

Based on the results obtained, by using K-Means clustering the dataset was clustered into 4 clusters, where cluster 0 is a cluster consisting of lower class shoe model with bigger shoe size (total of 19.274 data points), cluster 1 is a cluster consisting of upper class shoe model with smaller shoe size (total of 29.687 data points), cluster 2 is a cluster consisting of lower class shoe model with smaller shoe size (total of 28.094 data points), and cluster 3 is a cluster consisting of upper class shoe model with bigger shoe size (total of 19.043 data points). The K-Means clustering evaluation shows a silhouette score of 0.4716 and davies-bouldin index (DBI) of 0.7397.

Support Vector Clustering (SVC)

The pseudocode used in the Support Vector Clustering (SVC) algorithm is shown in Fig. 5.

1. Load the dataset:
   - read the dataset from CSV file
2. Feature selection:
   - select relevant features from the dataset: ‘po_number’, ‘category_id’, ‘model_id’, ‘article_id’, ‘main_size’
3. Data preprocessing:
   - remove duplicate rows
   - drop rows with missing values

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14332
e-ISSN : 2541-2019
p-ISSN : 2541-044X

- apply label encoding to convert categorical features to numeric
- standardize features to scale the value into a similar range
- apply Principal Component Analysis (PCA) to reduce the dataset into 2 dimensions

4. Apply Support Vector Clustering (SVC) with different kernel types:
   - perform clustering using different kernels: 'rbf', 'poly', 'sigmoid', 'linear'
   - fit the model on PCA-reduced data for each kernel
5. Support Vector Clustering (SVC):
   - perform Support Vector Clustering (SVC) using the optimal kernel types
   - assign cluster labels to each data point
   - plot the data points colored by their assigned cluster
6. Cluster evaluation:
   - extract the number of points in each cluster
   - compute the Silhouette Score
   - compute the Davies-Bouldin Score
7. Cluster identification:
   - use heatmap to visualize the distribution of each feature across clusters

Fig. 5 Support Vector Clustering (SVC) Pseudocode

After applying Support Vector Clustering (SVC) using different kernels, the optimal kernel for the dataset is linear. The clustering results and the identification of each cluster are visualized in Fig. 6. For the cluster identification, the darker the heatmap indicates more data points in those side of the cluster, while the brighter the heatmap indicates less data points in those side of the cluster. The x-axis in the model identification was arranged from upper to lower class. The upper-class shoe models were models used by semi-professional to professional, while the lower-class shoe models were models used by amateurs to semi-professional. The x-axis in the main size identification was arranged from biggest to smallest shoe size.

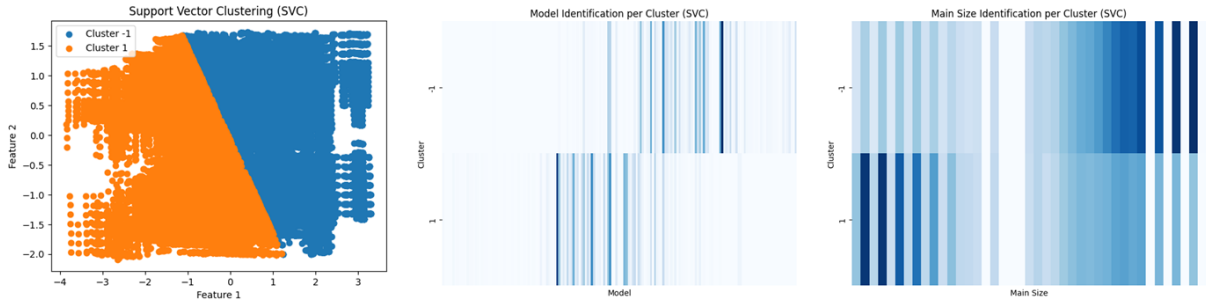

Fig. 6 Support Vector Clustering (SVC) visualization and identification of clusters

Based on the results obtained; by using Support Vector Clustering (SVC) the dataset was only able to be clustered into 1 cluster (cluster 1, consisting of 48.057 data points), while the rest were interpreted as noise (cluster -1, consisting of 48.041 data points). The cluster also cannot be identified because the model and main size in cluster 1 and the noise cluster are overlapped with each other. The Support Vector Clustering (SVC) evaluation shows a silhouette score of 0.3728 and davies-bouldin index (DBI) of 1.0846.

Deep Autoencoder

The pseudocode used in the Deep Autoencoder algorithm is shown in Fig. 7.

1. Load the dataset:
   - read the dataset from CSV file
2. Feature selection:
   - select relevant features from the dataset: 'po_number', 'category_id', 'model_id', 'article_id', 'main_size'
3. Data preprocessing:
   - remove duplicate rows
   - drop rows with missing values
   - apply label encoding to convert categorical features to numeric

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

- standardize features to scale the value into a similar range
- apply Principal Component Analysis (PCA) to reduce the dataset into 2 dimensions

4. Elbow method for optimal clusters:
   - perform K-Means clustering for cluster between 1 and 10
   - calculate the Sum of Squared Errors (SSE) for each number of clusters
   - plot the SSE values against the number of clusters
   - identify the "elbow" point to find the optimal number of clusters
5. Autoencoder model hyperparameter tuning:
   - define list of hyperparameters to test: dropout rates, activation functions, optimizers, learning rates
   - build autoencoder model using different combination of hyperparameters
   - train the model with each combination
   - print the combination testing results and show the best hyperparameter combination
6. Train autoencoder with best autoencoder model:
   - train the autoencoder using the best hyperparameters combination
7. K-Means clustering:
   - perform K-Means clustering using the optimal number of clusters
   - assign cluster labels to each data point
   - plot the data points colored by their assigned cluster
8. Cluster evaluation:
   - extract the number of points in each cluster
   - compute the Silhouette Score
   - compute the Davies-Bouldin Score
9. Cluster identification:
   - use heatmap to visualize the distribution of each feature across clusters

Fig. 7 Deep Autoencoder Pseudocode

After applying autoencoder model hyperparameter tuning, the top 5 smallest losses hyperparameters combination were shown in Table 1.

Table 1 Autoencoder top 5 smallest losses hyperparameters combination

No	Dropout	Activation	Optimizer	Learning Rate	Final Loss
1	0.1	linear	nadam	0.0010	0.043212
2	0.1	linear	nadam	0.0001	0.043250
3	0.1	linear	nadam	0.0005	0.043325
4	0.1	linear	adam	0.0001	0.043540
5	0.1	linear	rmsprop	0.0001	0.043606

The dataset is then trained to the autoencoder model and applied to the K-Means clustering using each combination found in Table 1. The best clustering results and the identification of each cluster are visualized in Fig. 8. For the cluster identification, the darker the heatmap indicates more data points in those side of the cluster, while the brighter the heatmap indicates less data points in those side of the cluster. The x-axis in the model identification was arranged from upper to lower class. The upper-class shoe models were models used by semi-professional to professional, while the lower-class shoe models were models used by amateurs to semi-professional. The x-axis in the main size identification was arranged from biggest to smallest shoe size.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14332
e-ISSN : 2541-2019
p-ISSN : 2541-044X

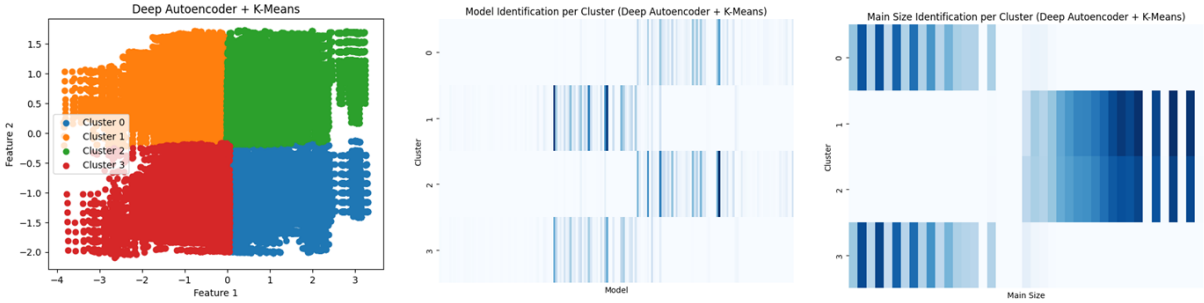

Fig. 8 Deep Autoencoder + K-Means clustering visualization and identification of clusters

Based on the results obtained, by using Deep Autoencoder and K-Means combination the dataset was clustered into 4 clusters, where cluster 0 is a cluster consisting of lower class shoe model with bigger shoe size (total of 19.276 data points), cluster 1 is a cluster consisting of upper class shoe model with smaller shoe size (total of 29.636 data points), cluster 2 is a cluster consisting of lower class shoe model with smaller shoe size (total of 28.140 data points), and cluster 3 is a cluster consisting of upper class shoe model with bigger shoe size (total of 19.046 data points). The K-Means clustering evaluation shows a silhouette score of 0.4822 and davies-bouldin index (DBI) of 0.6741 with a validation loss of 0.0062.

DISCUSSIONS

This study analyzed and compared three clustering algorithms: K-Means, Support Vector Clustering (SVC), and Deep Autoencoder + K-Means on a large and diverse shoe production dataset from XYZ Footwear. The results revealed that K-Means effectively grouped the dataset into 4 clusters, achieving a silhouette score of 0.4716 and a Davies-Bouldin Index of 0.7397. Support Vector Clustering (SVC) struggled with the dataset. It classified a significant portion of the data as noise, resulting in a silhouette score of 0.3728 and a Davies-Bouldin Index of 1.0846, making it unsuitable for this application. Deep Autoencoder + K-Means outperformed the standalone algorithms. It clustered the dataset into 4 distinct groups with the highest silhouette score of 0.4822 and the lowest Davies-Bouldin Index of 0.6741. The comparison result was shown in Table 2.

Table 2 Clustering evaluation comparison

Evaluation Metric	K-Means	Support Vector Clustering	Deep Autoencoder + K-Means
Silhouette score	0.4716	0.3728	0.4822
Davies bouldin index	0.7397	1.0846	0.6741
Number of clusters	4	1 + Noise	4

CONCLUSION

Based on the study conducted, it was concluded that Support Vector Clustering (SVC) is not suitable to identify patterns in shoe characteristics. While K-Means and Deep Autoencoder both perform reasonably well, the combination of Deep Autoencoder and K-Means have better results than just K-Means alone in identifying patterns in shoe characteristics. This highlights the potential of combining deep learning and clustering for handling complex datasets, such as the given dataset where the data were diverse and high-dimensional.

Further works, such as building a better autoencoder model that is more suitable for identifying patterns in shoe characteristics; or implementing the model built in this study in the shoe manufacturing companies' production planning systems could be done to develop more effective production planning systems.

REFERENCES

Badan Pusat Statistik Indonesia. (2024). *Ekspor Sepatu Olahraga menurut Negara Tujuan Utama, 2012-2023*. Retrieved October 23, 2024, from Badan Pusat Statistik: https://www.bps.go.id/id/statistics-table/1/MjAzMiMx/ekspor-sepatu-olahraga-menurut-negara-tujuan-utama--2012-2023.html

Chaudhary, R. A. (2023). *An Introduction to Data Encoding and Decoding in Data Science*. Retrieved October 23, 2024, from SitePoint: https://www.sitepoint.com/data-encoding-decoding-data-science-introduction

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Chen, S., & Guo, W. (2023). Auto-Encoders in Deep Learning—A Review with New Perspectives. *Mathematics, 11*. doi:10.3390/math11081777

Coraggio, L., & Coretto, P. (2023). Selecting the number of clusters, clustering models, and algorithms. A unifying approach based on the quadratic discriminant score. *Journal of Multivariative Analysis, 196*. doi:10.1016/j.jmva.2023.105181

Drid, A., Abdelhamid, D., & taleb-ahmed, A. (2022). Support vector machine based clustering: A review. *International Symposium on iNnovative Informatics of Biskra (ISNIB)*. doi:10.1109/ISNIB57382.2022.10076027

Ezugwu, A., Ikotun, A., Olaide, O., Abualigah, L., Agushaka, O., Eke, C., & Akinyelu, A. (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence, 110*. doi:10.1016/j.engappai.2022.104743

Feng, G., Lin, J., & Wang, K. (2022). Researches Advanced in Clustering Algorithms. *Highlights in Science, Engineering and Technology, 16*, 168-177. doi:10.54097/hset.v16i.2498

Firtikiadis, L., Manavis, A., Kyratsis, P., & Efkolidis, N. (2024). Product Design Trends within the Footwear Industry: A Review. *Designs, 8*. doi:10.3390/designs8030049

Mutinda, J., & Langat, A. (2024). Exploring the Role of Dimensionality Reduction in Enhancing Machine Learning Algorithm Performance. *Asian Journal of Research in Computer Science, 17*, 157-166. doi:10.9734/AJRCOS/2024/v17i5445

Ndung'u, R. (2022). Data Preparation For Machine Learning Modelling. *International Journal of Computer Applications Technology and Research,, 11*, 231-235. doi:10.7753/IJCATR1106.1008

Oyewole, G., & Thopil, G. (2022). Data clustering: application and trends. *Artificial Intelligence Review, 56*. doi:10.1007/s10462-022-10352-y

World Footwear. (2023). *The World Footwear Yearbook 2023*. Retrieved October 23, 2024, from World Footwear: https://www.worldfootwear.com/yearbook/the-world-footwear-2023-Yearbook/231.html

World Footwear. (2024). *The World Footwear Yearbook 2024*. Retrieved October 23, 2024, from World Footwear: https://www.worldfootwear.com/yearbook/the-world-footwear-2024-Yearbook/232.html

Yin, H., Aryani, A., Petrie, S., Nambissan, A., Astudillo, A., & Cao, S. (2024). A Rapid Review of Clustering Algorithms. *ArXiv*. Retrieved from https://arxiv.org/abs/2401.07389

Žagar, A. P., & Demšar, J. (2022). Model Evaluation: How to Accurately Evaluate Predictive Models. 253-274. doi:10.1007/978-3-030-88389-8_13

*name of corresponding author