Clustering in Game Analysis on FIFA22 Official Players Data

Muhammad Faidi Akif Md Ali
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
Shah Alam, Malaysia
2021196337@isiswa.uitm.edu.my

Muhammad Amir Fahmy Muhalith
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
Shah Alam, Malaysia
2021149601@isiswa.uitm.edu.my

Muhammad Syazwan Fikri Sahran
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
Shah Alam, Malaysia
2021100233@isiswa.uitm.edu.my

Zubli Quzaini Zubli
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
Shah Alam, Malaysia
2021119947@isiswa.uitm.edu.my

Siti Nur Kamaliah Kamarudin
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
Shah Alam, Malaysia
*kamaliah@fskm.uitm.edu.my

Shuzlina Abdul Rahman
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
Shah Alam, Malaysia
shuzlina@fskm uitm.edu.my

2022 3rd International Conference on Artificial Intelligence and Data Sciences (AiDAS) | 978-1-6654-9164-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/AIDAS56890.2022.9918708

***Abstract*— This study investigates the similarities of skill sets based on football players' position and performance of the clustering algorithm of two unsupervised learning methods: the k-Means and the Hierarchical Clustering. The study aims to discover the number of clusters and performance of machine learning algorithms, k-Means, and Hierarchical Clustering to group players based on their position. The FIFA22 dataset was obtained from Kaggle, which had undergone a dimensionality reduction method, the Principal Component Analysis (PCA). The comparison of the model for both methods with and without PCA showed a better silhouette score using the PCA data reduction. The hyperparameter tuning was applied to the models with k=3, k=4, k=8 and k=14, where k is the number of clusters, and the value of k is selected based on the player's positioning in the football match. The best model with different hyperparameters was evaluated using the Sum of Squared Error (SSE) and the Silhouette Analysis. The findings indicate k-Means and Hierarchical Clustering model can achieve a silhouette score greater than 0.5. In addition, the k-Means model scores 16% higher than Hierarchical Clustering on k=14. Results found that the k-Means algorithm with k=3 is the best performing model. Based on the observation, each cluster is grouped with similar skills values of different players' attributes.**

Keywords—FIFA22 dataset, k-Means, Hierarchical Clustering, Unsupervised Learning, Machine Learning, Data Mining

I. INTRODUCTION

Skills are essential in measuring an individual's ability to perform specific tasks. Skills involve implementing physical action based on learning or experience through the motor and mental understanding [1]. In a competitive football league, skill assessment assist in talent identification and selecting the best position that matches the player's skill. The rise of football players' data in recent years is influenced by popularity and technological advancement [2].

The main objective of this study is to discover the number of clusters and the performance of machine learning algorithms, k-Means and Hierarchical Clustering to group players based on their position. This study will compare the two algorithms' performance and evaluate them using Sum of Squared Error (SSE) and Silhouette Analysis.

This study helps to give insights for a football club to conduct strategy, virtual simulated training, and find a replacement for players due to injuries or approaching retirements. Applying machine learning techniques can benefit football clubs, and players can better understand their roles on the pitch and the relevant attributes needed for each position.

This study discerns relevant skills needed for each position in the football match using unsupervised machine learning techniques (k-Means and Hierarchical Clustering) and compares the performance of both models. The problem identified in this research involves the specific skills similarities between football players, where a single player can play more than one position in a game. However, the player's abilities may not fit their preferred position. Hence this study aims to identify the relevant skills needed for each position in the football match using unsupervised machine learning techniques.

II. RELATED WORKS

There is an increase in research on finding the similarities of skillset among football players. Most studies applied their analysis to unsupervised machine learning techniques such as Expectation Maximum, k-Means, or Hierarchical Clustering. Unsupervised machine learning uses machine learning algorithms to analyse and cluster unlabelled datasets [3]. Previous research evaluated the clustering models as either Expectation Maximum against k-Means [4] and Expectation Maximum against Hierarchical Clustering [5]. In [4] and [5], both found that k-Means and Hierarchical Clustering outperformed Expectation Maximum in grouping the similarities in the player's skills.

Another research by [6] investigated the clustering of football players' performance, where the authors customised a dissimilarity measure clustering method. In [6], six clustering algorithms were applied using two different weights. The research successfully clustered players according to their skills. However, expert assessments were needed to validate the clustering for this research. In [7], a mixed-attributes fuzzy clustering model was applied to cluster football players based on their performance and positional attributes. For both [6] and [7], reviews and comparisons were made with other research work on clustering tasks.

Moreover, in the research by [8], several players are examined using a dataset of events to ascertain the coach's general strategy on the squad. During a soccer match, it

computes event streams and deduces phases rather than single occurrences. In addition, the research by [9] concentrates on the unsupervised learning techniques: k-Means and spectral clustering. The research was only concerned with the drills and the players' training data. However, there is no significant conclusion achieved by the authors in the research. Similarly, [10] applied the machine learning approaches and human experts by building a player vector that uses clustering and nearest neighbours to describe a player's playing style.

According to [11], the research used Spatio-Temporal Trajectory clustering that did not require erroneous human annotations and could automatically recognise counterattacks and counter-pressing. Besides, research done by [12] created a mechanism for rating soccer players based on weights determined by a Linear Support Vector Classifier (LSVC) model. The weights of the performance aspects, referred to as the significant variable coefficient, have been determined in this research. In this context, the feature weights determined by the machine learning models offer the opportunity to use machine learning rating and ranking techniques for the association of football performance [13].

III. METHODS

The experiment setup for the research is shown in Fig. 1, where the processes are divided into data collection and model design and development.

A. Data Collection

This dataset is gathered from Kaggle [14] with 16,710 instances; 65 attributes consist of 51 numeric attributes and 14 nominal attributes. Table 1 shows the sample of attribute information of the data.

TABLE I. Attribute Information

	Attribute Information	
	Attributes	***Description***
1	ID	Unique ID for the player
2	Name	Name of the player
3	Age	Age of the player
4	Preferred foot	Player's preferred foot
5	Dribbling	Player's dribbling skills

B. Model Design and Development

This research uses the development platform of the machine learning library, Python and Waikato Environment for Knowledge Analysis, WEKA. The clustering algorithms used in this research are k-Means and Hierarchical Clustering to find each player's best position and similarities. There are changes in distance function for the model development using Euclidean distance and Manhattan distance.

The dataset will undergo another iteration of the same dataset with the Principal Component Analysis (PCA). The research uses different clusters such as k=3, k=4, k=8 and k=14 for the hyperparameter tuning. Each number of k is selected based on the positioning of players on the pitch. For example, the player's position includes Attacking, Midfield and Defensive for k=3.

The selection of parameters followed the research done by [4] for k=4 and k=8, while k=14 was selected to include all unique football positions available in the dataset.

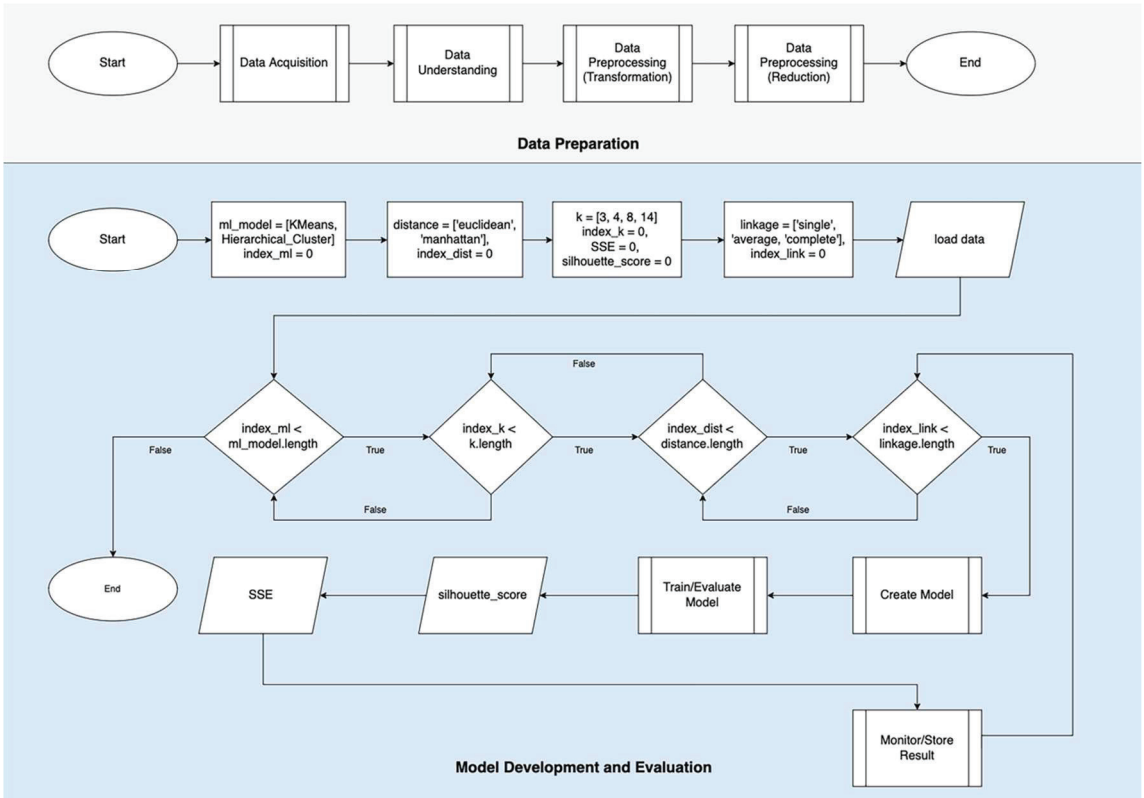

Fig. 1. Flowchart of the Experiment Setup

C. Clustering Measures

There is two formula that was used to evaluate the clustering results. The first formula is Sum of Squared Error (SSE), as shown in (1)

$$SSE = \sum_{i=1}^{n} (x_i - \overline{x})^2 \quad (1)$$

In this equation, n refers to the number of observations to be made whereas x_i refers to the value of the observed x, at ith iteration. The value inside the bracket will be the resulting difference between the individual data point x_i and the mean of all observations, $\overline{x}$. The squared differences will then be summed to produce the Sum of Squared Error for the given cluster.

The second formula used to measure the cluster quality is the Silhouette Coefficient, as shown in (2)

$$s = \frac{(b - a)}{\max(a, b)} \quad (2)$$

The Silhouette Coefficient score of clusters can be calculated using the mean intra-cluster distance denoted by a and the mean nearest-cluster distance denoted by b. The value denoted by a can be interpreted as how well each point in a is assigned, while the value b is the distance between a sample and the nearest cluster.

IV. Result and Discussion

The results of the experiment for the research are divided into three sections, the data preparation results, experimental results and cluster evaluation.

A. Data Preparation Results

Data cleaning is a part of pre-processing to remove noisy data in data preparation. Besides that, several missing data in other attributes have been replaced with the mean of the data. The data also undergo a transformation process such as normalisation of the player's skills value and reduction process to select the essential attributes using Information Gain evaluation.

As a result, 4866 instances and 38 attributes remain for the model training and development. Principal Component Analysis (PCA) has helped improve the clustering process by reducing data dimensionality. Thus, increasing the interpretability. The silhouette score for the PCA dataset obtained 0.4947, whereas, without PCA, the score is 0.1736 on the Hierarchical Clustering model with k=3. The comparison of the silhouette score is shown in Fig. 2.

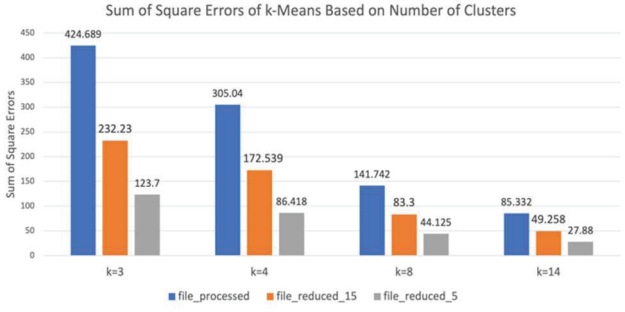

Fig. 2. Result of Silhouette Score for Processed vs Processed with PCA Data

B. Experimental Results

For the model performance results, we investigate the performance of clustering algorithms based on the number of attributes, including all attributes, 15 best attributes and five best attributes. First, the Sum of Squared Error (SSE) evaluation is based on the number of clusters using the k-Means algorithm. The results are shown in Fig. 3.

Second, the Silhouette Score evaluation for k-Means and Hierarchical Clustering is based on the number of clusters. The results are as shown in Fig. 4 and Fig. 5.

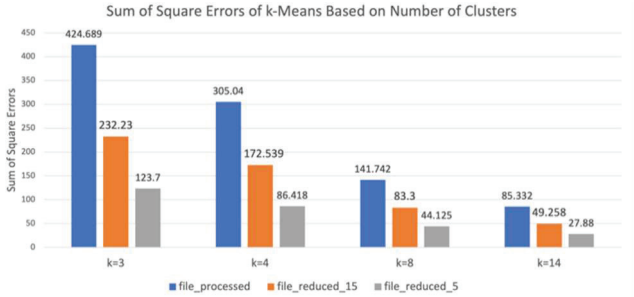

Fig. 3. Sum of Square Errors of k-Means based on the Number of Clusters

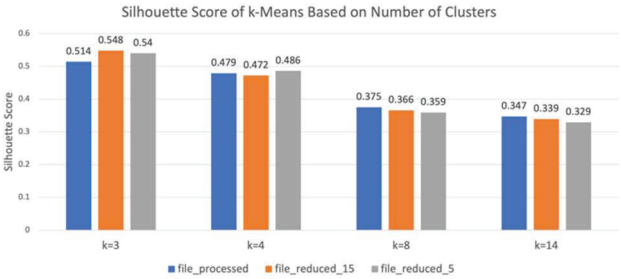

Fig. 4. Result of Silhouette Score of k-Means based on the Number of Clusters

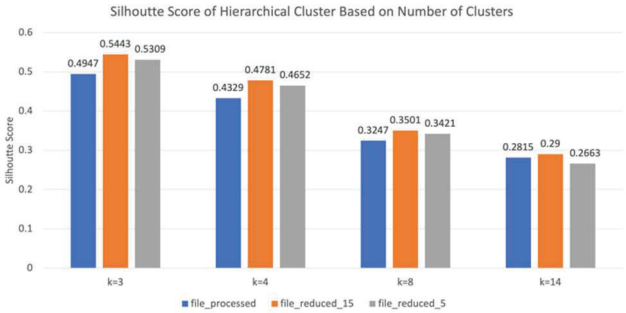

Fig. 5. Result of Silhouette Score of Hierarchical Clustering based on the Number of Clusters

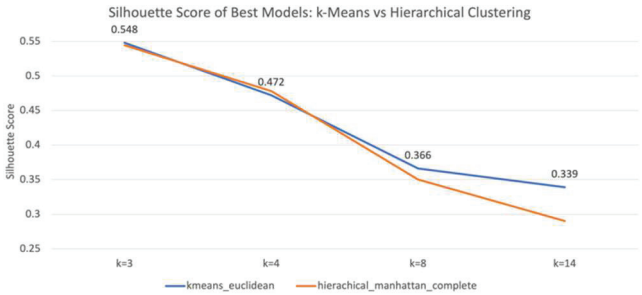

Fig. 6. Silhouette Score of Best Models: k-Means vs Hierarchical Clustering

Based on Fig. 6, this study found that the k-Means algorithms did better overall than the Hierarchical Clustering model. The silhouette score difference between k-Means and Hierarchical Clustering is 0.00037 on k=3, whereas the silhouette score for k=14, the k-Means algorithm produced 16% higher than Hierarchical Clustering. Using the silhouette evaluation method to evaluate the quality of clusters, the model (k=3) did better by clustering the players into three groups.

Other clusters' silhouette score is below 0.5, indicating a lack in terms of the separability of different players. This study concludes that other specific positions in the game do not carry much difference among the three types. Therefore, the k-Means model with k=3 got a higher silhouette score of 0.548 by reducing the data to 15 best attributes with the implementation of Principal Component Analysis.

It is discovered that with k=3, the model can classify players into three types, Attacking, Midfield and Defensive positions, as shown in Fig. 7, where the cluster in green shows the players with Attacking positions, followed by Midfield and Defensive positions in orange and blue.

Based on Fig. 7, it is shown that some positions may overlap with each other. For example, the position Centre Attacking Midfield, CAM is clustered in both Attacking and Midfield, indicating a similarity of the players' attributes for both positions.

Similarly, the Centre Back, CB position appeared in both Midfield and Defensive clusters showing a player's attribute shares the same pattern in terms of skills and playing style. The result is proven to show that some of the Centre Back players can interchange their position with Centre Defensive Midfield, CDM players.

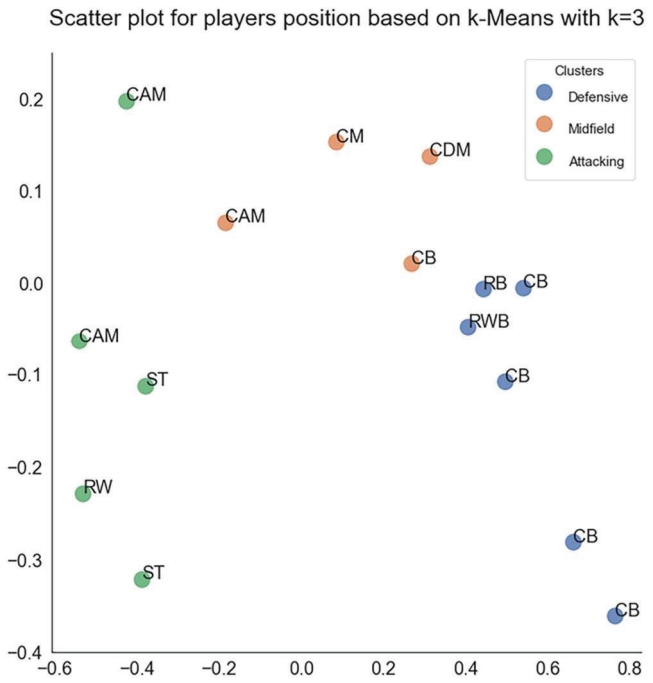

Fig. 7. Scatter Plot for Players Position based on k-Means with k=3

For visualisation purposes, the five most meaningful attributes are used to explain the disparity in the clusters; Positioning, Finishing, Interception, Slide Tackle and Standing Tackle, as shown in Fig. 8, Fig. 9, Fig. 10, Fig. 11, and Fig. 12.

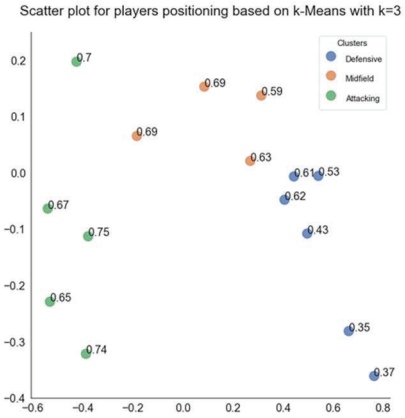

Fig. 8. Scatter Plot for Players Positioning based on k-Means with k=3

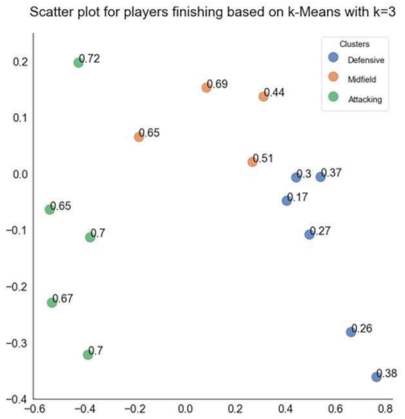

Fig. 9. Scatter Plot for Players Finishing based on k-Means with k=3

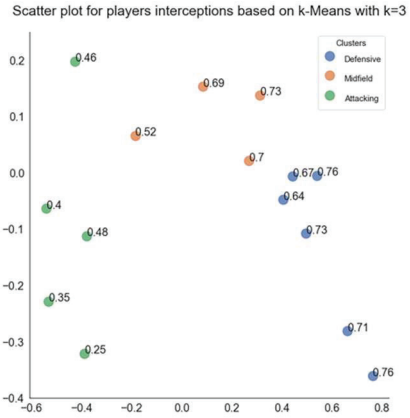

Fig. 10. Scatter Plot for Players Interceptions based on k-Means with k=3

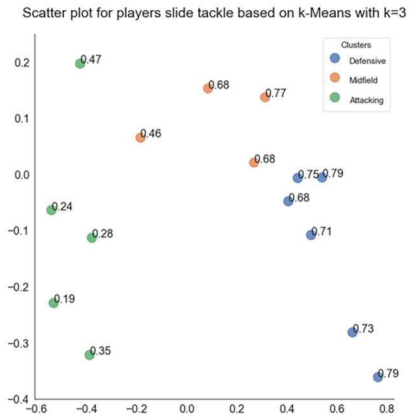

Fig. 11. Scatter Plot for Players Slide Tackle based on k-Means with k=3

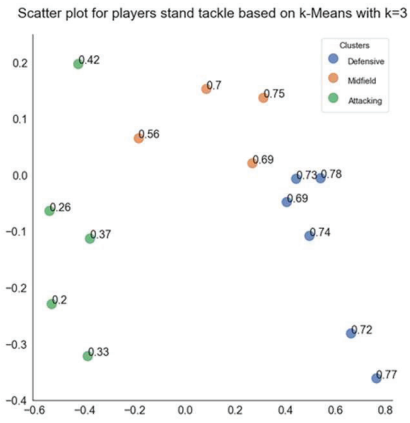

Fig. 12. Scatter Plot for Players Stand Tackle based on k-Means with k=3

The study extracted the skill values for each position based on the clusters based in the figures. Based on Fig. 9, the Attacking cluster in green has higher skill values for finishing. The range of values from 0.65 to 0.75 indicates that the players are mainly strikers and some midfielders focusing on attacking roles. Finishing can be stated as a vital attribute for attacking players.

Besides that, based on Fig. 10, the Defensive cluster in blue has a more defensive playing style with higher values for interception skills. The range of interception skills values from 0.64 to 0.76 proves that the players play more profoundly on the field and are primarily responsible for their team's defence.

A similar pattern of findings can be found in Fig. 11 and Fig. 12, showing higher values for the Defensive cluster, indicating the players in the cluster are grouped into defensive playing styles.

C. Cluster Evaluation

Silhouette analysis is a method to compare which graph displays the best characteristics out of all clusters. The evaluation is conducted by measuring the size and depth of each cluster. In addition, as shown in Fig. 13 and Fig. 14, the red dotted line represents the average silhouette coefficient.

Based on Fig. 13 and Fig. 14, each cluster in Fig. 14 has a wide fluctuation in size and width compared to clusters in Fig. 13. Therefore, the most compatible number of clusters is 3. The size and width are almost uniform throughout the clusters, confirming the findings in selecting k=3 using silhouette score over SSE.

This study found that the k-Means algorithm can group players according to the best position based on skill set. In addition, the study also found that using the k-Means model using k=3 can achieve a silhouette score greater than 0.5.

Results show that the k-Means model displayed a better silhouette score stability as the number of clusters increased according to the hyperparameter tuning. The PCA method assists in preserving the information throughout the clustering model development.

In this study, the k-Means can perform better across all numbers of k especially higher numbers of k. The results might be due to the differing nature of how these algorithms perform clustering. Hierarchical Clustering starts by combining the most similar data points.

On a higher number of k, the result shows that the data points are isolated and not clustered well. On the other hand, k-Means starts randomly with each data point set to random clusters initially. In contrast, k-Means clusters, regardless of the number of clusters, will always mean that all clusters are within the minimum distance of the centroid.

Therefore, k-Means is observed to perform with isolated data points. This difference in performance might also be due to the properties of the data, which contains little to no hierarchy in its nature which is unfavourable for the Hierarchical Clustering algorithm.

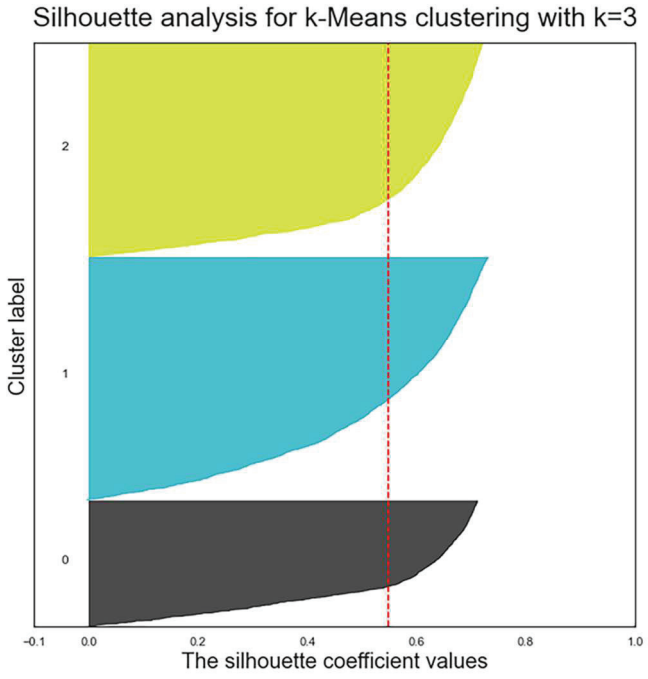

Fig. 13. Silhouette Plot for k-Means with k=3

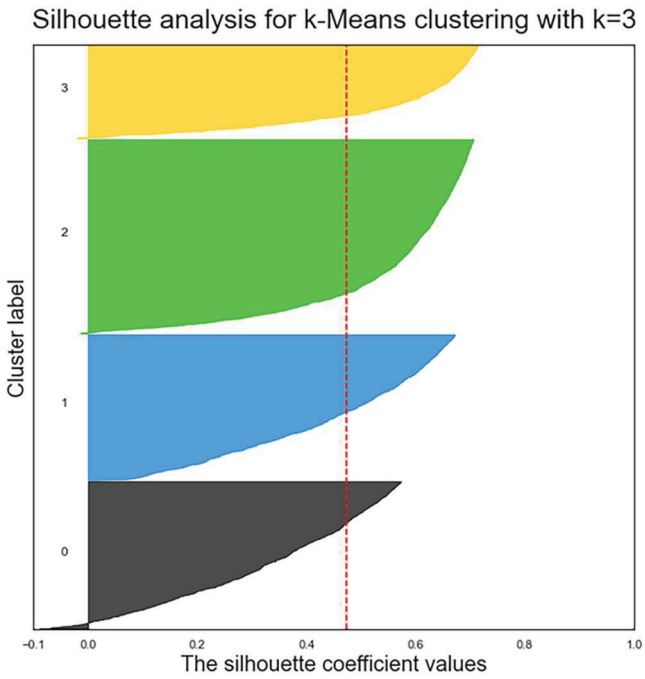

Fig. 14. Silhouette Plot for k-Means with k=4

V. Conclusion and Future Works

The primary purpose of this study is to find similarities in skillsets for football players based on the player's position and compare the performance of machine learning algorithms, the k-Means and Hierarchical Clustering. As a result, k-Means performed slightly better than Hierarchical Clustering.

The study proposed the k-Means model, with k=3, which is considered effective for clustering players' skills. The model can cluster players' data according to specific skill sets, validated through the findings. The study suggests that using PCA for dimensionality reduction reduces the number of attributes in the dataset by extracting essential attributes relevant for skills evaluation.

This research can benefit many parties, including researchers, football clubs, and players, in understanding football players' skill sets according to the best position. However, some limitations were found in the study. First, the Manhattan distance function is only applied to Hierarchical Clustering, and the dataset used was the FIFA22 dataset.

In conclusion, future work should include FIFA21, FIFA20, and FIFA19 datasets which may contribute to a better clustering model. Secondly, an algorithm such as k-medoids that can use the Manhattan distance function should be implemented to get a fairer result. Lastly, the number of clusters must be reduced in terms of gaps so it can be concluded as more concrete evidence.

Acknowledgment

The authors would like to thank the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Malaysia, for the support throughout this research.

References

[1] 'Skills Definition & Meaning - Merriam-Webster'. https://www.merriam-webster.com/dictionary/skills (accessed Aug. 12, 2022).

[2] C. Soto-Valero, 'A Gaussian mixture clustering model for characterizing football players using the EA Sports' FIFA video game system. [Modelo basado en agrupamiento de mixturas Gaussianas para caracterizar futbolistas utilizando el sistema de videojuegos FIFA de EA Sports].', *Ricyde. Revista Internacional De Ciencias Del Deporte*, vol. 13, pp. 244–259, 2017.

[3] H. U. Dike, Y. Zhou, K. K. Deveerasetty, and Q. Wu, 'Unsupervised Learning Based On Artificial Neural Network: A Review', *2018 IEEE International Conference on Cyborg and Bionic Systems, CBS 2018*, pp. 322–327, Jan. 2019, DOI: 10.1109/CBS.2018.8612259.

[4] G. W. A. Wijngaard, 'Clustering soccer players: investigating unsupervised learning on player positions', 2020, Accessed: Aug. 12, 2022. [Online]. Available: https://studenttheses.uu.nl/handle/20.500.12932/35795

[5] S. Roy and B. Sasmal, 'Integration of Hierarchical Clustering method and Dendrogram method with Expectation Maximisation for identification of the best player cluster', *2021 International Conference on Optimisation and Applications, ICOA 2021*, May 2021, DOI: 10.1109/ICOA51614.2021.9442644.

[6] S. E. Akhanli and C. Hennig, 'Clustering of football players based on performance data and aggregated clustering validity indexes', Apr. 2022, DOI: 10.48550/arxiv.2204.09793.

[7] P. D'Urso, L. de Giovanni, and V. Vitale, 'A robust method for clustering football players with mixed attributes', *Annals of Operations Research*, pp. 1–28, Feb. 2022, DOI: 10.1007/S10479-022-04558-X/FIGURES/9.

[8] T. Decroos, J. van Haaren, and J. Davis, 'Automatic Discovery of Tactics in Spatio-Temporal Soccer Match Data', in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 223–232. DOI: 10.1145/3219819.3219832.

[9] R. Mahfuz, Z. Mourad, and A. el Gamal, 'Analysing Sports Training Data with Machine Learning Techniques', *Methods*, vol. 9, pp. 0–38104976, 2016.

[10] J. Decroos Tom and Davis, 'Player Vectors: Characterizing Soccer Players' Playing Style from Match Event Streams', in *Machine Learning and Knowledge Discovery in Databases*, 2020, pp. 569–584.

[11] J. Hobbs, P. Power, L. Sha, H. Ruiz, and P. Lucey, 'Quantifying the value of transitions in soccer via spatiotemporal trajectory clustering', 2018.

[12] L. Pappalardo, P. Cintia, P. Ferragina, E. Massucco, D. Pedreschi, and F. Giannotti, 'PlayeRank: data-driven performance evaluation and player ranking in soccer via a machine learning approach', *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 5, pp. 1–27, 2019.

[13] Y. Li, R. Ma, B. Gonçalves, B. Gong, Y. Cui, and Y. Shen, 'Data-driven team ranking and match performance analysis in Chinese Football Super League', *Chaos, Solitons & Fractals*, vol. 141, p. 110330, 2020, DOI: https://doi.org/10.1016/j.chaos.2020.110330.

[14] 'FIFA22 OFFICIAL DATASET | Kaggle'. https://www.kaggle.com/datasets/bryanb/fifa-player-stats-database (accessed Aug. 12, 2022).