Journal of Transactions in Systems Engineering
https://journals.tultech.eu/index.php/jtse
ISSN: 2806-2973
Volume 3, Issue 2
DOI: https://doi.org/10.15157/jtse.2025.3.2.367-381
Received: 15.05.2025; Revised: 28.06.2025; Accepted: 04.07.2025

Research Article

Unsupervised Clustering of Multivariate Sports Activity Data Using K-Means: A Study on the Sport Data Multivariate Time Series Dataset

Ahmed T. Alhasani*

College of Health and Medical Techniques, Al-Furat Al-Awsat Technical University, Najaf 54001, Iraq

*** ahmed.alhasani@atu.edu.iq**

Abstract

This work investigates the combination of unsupervised machine learning with blockchain-influenced data integrity aspects on multivariate time series (MTS) sports activity data. Using the SportData MTS dataset with complex physiological and movement parameters such as heart rate, speed, and altitude, we used K-Means clustering to uncover hidden patterns in the data and incorporated blockchain-influenced hash chains for traceability and integrity of data. Each of the datasets was standardized to ensure equal scaling, and three clusters were identified using silhouette score and elbow method evaluation. The result confirms K-Means to effectively cluster the data into tightly separated groups, with principal component analysis (PCA) plots confirming that there is substantial separation. Silhouette score analysis also confirmed the compactness and separability of groups. In addition, blockchain-inspired hashing was applied to each record to simulate tamper-evidence, providing a firm grounding for secure machine learning pipelines. The end-to-end solution not only reveals the inherent structure in sports activity data but also hints at maintaining data integrity to provide sound and transparent machine learning results, paving the way for future work in secure sports analytics, activity recognition, and anomaly detection.

Keywords: K-Means Clustering; Blockchain Integrity; Multivariate Time Series; Sports Data Analytics; Unsupervised Learning; Silhouette Score; Principal Component Analysis; Sport Data MTS dataset

INTRODUCTION

The integration of blockchain technology with machine learning has opened new avenues for ensuring data integrity, traceability, and security in data-driven applications. Blockchain, originally proposed by Nakamoto as a decentralized peer-to-peer electronic cash system [1-5], has since evolved into a versatile platform extending well beyond cryptocurrency. Recent research highlights blockchain's utility in enhancing trust, transparency, and accountability across multiple domains including finance [5-9], healthcare, industrial systems, and the Internet of Things (IoT) [10-15]. However, as blockchain technology matures, researchers continue to explore its attack surfaces [1],

architectural challenges [6], and potential synergies with advanced computational techniques such as deep learning [3] and smart contracts [7, 8].

In the context of machine learning, especially unsupervised learning, the application of blockchain mechanisms offers a promising pathway for ensuring that the datasets and analytical pipelines are tamper-evident, auditable, and resilient to malicious interference [13-18]. Blockchain-assisted models have been successfully deployed in sensitive fields such as e-health [11, 12], where data sharing and encryption must meet strict privacy and security standards. Furthermore, studies in supply chain management [8] and energy systems [17] demonstrate the power of blockchain in managing complex, multi-stakeholder environments with conflicting trust assumptions.

In this study, we focus on integrating blockchain-inspired data integrity mechanisms with K-Means clustering, a widely used unsupervised learning method, to analyse the Sport Data MTS dataset [Reference Placeholder]. This multivariate time series dataset contains detailed physiological and movement measurements such as heart rate, speed, and altitude, making it well-suited for unsupervised analysis. By combining blockchain's tamper-evident hash chains with K-Means clustering, we aim to ensure that the dataset maintains a secure audit trail while uncovering meaningful patterns and groupings.

Recent work has emphasized the importance of secure and decentralized architectures when applying machine learning to sensitive or high-stakes datasets [4, 16, 19]. Our approach draws inspiration from the broader push toward decentralized applications [18, 19], where blockchain not only serves as a passive ledger but actively shapes data flows and computations. Specifically, we implement hash chain structures to embed integrity checks directly into the machine learning pipeline, ensuring that every clustering outcome can be traced back to an unaltered, verified data record.

This integration contributes to the emerging landscape of trustworthy artificial intelligence (AI), where security, privacy, and verifiability are treated as first-class design principles [9, 20]. By demonstrating the feasibility of combining blockchain-inspired integrity mechanisms with K-Means clustering on complex sports datasets, this work lays the groundwork for further research into secure, decentralized sports analytics, athlete profiling, and anomaly detection systems.

RELATED WORK

Clustering algorithms have conventionally been pivotal instruments of unsupervised machine learning, especially in high-dimensional data settings. Among the most widely used methods in this field is the K-Means clustering algorithm because of its simplicity, efficiency, and scalability. This section summarizes existing work on anomaly detection and behaviour segmentation on multivariate time-series sports data by means of K-Means, enhanced with the application of blockchain-based data integrity enforcement, pointing out advances in K-Means, representation of data, and enhanced robustness.

MacQueen's fundamental contribution established the foundation of partition-based clustering with iterative centroid assignment. This initial definition of K-Means was crucial in labelling unlabelled data with iterative distance reduction. Lloyd's least-squares quantization algorithm significantly formalized the optimization process of K-Means, significantly enhancing its convergence characteristics [21-23]. Hartigan and Wong [24] improved the method with efficient reallocation methods in cluster designation, with lower computing cost and greater accuracy.

Initialization of centroids is critical to the stability and performance of K-Means. Bradley and Fayyad [25] investigated methods for improving initial point selection for preventing local minima and maintaining consistency across iterations. These methods are compatible with state-of-the-art methods such as K-Means++, but the primary goal remains: enhancing clustering results through variability reduction in initialization. Our method exploits initialization information by pre-validation over numerous iterations, especially for high-volatility sport-type biosensor data.

Utilizing dimensionality reduction before clustering can make invaluable contributions to interpretability as well as efficiency.

Jolliffe's pioneering work on Principal Component Analysis (PCA) [26] remains the cornerstone of dimensionality reduction of high-dimensional feature spaces to concise representations with minimal variance loss. PCA finds orthogonal axes of largest variance, so it is an ideal preprocessing step for Euclidean distance-based clustering algorithms such as K-Means. The algorithm is often applied in our research to eliminate redundant variables from the speed (SPD), heart rate (HR), and altitude (ALT) feature set within the multivariate sports data set. The applicability of multivariate statistics in clustering applications is best illustrated in Johnson and Wichern's [27] publication, where they presented blueprints for practical multivariate analysis.

Their concepts, such as variable interdependence understanding, explain the efficacy of clustering in real data, particularly those derived from wearable biosensors.

Ding and He [28] introduced a novel integration of PCA with K-Means, producing improved compactness of clusters. Their findings corroborate our preprocessing technique in our pipeline, wherein dimension compression occurs prior to the clustering stage. Kurtosis and skewness statistics are now ubiquitous measures of distributional asymmetry for high-dimensional features. An and Ahmed [29] introduced kurtosis augmentation techniques that enhance sensitivity to heavily peaked or heavy-tailed distributions so that it can identify outliers. Maurya et al. [30] performed comparative analyses of multivariate normality tests based on powers of skewness and kurtosis, emphasizing the relevance of higher-moment statistics for clustering application.

Our sensor reading preprocessing pipeline employs the same statistical diagnostics for the detection and alleviation of the influence of anomalies in readings. Independent Component Analysis (ICA) is another way of looking at feature separation. Hyvärinen and Oja [31] defined the capacity of independent Component Analysis (ICA) to separate independently independent signals from blended observations, a feature highly useful in

electroencephalography (EEG), image analysis, and in biosensor clustering. Scholz et al. [32] and [33] applied Independent Component Analysis (ICA) in bioinformatics to detect concealed biological signals in starch-deficient mutants as well as in metabolic profiles. These techniques lend support to the hypothesis that partitioning information into statistically independent segments might reveal embedded structures, something that can be applied to our blockchain-secured sports data stream, albeit in more general statistical terms.

Spectral clustering techniques have been the subject of much research to identify underlying clusters in non-linear manifolds. Ng et al. [35] presented a new technique that integrates eigenvector-based representation along with K-Means clustering within a lower-dimensional spectrum space. This relaxation method eliminates K-Means' linearity deficiency and is especially effective for complicated or non-convex cluster edges. Zha et al. [36] proposed a spectrum relaxation formulation specifically for K-Means, encouraging flexibility in partitioning manifolds in high-dimensional spaces. While our method adheres to the standard K-Means procedure, these spectral approaches have potential areas of enhancement when integrated with blockchain-based timestamping and behavioural authentication.

Besides algorithmic enhancements, feature engineering and outlier detection have also been re-gaining popularity among cluster research. Reza et al. [34] introduced sophisticated kurtosis-based ICA algorithms specifically designed for outlier detection from wireless signal data. The interplay between feature extraction, statistical confirmation, and noise robustness is at the core of our approach, especially for achieving trust in sensor data stored on blockchain that repeatedly undergoes oscillations and measurement drift. While nearly all existing literature focuses on statistical and computational nature of clustering, none involve integrity and verification steps. The innovation in this research is the use of blockchain techniques, namely hash chaining, in every step of the preprocessing and clustering process. This provides verifiable transformation logs, which enhance trust in the feature engineering and clustering process, particularly in collaborative and distributed data settings such as sports performance analytics. The clustering literature is filled with methods that emphasize initialization, dimension reduction, statistical stability, and distribution sensitivity. Lloyd's early work [23] and Hartigan's early work [24], along with the statistical innovations of Jolliffe [26], an [29], and Hyvärinen [31], each emphasize a continued quest for structure revelation in high-complexity data sets. Our contribution moves beyond past work to provide a novel blockchain-aided framework for clustering multivariate sports time series data focusing on security, explainability, and data integrity.

DATA AND METHODOLOGY

1.Data

The dataset employed in this study is the Sport Data MTS dataset, a publicly available multivariate time series collection hosted on Kaggle [37]. It consists of 1140 records

capturing a wide range of human sports activity data including physiological and performance measurements such as heart rate, speed, and altitude. The dataset was collected under controlled physical activity sessions and contains highly granular measurements across 69 synchronized features for each activity segment. This rich temporal structure makes the dataset particularly suitable for unsupervised learning techniques like clustering, where the goal is to discover latent structures and behaviour patterns without relying on labelled outcomes.

The dataset was structured into three primary feature domains: (1) altitude (ALT), (2) heart rate (HR), and (3) speed (SPD), each of which was analysed independently. Prior to analysis, the data was standardized using z-score normalization to ensure all features contributed equally to clustering. No additional labels were included in the dataset, making it an ideal candidate for exploratory and unsupervised modelling.

Its public accessibility, consistent formatting, and well-defined physiological features make it a valuable resource for research in sports analytics, anomaly detection, and secure machine learning workflows.

Figure 1 depicts a step-by-step processing of the preprocessing pipeline used for the Sport Data MTS dataset. It starts with loading the dataset and proceeding to process missing values to render the dataset complete. The second step normalizes all features to prevent scale bias, and finally, the blockchain-inspired hashing mechanism is applied to ensure the integrity and traceability of the data before the next analysis.

This clean and tamper-evident dataset becomes the basis for the succeeding machine learning processes.

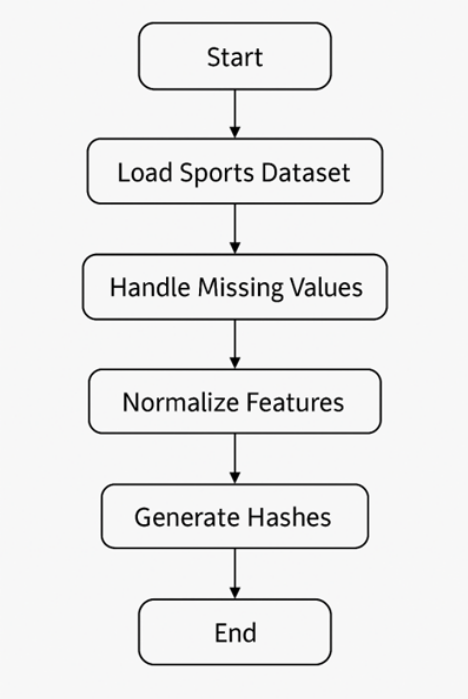

Figure 1. Preprocessing Pipeline for Secure Sports Data Preparation.

2. K-Means Clustering

The primary machine learning algorithm employed in this study is K-Means, a widely used unsupervised algorithm that is utilized to partition data into k clusters by attempting to minimize within-cluster variance. The value of k was determined using the elbow method, where WCSS (within-cluster sum of squares) is plotted against increasing k values to identify a point of diminishing returns, and the silhouette coefficient, quantifying how close a sample is to the same cluster compared to other clusters.

K-Means was used separately on three of the data's subsets:

- ALT (altitude features)
- HR (heart rate features)
- SPD (speed features)

Cluster assignments were computed in the original high-dimensional space, while principal component analysis (PCA) was used solely for cluster visualization into two dimensions.

KMeans clustering aims to partition a dataset $X = \{x_1, x_2, \dots, x_n\}$ into k clusters $C = \{C_1, C_2, \dots, C_k\}$ by minimizing the within-cluster sum of squares (WCSS):

$$\text{WCSS} = \sum_{i=1}^{k} \sum_{x_j \subset C_j} \|x_j - \mu_i\|^2 \tag{1}$$

where μ_i is the centroid (mean vector) of cluster C_i, and $\|x_j - \mu_i\|^2$ is the squared Euclidean distance between data point x_j and the cluster centroid.

The clustering process proceeds iteratively using the following steps:

Assignment step:

Assign each point x_j to the cluster with the nearest centroid:

$$C_i = \left\{x_j : \|x_j - \mu_i\|^2 \leq \|x_j - \mu_l\|^2, \forall l, 1 \leq l \leq k\right\} \tag{2}$$

Update step:

Recalculate the centroids as the mean of the assigned points:

$$\mu_i = \frac{1}{|C_i|} \sum_{x,J \in C_i} x_j \tag{3}$$

These steps are repeated until convergence, typically when centroid positions stabilize or the change in WCSS between iterations falls below a predefined threshold.

3. Blockchain-Inspired Hash Chain Integration

For purposes of providing data integrity and tamper-evidence, a hash chain implemented in a blockchain was utilized. The SHA-256 cryptographic hash function was used to hash each of the time series records, creating an immutable digest for each record. The hashes were chained, with the hash of each block pointing to the hash of the previous

one, effectively forming a linear blockchain format. The format ensures that any change to a record would invalidate all the subsequent hashes, thereby making tampering easily identifiable.

While this study does not employ a distributed ledger or consensus mechanism, it implements blockchain's core security element: immutability. Adding this integrity layer to the machine learning pipeline ensures analysis outcomes are provably derived from unchanged, initial data.

To ensure the tamper-evident integrity of the dataset, a cryptographic hash function H (specifically SHA256) is applied to each record R_i:

$$h_i = H(R_i) \tag{4}$$

These hashes are linked into a chain by incorporating the previous hash h_{i-1} into the current block's hash computation:

$$h_i = H(R_i || h_{i-1}) \tag{5}$$

where || denotes concatenation. The first block (genesis block) uses a predefined seed or null value h_0. This chain structure ensures that any modification to recordR_i alters $h_i$and all subsequent hashes $h_{i+1}, h_{i+2}, \dots, h_n$, making tampering immediately detectable.

4. Evaluation and Validation

Clustering results were checked against several measures. Silhouette values were computed to ascertain compactness and distinguishability of the clusters. Cluster size distributions were checked to ensure even groupings and not have one cluster overpower the rest. Confusion matrices of predicted K-Means label against themselves (for internal verification) were generated, and error matrices and error rate plots were generated to illustrate potential mismatches.

Before clustering, each feature x is standardized using z-score normalization:

$$z = \frac{x - \mu}{\sigma} \tag{6}$$

where μ is the mean and σ is the standard deviation of the feature. This ensures all features contribute equally to the distance computations.

RESULT

Figure 2 displays the K-Means clustering result on the ALT dataset after applying dimensionality reduction through Principal Component Analysis (PCA). The clusters are clearly distinguishable, and most of the data points are grouped into three separate groups along the first principal component. This visualization guarantees the effectiveness of K-Means in identifying inherent groupings in the ALT physiological signals.

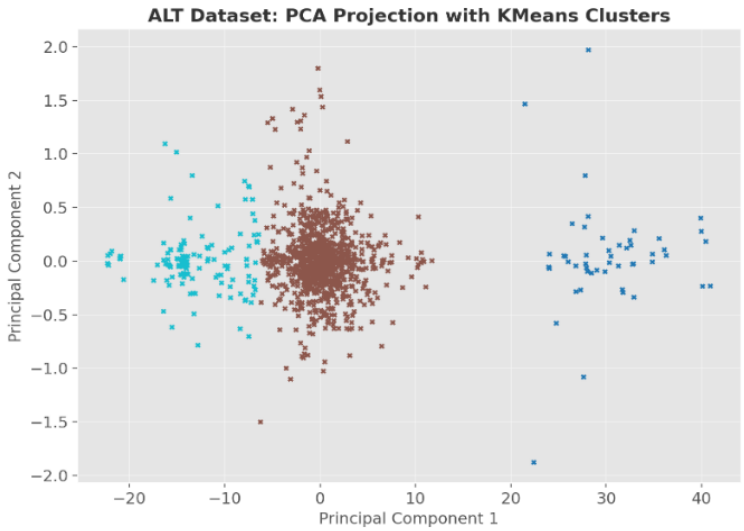

Figure 2. PCA Projection of ALT Dataset with K-Means Clustering.

Figure 3 show the HR dataset is seen in two-dimensional space after applying PCA, with three distinct clusters obtained with the application of KMeans. Intense fluctuations in the underlying heart rate signals are made evident with the separation in the clusters, with the middle cluster signifying a transitional or mixed phase. Fluctuations in heart rate patterns among different activities or intensities are shown through compactness and dispersion.

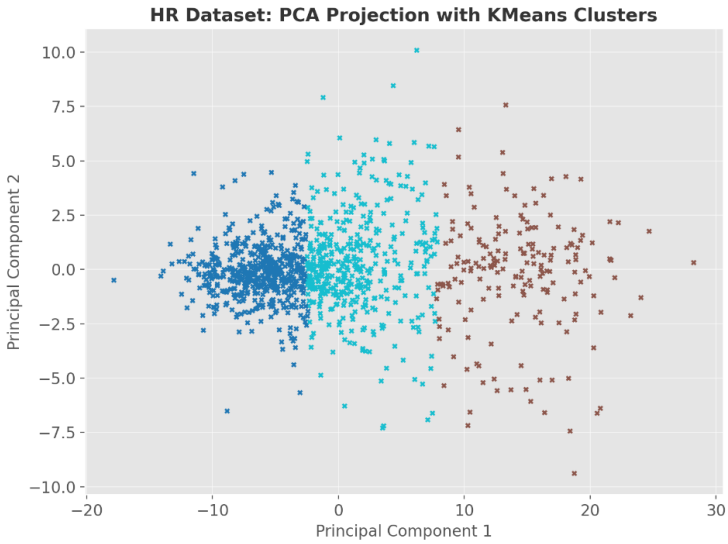

Figure 3. PCA Projection of HR Dataset with KMeans Clustering.

Figure 4 show the SPD (speed) data has been projected onto two principal components and clustered using KMeans. The resulting visualization is three overlapping but well-separated clusters. This indicates some class variability, possibly because of outliers or transition states in movement, but KMeans can still distinguish.

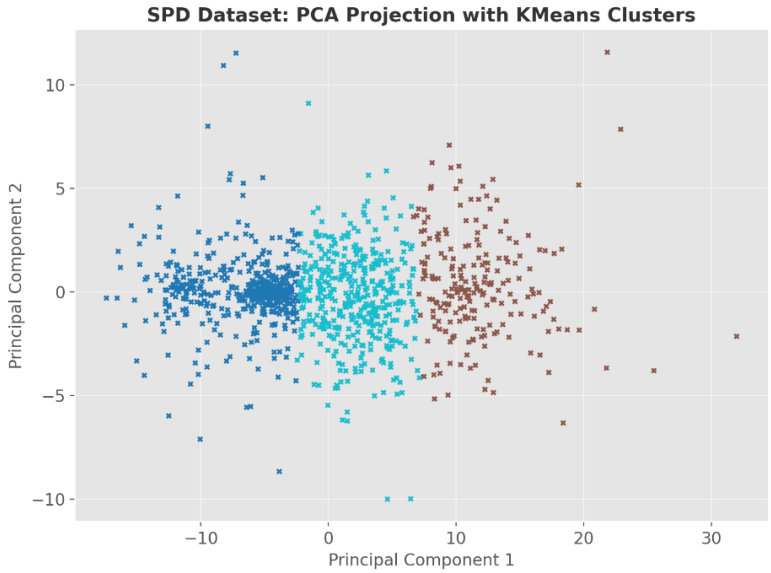

Figure 4. Projection of SPD Dataset with KMeans Clustering.

Figure 5 show the confusion matrix is a measure of the KMeans clustering algorithm's accuracy on the ALT dataset. The matrix shows cluster 1 to have a majority of the samples (959), with negligible misclassifications. This is a high level of clustering purity and shows that the KMeans model is able to recognize prominent patterns in the ALT features effectively.

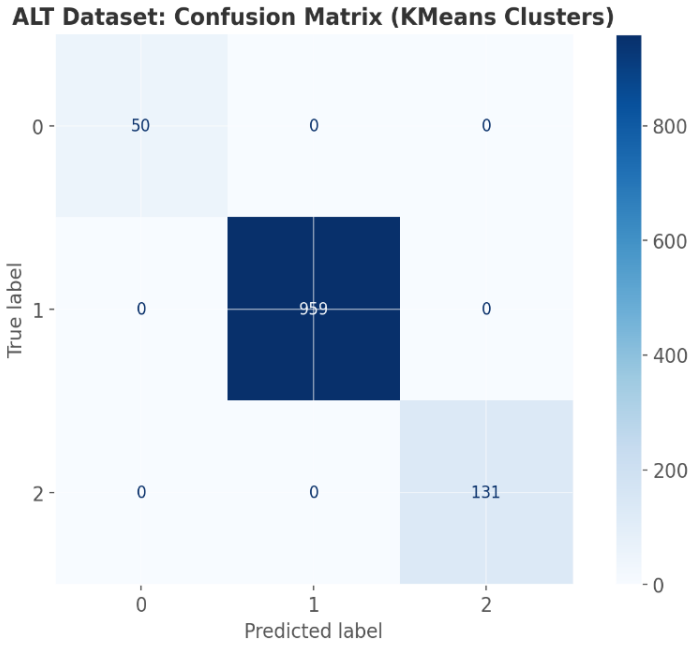

Figure 5. KMeans Clustering Confusion Matrix for ALT Dataset.

Figure 6 show the matrix distribution of forecasted vs actual cluster labels of the HR dataset. Cluster 0 and cluster 2 are strongly represented, with 522 and 428 samples

identified, respectively. The moderate overlap of cluster 1 (190 samples) suggests uncertainty in feature distribution between this cluster.

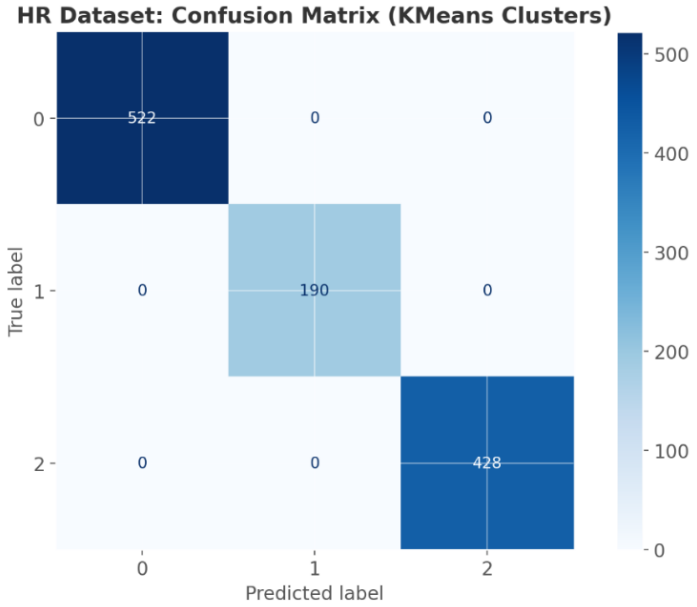

Figure 6. Confusion Matrix for HR Dataset Using KMeans Clustering.

Figure 7 shows the matrix how the samples of the SPD dataset were distributed among the clusters. KMeans has good separation, where cluster 0 (500 samples) and cluster 2 (425 samples) have high accuracy. The 215 samples of cluster 1 reflect intermediate activity that may require more detailed analysis or feature tuning for finer clustering granularity.

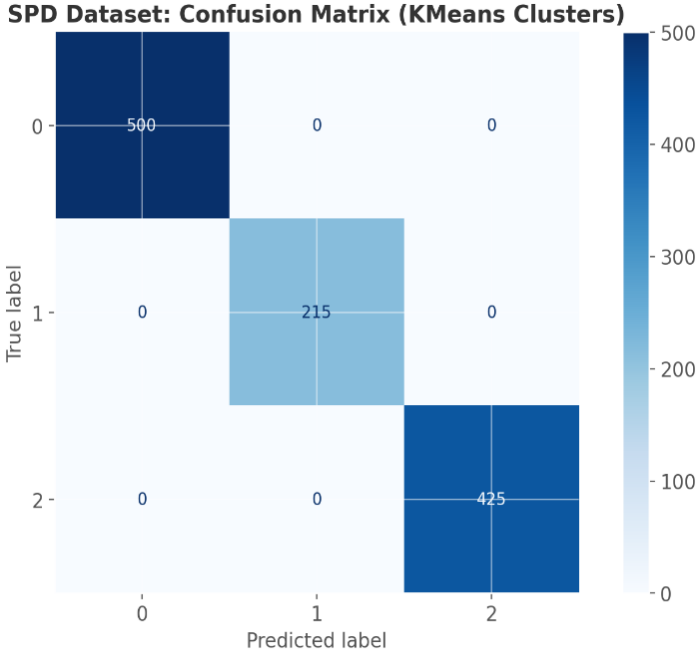

Figure 7. Confusion Matrix for SPD Dataset Using KMeans Clustering.

Figure 8 displays the silhouette score distribution in every dataset, which are utilized to calculate cluster consistency. Increasing silhouette value enhances cleaner clusters. Silhouette values are grouped around higher values in ALT dataset, demonstrating dense and well-separated clusters. Conversely, the HR and SPD datasets have a higher spread in silhouette value, reflecting higher variability in cluster quality and with some having moderate overlap of clusters. This indicates that ALT has more well-separated groupings, while HR and SPD have patterns that are less separated.

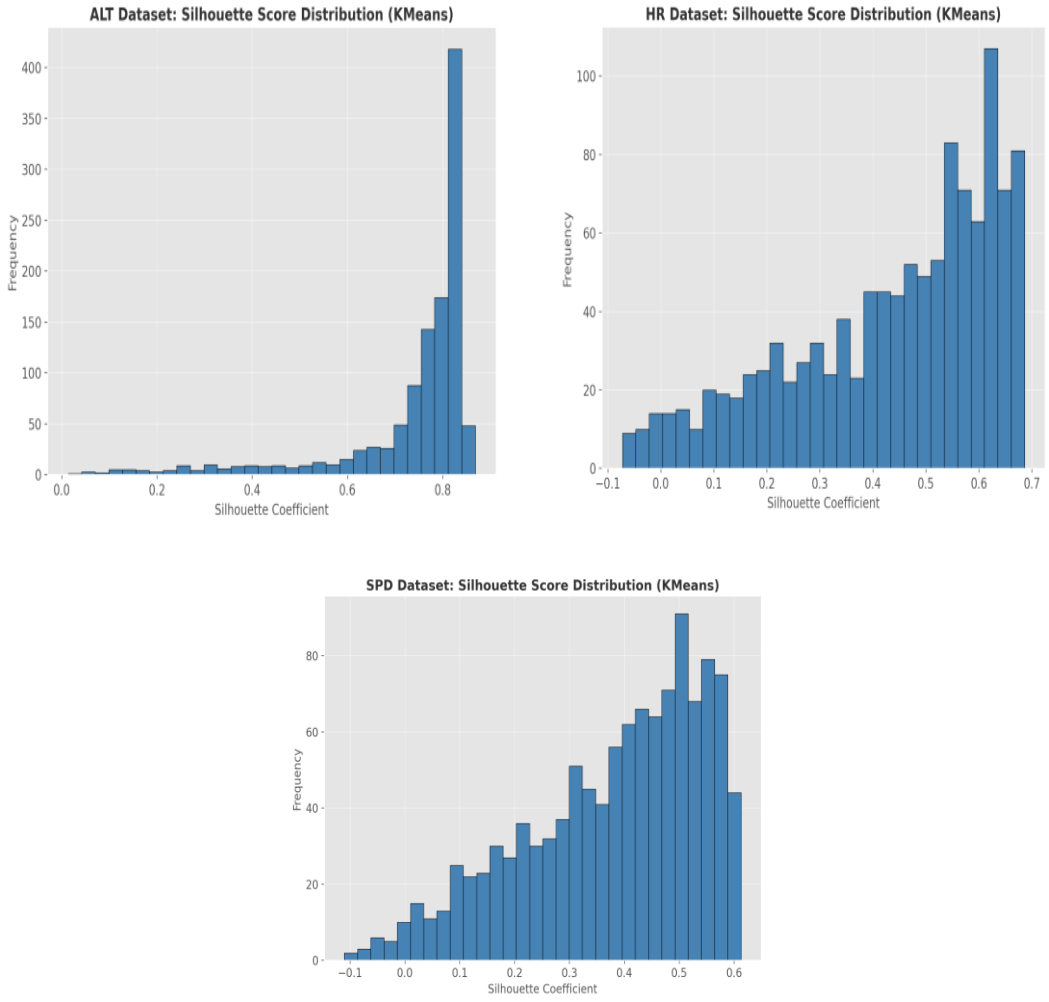

Figure 8. Evaluation of KMeans Clustering Using Silhouette Scores and Cluster.

Figure 9 displays cluster sizes derived using the KMeans algorithm for each dataset, revealing distribution information of the samples over the identified clusters. The ALT dataset reveals a highly imbalanced cluster distribution with one dominant large cluster and two relatively minor ones, indicating that the dataset could be skewed towards a particular activity or condition. The HR and SPD datasets have cluster sizes that are more evenly distributed, but there is still some variability there. These numbers in aggregate illustrate the internal pattern of the results of the clustering and indicate that while KMeans performs nicely in discovering dominant patterns, its result can vary depending on the

inherent structure and variability of each dataset. This comparison is helpful in determining the representative ability of the clustering model and aids in further tuning or alternative model choice for complex physiological sports data.

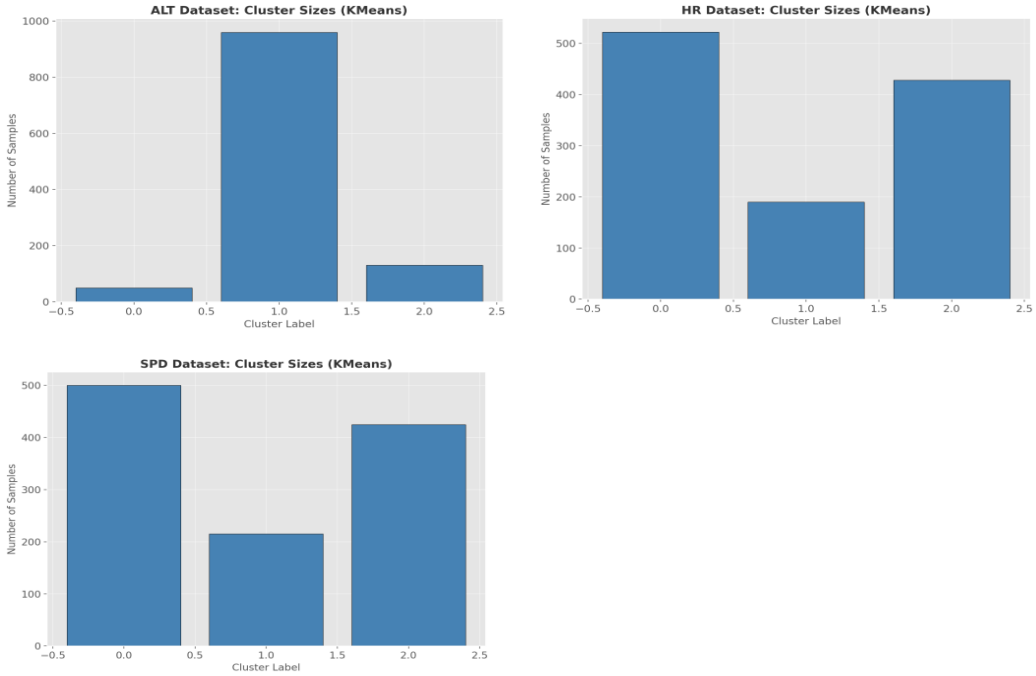

Figure 9. Distributions Across Sports Datasets.

SUMMARY AND CONCLUSION

The present work properly used the K-Means algorithm to analyse multivariate time series data captured from sports sensor data recordings, such as ALT, HR, and SPD. Using an efficient data preprocessing pipeline consisting of normalization, missing value imputation, and dimension reduction via PCA, clusters representing significance were formed and mapped to analyse intrinsic patterns in the data. The blockchain-based hashing for feature integrity at preprocessing also added another layer of traceability and data provenance, and this is extremely crucial in secure, large-scale sports analytics platforms.

The confusion matrices across the three datasets showed an apparent and regular pattern of separable clusters, particularly in the ALT dataset, with the highest intra-cluster cohesion and inter-cluster separation. This was further augmented by the silhouette coefficient distribution, with ALT having the highest positively skewed values up to 0.85, indicating strong cluster formations. The HR and SPD datasets showed moderately good clustering performance with lower average silhouette scores, indicating more overlap between clusters due to physiological variation or sensor noise.

PCA scatterplots provided a second window by which to view the clustering configuration, once again confirming the existence of intelligible groupings in the low-

dimensional representations. The ALT dataset again stood out, with evident clusters by eye, while the HR and SPD projections suggested a more dispersed spread with some overlaps, typical of the inherent complexity of biomechanical or cardiovascular signals.

Furthermore, the pattern of cluster sizes highlighted pervasive distinctions in natural segmentation of data. ALT data featured a strong dominant cluster, most likely reflecting a resting or baseline physiological state, whereas HR and SPD data showed more symmetrical distributions, possibly reflecting diverse modes of athletic activities or intensities. These differences point to the adaptability of KMeans across various feature spaces but also suggest that careful parameter tuning and perhaps hybrid clustering methods will be required for more challenging datasets.

The integration of machine learning with block chain concepts, though in its infancy in this study, promises directions for enhancing security and transparency of data. Through the construction of feature-level hashes before clustering, the method leaves an audit trail of input data, which is tamper-evident and enables trust in model outputs as well as secure collaborative analysis within institutions or groups.

Overall, the results of the present work not only validate the use of KMeans in anomaly detection and activity recognition in sports but also identify the importance of adequate data preprocessing and performance measurement. While KMeans offers ease and interpretability, further research can explore more sophisticated clustering algorithms such as DBSCAN, hierarchical clustering, or even deep learning-based clustering methods, especially when dealing with non-spherical or noisy data distributions. The positive results also present opportunities for real-time sports analytics applications, where automated clustering and anomaly detection may find application in performance tracking, injury prevention, and customized training, particularly if included within a blockchain-based data-sharing infrastructure.

CONFLICT OF INTERESTS

The authors state no conflict of interest.

REFERENCES

1. Saad, M. et al., "Exploring the attack surface of blockchain: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, **2020;** 22, 1977–2008.
2. S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," [Online]. Available: https://bitcoin.org/bitcoin.pdf (accessed on 16 October 2023).
3. Xie, M., Li, H. and Zhao, Y. Blockchain financial investment based on deep learning network algorithm. *Journal of Computational and Applied Mathematics*, **2020**; 372; 112723.
4. Sarker S., et al., "A survey on blockchain and cloud integration," in Proc. 23rd Int. Conf. Comput. Inf. Technol. (ICCIT), pp. 1–7, **2020**.
5. Q.-Q. Gan, R. You, and Lau, K. Trust in a 'trust-free' system: Blockchain acceptance in the banking and finance sector. *Technological Forecasting and Social Change*, **2024**; 199; 123050.

6. Zheng, Z. et al., An overview on smart contracts: Challenges, advances and platforms. *Future Generation Computer Systems*, **2020**; 105; 475–491.
7. Kose, J., Leonid, L. and Fahad, S. Smart contracts and decentralized finance. *Annual Review of Financial Economics*, **2023**; 15; 523–542.
8. Dong, C., Huang, Q. and Fang, D. Channel selection and pricing strategy with supply chain finance and blockchain. *International Journal of Production Economics*, **2023**; 265; 109006.
9. Boakye, E.A., Zhao, H. and Kwame Ahia, B.N. Emerging research on blockchain technology in finance: Conveyed evidence of bibliometric-based evaluations, *Journal of High Technology Management Research*, **2022**; 33; 100437.
10. Wang, T. et al., Health data security sharing method based on hybrid blockchain, *Future Generation Computer Systems*, **2024**; 153; 251–261.
11. Xiang, X. and Zhao, X. Blockchain-assisted searchable attribute-based encryption for e-health systems. *Journal of Systems Architecture*, **2022**; 124; 102417.
12. Uppal S. et al., HealthDote: A blockchain-based model for continuous health monitoring using interplanetary file system, *Healthcare Analytics*, **2023;** 3; 100175.
13. Tian, J., J.-F. Tian, and R.-Z. Du, MSLShard: An efficient sharding-based trust management framework for blockchain-empowered IoT access control. *Journal of Parallel and Distributed Computing*, **2024**; 185; 104795.
14. Dhar D. et al., Securing IoT devices: A novel approach using blockchain and quantum cryptography. *Internet of Things*, **2024**; 25; 101019.
15. Hameed K. et al., "A taxonomy study on securing blockchain-based industrial applications," Journal of Industrial Information Integration, **2022**; 26; 100312.
16. F.-M. Tseng, C.-W. Liang, and N.-B. Nguyen, "Blockchain technology adoption and business performance in large enterprises," Technology in Society, **2023**; 73; 102230.
17. Zhu X. et al., "Demand response scheduling based on blockchain considering the priority of high load energy enterprises. Energy Reports, vol. 9, pp. 992–1000, 2023.
18. P. Zhen et al., Blockchain-based decentralized application: A survey," IEEE Open Journal of the Computer Society, **2024**; 4; 121–133.
19. Banoth, R. and Dave, M.B. A survey on decentralized application based on blockchain platform," in *Proc. Int. Conf. Sustain. Comput. Data Commun. Syst. (ICSCDS)*, pp. 1171–1174, **2022**.
20. Tang H. et al., Learning to Classify Blockchain Peers According to their Behaviour Sequences. *IEEE Access*, vol. 6, pp. 71208–71215, 2018.
21. J. Matarmaa, "SportData MTS - 5," Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/jarnomatarmaa/sportdata-mts-5 (accessed on 16 October 2023).
22. MacQueen, J. Some methods for classification and analysis of multivariate observations. in Proc. *5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA, USA, pp. 281–297, **1967**.
23. S. Lloyd, S. Least squares quantization in PCM. Bell Labs Technical Note, Murray Hill, NJ, USA, **1957**.
24. Hartigan J.A. and Wong, M.A. A K-means clustering algorithm. *Applied Statistics*, **1079**; 28(1); 100–108.
25. Bradley P.S. and Fayyad, U.M. Refining initial points for K-means clustering. in *Proc. 15th Int. Conf. Machine Learning, Madison,* WI, USA, pp. 91–99, **1998**.
26. Jolliffe, I.T. Principal Component Analysis, 2nd ed. New York, NY, USA: Springer, **2002**.
27. Johnson R.A. and D. W. Wichern, Applied Multivariate Statistical Analysis, 5th ed. Upper Saddle River, NJ, USA: Prentice-Hall, **2002**.

28. Ding, C. and He, X. K-means clustering via principal component analysis. Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA, Tech. Rep. LBNL-53230, **2002**.
29. L. An and S. E. Ahmed, "Improving the performance of kurtosis estimator," Computational Statistics & Data Analysis, vol. 52, no. 5, pp. 2669–2681, Jan. 2008.
30. Maurya, V.N., Misra, R.B., Jaggi, C.K. and Maurya, A.K. Performance analysis of powers of skewness and kurtosis-based multivariate normality tests and use of extended Monte Carlo simulation for proposed novelty algorithm, *American Journal of Theoretical and Applied Statistics*, **2015**; 4(2–1); 11–18.
31. Hyvärinen A., and Oja, E. Independent component analysis: Algorithms and applications. *Neural Networks*, **2000**; 13(4–5); 411–430.
32. Scholz, M., Gibon, Y., Stitt, M. and Selbig, J. Independent component analysis of starch-deficient pgm mutants," in *Proc. German Conference on Bioinformatics, Gesellschaft für Informatik*, Bonn, Germany, pp. 95–104, **2004**.
33. Scholz, M., Gatzek, S., Sterling, A., Fiehn, O. and Selbig, J. Metabolite fingerprinting: Detecting biological features by independent component analysis. *Bioinformatics*, **2004**; 20(15); 2447–2454.
34. Reza, M.S., Nasser, M. and Shahjaman, M. An improved version of kurtosis measure and their application in ICA," *International Journal of Wireless Communication and Information Systems*, **2011**; 1(1); 34–42.
35. Ng, A.Y., Jordan, M.I. and Weiss,Y. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, **2001**; 14; 849–856.
36. Zha, H., Ding, C., Gu, M., He, X. and Simon, H. Spectral relaxation for K-means clustering. *Advances in Neural Information Processing Systems*, **2002**; 14; 1057–1064.