Journal of Engineering, Project, and Production Management
2026, 17(1), 2026-0017

Cluster Analysis of College Students' Physical Education Behavior Based on a Hybrid Algorithm of K-means and DBSCAN

Hailong Zhang
Associate Professor, Sport College, Xinxiang University, Xinxiang 453003 China,
E-mail: HailongZhangg@outlook.com

Project Management
Received April 22, 2026; revised June 10, 2026; accepted September 5, 2026
Available online September 10, 2026

Abstract: Insight into students' athletic behavior patterns is critical to efficient resource deployment and enhanced engagement within higher education. This study develops a clustering method for university student sports behavior based on a hybrid algorithm of K-means and Density-Based Spatial Clustering of Applications with Noise (DBSCAN). The hybrid algorithm first uses K-means for data pre-segmentation to identify dense core regions, then uses DBSCAN to fine-tune cluster boundaries. Experimental data come from a university's smart sports platform, covering the exercise behavior records of 5,000 to 10,000 students over an academic year, including features such as exercise frequency, single-session duration, and activity preferences. The hybrid algorithm achieved a silhouette coefficient of 0.71 and successfully identified four typical groups with 89.3% accuracy. Compared with the single K-means and DBSCAN algorithms, this hybrid method improves both clustering accuracy and noise handling capabilities. The research results provide data support for the design of personalized sports curricula and dynamic venue scheduling.

Keywords: K-means; density-based spatial clustering of applications with noise DBSCAN; college athletics; cluster analysis.

DOI 10.32738/JEPPM-2026-0017

1. Introduction

Under the background of the national strategy of "Healthy China" and the continuous deepening of the national fitness upsurge, college physical education, as a key position to improve young people's physical health level and cultivate lifelong sports awareness, has received increasing attention on its development trend and internal law (Ke et al., 2025). College students' sports participation behavior contains rich information about individual preferences, environmental influences, and the characteristics of campus sports culture. Accurately identifying and scientifically analyzing the sports participation mode, temporal and spatial distribution characteristics, project preference trend, and potential group structure contained in it is of great practical value for college sports management departments to optimize resource allocation (such as venue and facilities planning, curriculum setting and community activity guidance), implement accurate intervention (such as targeted promotion of specific group participation and design personalized exercise programs), improve service quality and promote the prosperity and development of campus sports culture (Wang and Zhang, 2021). As universities rapidly implement smart campuses, massive amounts of sports behavior data collected by various sports punch-in systems, venue reservation platforms, physical health test databases, and wearable devices have provided an unprecedented data basis for in-depth insights into campus sports trends (Rao, 2024). However, these data usually have the characteristics of high dimension, large scale, complex structure (such as time series, spatial location, category attribute interleaving), significant noise interference (such as sporadic motion, equipment false recording), and efficient and robust data mining technology is urgently needed for in-depth analysis and knowledge discovery (Chen et al., 2022). As a powerful unsupervised learning method, cluster analysis can automatically divide complex individual motor behavior data into several highly homogeneous groups (clusters) according to the intrinsic similarity among data objects without relying on prior labels, thus revealing hidden motion participation modes and group structure characteristics, which has become the core technical path to solve the above problems (Wu, 2025). Therefore, it is not only of theoretical importance to conduct fine clustering research on college sports data but also an urgent need to improve the level of scientific and intelligent decision-making in college sports work (Zhao and Sun, 2023).

Although clustering analysis has shown great power in many fields (Seo et al., 2025), when applied directly to complex and dynamic college sports data, traditional single-algorithm clustering often faces significant challenges. It is not easy to fully meet the actual needs (Zeng and Bao, 2025). Taking the most widely used K-means algorithm as an example (Tang et al., 2025), its principle is simple, its computational efficiency is high, and it performs well on data with spherical cluster distributions and moderate scale (Wu et al., 2025). However, its inherent defects in sports data analysis are fully revealed: firstly, the algorithm requires preset cluster count K, but lacks prior knowledge for determining optimal K in sports trend analysis. "positive", "general" and "negative", or are there more subdivided groups?), Subjective setting easily leads to the results deviating from the real structure; Secondly, K-means is highly sensitive to initial centroids, causing unstable/variable clustering results (Bíró et al., 2025); Critically, it assumes spherical, equally-sized clusters, and cannot effectively identify non-spherical clusters ubiquitous in real-life data (such as long-tail distribution groups with specific time preferences or project portfolio preferences) and cluster structures with significant density differences (such as high density of core sports enthusiasts and sparse distribution of occasional participants); Finally, K-means is extremely sensitive to noise points and outliers (Han et al., 2025), and there must be a large number of short-term, accidental, and abnormal participation records in motion data (such as physical assault training, temporary participation in fun activities). These "noises" will significantly distort the cluster center position and reduce the clustering quality (Nahoujy, 2025). Existing hybrid clustering studies, such as the K-means and DBSCAN Hybrid Algorithm (K-DBSCAN), primarily focus on using K-means to partition the data to improve DBSCAN's computational efficiency. This study proposes a different hybrid approach: first, K-means is used to pre-cluster the entire dataset to identify high-density core regions; then, DBSCAN is employed to refine the boundaries and remove noise from each coarse clustering result. This collaborative strategy aims to address K-means' sensitivity to initial centers and DBSCAN's difficulty with parameter selection in high-dimensional spaces, rather than simply improving computational speed.

DBSCAN handles arbitrary-shaped clusters without a preset K, identifies core points, filters out noise, and has great potential for spatial aggregation of sports data (e.g., popular venues during specific periods) and for identifying sparse groups (Mehmood and Wang, 2025). However, DBSCAN is not omnipotent: its performance depends on properly setting two key parameters: neighborhood size (Eps) and minimum density (MinPts). Improper parameter selection (especially when data density is unevenly distributed) can easily lead to "over-clustering" (incorrectly splitting a large cluster) or "under-clustering" (ignoring a cluster that should be identified) (Kim et al., 2024). At the same time, when faced with high-dimensional motion data (such as multi-dimensional features such as time, item, duration, frequency, intensity, location, etc. are considered at the same time), DBSCAN is susceptible to "dimensional disasters", the distance measurement fails, and the clustering effect drops sharply; In addition, it also faces difficulties when dealing with clusters with greatly different densities (such as identifying structures where high-density "runner enthusiast groups" coexist with extremely sparse "niche project attempter groups"). Therefore, it is difficult to comprehensively, consistently, and accurately describe the complex, diverse, and dynamically evolving modes and group structures of sports participation on university campuses by relying solely on K-means or DBSCAN.

To overcome the limitations of a single algorithm and effectively integrate the complementary advantages of different clustering paradigms, Hybrid clustering is key to analyzing complex data. Aiming at the core characteristics of college sports data, such as high dimension, heterogeneity, noise, and complex cluster structure (different shapes and uneven density), this study proposes and deeply explores a hybrid clustering method that creatively integrates the K-means algorithm and DBSCAN algorithm, aiming at constructing an analysis framework with high efficiency, robustness, adaptability, and explanation. The core idea of this hybrid strategy lies in giving full play to the synergistic effect of the two algorithms: firstly, using the fast convergence and dimensionality reduction capabilities of K-means, the high-dimensional motion data is preliminarily partitioned in the preprocessing stage or a specific subspace (even if the K value is not the final true cluster number), and its output results (such as preliminary cluster partition, cluster center point, sample-to-center distance) can provide key information guidance for subsequent steps. For example, analyzing the silhouette Coefficient or distance distribution in the K-means preliminary clustering results can more scientifically assist in determining the neighborhood radius (Epsilon, Eps) parameter range required by DBSCAN (Sheng et al., 2025). Alternatively, the high-density core region identified by K-means serves as a "seed point" to guide DBSCAN in performing more refined density expansion, thereby alleviating its sensitivity to parameters in high-dimensional space. Then, on this basis, DBSCAN's powerful noise recognition ability and non-spherical cluster discovery ability (Ueki, 2025) are introduced to optimize and refine the preliminary results of K-means deeply: it can effectively eliminate a large number of noise points (such as accidental motion records) contained in K-means results, and significantly improve the "purity" of clusters (Lu et al., 2025); More importantly, DBSCAN can break through the spherical hypothesis limitation of K-means (Bajpai et al., 2025), and further detect and identify fine substructures with arbitrary shapes and different densities within each coarse-grained cluster (or specific feature subspace) divided by K-means (for example, in a preliminary category of "ball enthusiasts", finer-grained patterns such as "fixed group of evening basketball", "social group of weekend badminton" and "fragmented table tennis practitioners" are identified), to more truly and finely reflect the internal differences and complexity of students' sports behavior. This hybrid mode of "first coarse scoring, then fine screening; first dimensionality reduction guidance, then density exploration" can theoretically solve the problems of K-means sensitivity to noise, shape limitation, and K value dependence, as well as the dilemma of DBSCAN in high-dimensional parameter selection (Wang et al., 2025), and realize multi-scale, multi-level and more accurate clustering insight into college sports trends (Li et al., 2025).

This study aims to answer the following three questions: First, does the hybrid algorithm improve clustering performance compared to a single algorithm? Second, what are the typical group structures in the sports behavior data of college students? Third, what are the actual effects of this hybrid method on sports interest identification and venue distribution analysis?

Based on the above analysis, this study focuses on "cluster analysis of college sports trends based on K-means and DBSCAN hybrid algorithm", aiming to provide powerful methodological tools and a decision-making basis for an in-depth

understanding of campus sports dynamics through algorithm innovation and empirical research. The main work of this paper is as follows:

(1) K-means is used for preliminary data partition to identify dense core areas.

(2) Using DBSCAN to finely optimize the cluster boundary, effectively overcoming the problems that traditional K-means are sensitive to the initial center, and DBSCAN is strongly dependent on parameters.

(3) The limitations of traditional single algorithms are addressed through fast initial clustering of K-means, combined with DBSCAN's recognition ability for complex distribution structures. Using data from smart sports platforms at colleges and universities, the experiment verifies the hybrid model's effectiveness in recognizing sports interests and analyzing venue distribution.

2. Cluster Analysis of College Sports Trends Based on K-Means and DBSCAN Hybrid Algorithm

2.1. Theoretical Basis of K-Means Algorithm

The k-means algorithm, a classical clustering method (Dai et al., 2025), aims to divide a dataset into k clusters via iterative optimization (Adjei et al., 2025), maximizing intra-cluster similarity and inter-cluster dissimilarity (Zaghloul et al., 2025). It minimizes the sum of within-cluster squared distances. The calculation process is shown in Eq. (1).

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} \| x - \mu_i \|^2 \tag{1}$$

Where C_i denotes the i-th cluster, μ_i denotes the center of the i-th cluster, x denotes the data point, and $\|x - \mu_i\|$ denotes the Euclidean distance from the data point to the cluster center. The K-means algorithm involves initialization, allocation, and update. It starts by randomly selecting A data point as an initial cluster center. Then, each data point is assigned to the nearest cluster center. and its calculation process is shown in Eq. (2).

$$Y_j^{'} = \frac{1}{|S_j|} \sum_{i \in S_j} X_i \tag{2}$$

Where, S_j denotes the set of all points assigned to cluster j and $|S_j|$ denotes the number of points in cluster j. Finally, repeat the steps until cluster centers stabilize or the iteration limit is reached. The k-means algorithm, a widely used clustering method in data mining and machine learning, involves key parameters such as the number of clusters (*k*), initial cluster centers, the number of iterations, and convergence criteria (Zhu et al., 2025). Clustering quality heavily depends on the preset *K*, which must be tuned per problem. Initial centers, random or *K*-means++ selected, affect speed and final partitions. Iteration proceeds until center stability is reached or the maximum number of steps is reached, minimizing intra-cluster distances.

2.2. DBSCAN

DBSCAN is a density-based clustering method that detects clusters of arbitrary shape and robustly handles outliers. Its operation depends on two parameters: a radius ε and a minimum point count MinPts. A point is designated as a core point if its ε-neighborhood contains at least MinPts other points. Boundary Points: Not core points but within the ε-neighborhood of a core point. Noise Points: Neither core nor boundary points.

DBSCAN uses concepts such as "directly density reachable" (a point within the ε-neighborhood of a core point) and "density reachable" (a chain of directly density reachable points) to form clusters. The choice of ε and MinPts significantly impacts clustering results. Specifically, the calculation for the core point is given by Eq. (3).

$$CP = \{P \| N_{\delta}(P)| \geq MinPts\} \tag{3}$$

Where CP denotes the core point and $N_{\delta}(P)$ denotes the set of points within the ∈-neighborhood of point P. Then, the calculation for the boundary point is given in Eq. (4).

$$BP = \{P \| N_{数}(P)| < MinPts\ AND\ \exists Q \in N\ (P)\ AND\ Q\} \tag{4}$$

Among them, BP represents the boundary point and Q represents the core point. At the same time, the calculation of the noise point is shown in Eq. (5).

$$NP = \{P \| N_{数}(P)| < MinPts\ AND\ \forall Q \notin N\ (P)\ OR\ Q\ \} \tag{5}$$

Where NP represents a noise point, and Q represents a non-core point. Finally, the calculation of direct density and reachability is given by Eq. (6).

$$DDUT = \{P, Q | Q \in N_{\delta}(P)\ AND\ P\} \tag{6}$$

Among them, DDUT means that the direct density is reachable, and P is the core point.

2.3. Trends in College Sports

As an important part of higher education, college sports strengthen students' physiques and cultivate their team spirit, competitive awareness and mental health. With growing social attention, e-sports is becoming increasingly diversified and specialized. It is now widely recognized by universities, which offer scholarships and build professional venues. Sports analytics has also become important, with many schools using data mining to help formulate training and competition strategies.

Cluster analysis, a data mining technology, groups similar objects into clusters and dissimilar ones into different clusters. In college sports, it identifies differences in students' participation, preferences and performance to support personalized teaching and promotion. For example, it can classify students by athletic performance data or analyze spectators' characteristics to aid targeted marketing.

Data mining in college sports focuses on three areas: 1) Information Assessment, which includes filtering sports-related data and excluding irrelevant ones. 2) Data Preparation and Modeling, such as cleaning data and selecting suitable clustering algorithms. 3) Result Evaluation and Strategy Implementation, including evaluating results with metrics and developing strategies to improve students' sports participation and performance.

2.4. Hybrid Model Based on K-Means and DBSCAN

A hybrid K-means and DBSCAN model is developed herein for analyzing college sports trends. Leveraging algorithmic refinements alongside empirical data, the framework offers methodological instruments and a conceptual basis for understanding campus sports dynamics. The system workflow, which includes preprocessing, clustering, and assessment stages, is schematized in Fig. 1.

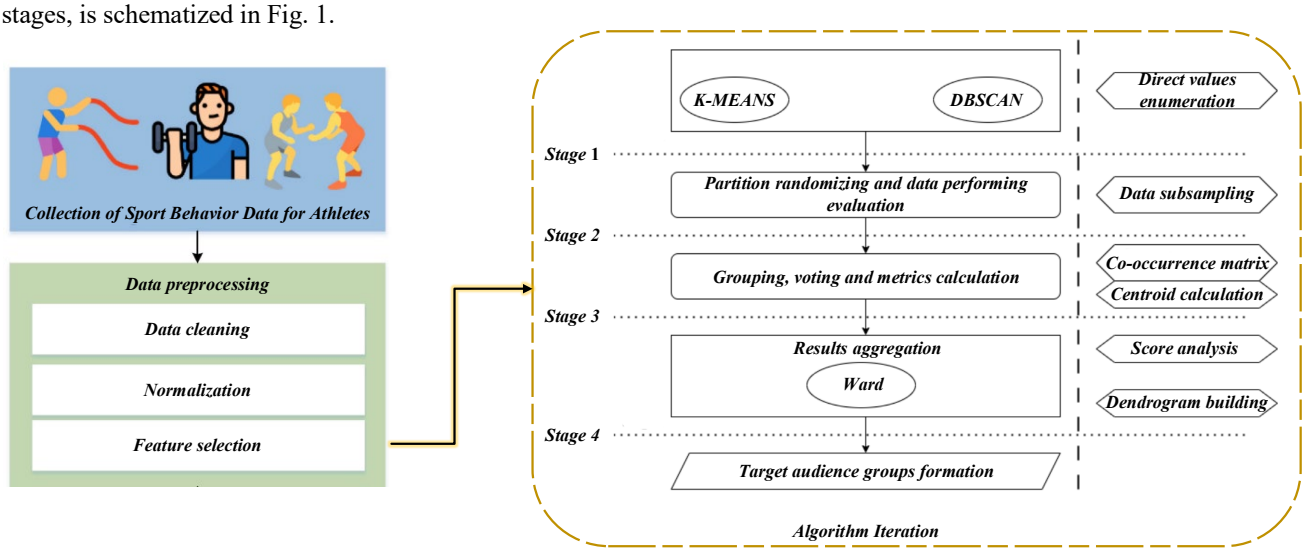


Fig. 1. Hybrid model based on K-means and DBSCAN

The preprocessing stage, encompassing cleaning, standardization, and dimensionality reduction, renders raw collegiate athletic data amenable to hybrid clustering. The core algorithm integrates K-means and DBSCAN: K-means furnishes rapid, coarse centroid initialization across the entire space, pinpointing high-density regions; DBSCAN then refines each coarse cluster as an independent subspace, thereby mitigating sensitivity to the global Eps parameter while preserving the capacity to detect arbitrary-shaped clusters. This synergy exploits K-means' scalability and DBSCAN's noise tolerance and shape flexibility, overcoming individual algorithmic deficiencies and enhancing the discrimination of latent trends in sports data. The procedural workflow is summarized in Algorithm 1.

Algorithm 1 K-means+DBSCAN Hybrid Clustering

Input: Dataset X, cluster number K, neighborhood radius ε , minimum points MinPts

Output: Final clusters C_{final} , noise points N

1: Initialize K cluster centers using K-means++

2: Assign each point in X to nearest center → coarse clusters C_{coarse}

3: For each coarse cluster *c* in C_{coarse} :

4: Apply DBSCAN on c with parameters ε and MinPts

5: Obtain refined subclusters and local noise points

6: End for

7: Collect all refined subclusters as C_{final}

8: Return C_{final} and N

The evaluation module's primary task is to assess the hybrid algorithm's clustering results and confirm its effectiveness and accuracy. Typical metrics used include the Silhouette Coefficient, Davies-Bouldin Index, and Adjusted Rand Index.

3. Experiment and Results Analysis

3.1. Dataset Description and Preprocessing

The experiment used a dataset from a comprehensive university that contained one academic year's exercise records for 5,000 to 10,000 students, ranging from freshmen to seniors. The students were enrolled in science, engineering, humanities, or arts programs. The raw data included exercise frequency, session duration, activity type, and time period. Missing entries were removed. Continuous features were normalized using Z-score transformation, and categorical features were encoded with one-hot encoding. The final feature set comprised basic attributes such as gender, grade and Body Mass Index (BMI), as well as dynamic behavioral attributes, including activity preference, weekly frequency, session duration, moderate-to-vigorous intensity ratio and time distribution. This preprocessing retained continuous numerical features suitable for K-means and spatiotemporal density features suitable for DBSCAN. The number of clusters for K-means was set to four based on the elbow method. For DBSCAN, the neighborhood radius ε was determined from the k-distance plot, and MinPts was set to five. The hybrid algorithm operated by first applying K-means to obtain coarse cluster centers, followed by DBSCAN to expand density within each subspace. This process identified mainstream exercise patterns, local subgroups and abnormal behaviors. Clustering performance was evaluated using the silhouette score, Davies-Bouldin index, Calinski-Harabasz index, and adjusted Rand index. Generalization of the method to data from other universities requires further cross-validation. Specifically, the silhouette coefficient is an evaluation metric for clustering performance that integrates intra-cluster cohesion and inter-cluster separation. The calculation of the silhouette Coefficient is given by Eq. (7).

$$Silhouette\ Score = \frac{b-a}{max(a,b)} \tag{7}$$

Where a is the average intra-cluster distance and b is the average distance to the nearest other cluster. The Silhouette Coefficient ranges from -1 to 1, with higher values indicating better clustering. The Davis-Bourdin Index was used to assess the separation and compactness of the clusters. The calculation process is shown in Eq. (8).

$$DBI = \frac{1}{n}\sum_{i=1}^{n} max_{j\neq i}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right) \tag{8}$$

Where n is the number of clusters. σ_i and σ_j are the standard deviations of clusters i and j, respectively, and $d(c_i, c_j)$ is the distance between cluster centers c_i and c_j. The lower the Davis-Bouldin Index, the better the clustering effect. The Calinski-Harabasz index is a measure of clustering effectiveness that evaluates cluster quality by comparing intra-class dispersion with inter-class dispersion. The calculation process is shown in Eq. (9).

$$CHI = \frac{Tr(B_k)}{Tr(W_k)} \times \frac{n-k}{k-1} \tag{9}$$

Where $Tr(B_k)$ represents the trace of the inter-class dispersion matrix, $Tr(W_k)$ represents the trace of the intra-class dispersion matrix, n is the total number of data points, and k is the number of clusters. The higher the Calinski-Harabasz index, the better the clustering performance. Accuracy is an important metric for measuring the performance of classification models, representing the proportion of correct predictions to the total number of predictions. The calculation process is shown in Eq. (10).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

Among them, TN stands for true negatives, the count of actually negative samples correctly predicted as negative by the model. FN signifies false negatives, the count of actual positive samples misclassified as negative. Calculating these evaluation metrics provides a comprehensive understanding of the strengths and weaknesses of different algorithms in terms of recommendation accuracy and comprehensiveness. In addition to K-means and DBSCAN, this experiment also compared Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) with Gaussian mixture models. K-DBSCAN outperformed HDBSCAN and the Gaussian Mixture Model (GMM) in silhouette coefficients. Paired t-tests were used to compare silhouette coefficients from five independent runs, and the differences between K-DBSCAN and K-means and DBSCAN were statistically significant.

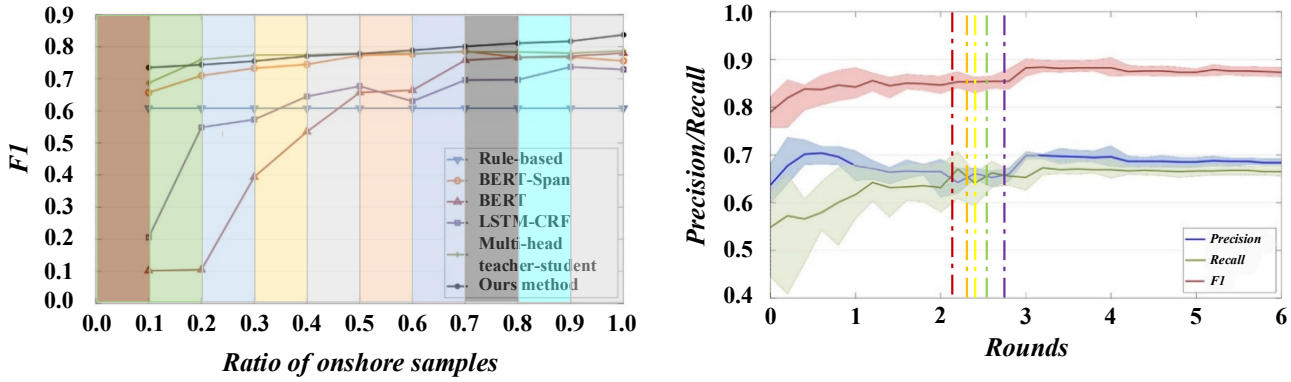
3.2. Comparative Performance of Clustering Algorithms

Table 1 shows that K-DBSCAN's silhouette coefficient is 0.71, higher than K-means' 0.62 and DBSCAN's 0.58. K-DBSCAN's Davis-Bourdin index is 0.89, which is lower than that of the other two algorithms. K-means detected no noise points, DBSCAN had a noise point ratio of 12.7%, and K-DBSCAN had 4.2%. In terms of clustering time, K-means was the shortest, with K-DBSCAN falling in between. Overall, K-DBSCAN demonstrates a balanced performance in clustering quality and noise reduction across all metrics.

Table 1. Performance metrics of clustering algorithms

Method	Silhouette Coefficient	DBI (Davies-Bouldin Index)	Proportion of noise points	Clustering time (s)
k-means (k = 4)	0.62	1.35	0%	3.2
DBSCAN	0.58	1.42	12.7%	8.5
K-DBSCAN	0.71	0.89	4.2%	6.8

Fig. 2 shows the F1 score and accuracy for different algorithms. The left figure evaluates the F1 score between the cluster boundary and the true label, while the right figure calculates the accuracy through classifier training to verify the stability of intra-cluster features. Analysis shows that the algorithm presented in this paper has a significant advantage in the aforementioned metrics.



(a) F1 score (b) Accuracy for different algorithms

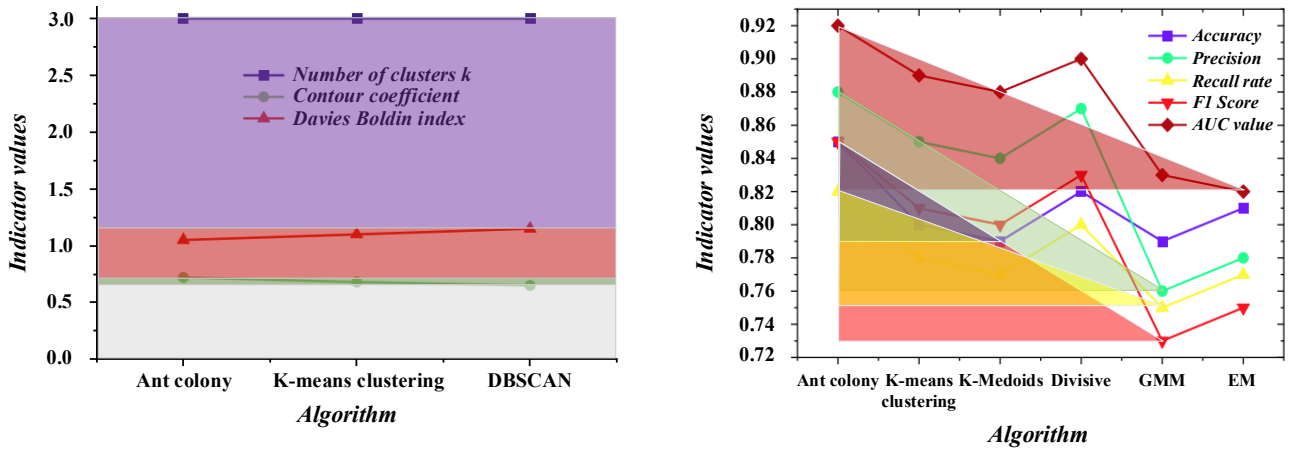
Fig. 2. Comparison of F1 scores and accuracy of different algorithms

3.3. Ablation Study and Parameter Sensitivity Analysis

Fig. 3 shows the evaluation indicators of different models. Analysis indicates that the mixed model proposed in this paper is significantly better than the comparative model across all indicators, particularly the Area Under the Curve (AUC), which has increased by more than 20%.

Fig. 4 shows the experimental results of the algorithm in this paper without samples. Analysis indicates that the algorithm in this paper can be classified effectively. Fig. 5 shows the experimental results of this algorithm using different sample points and different training steps. Analysis indicates that this algorithm performs well.

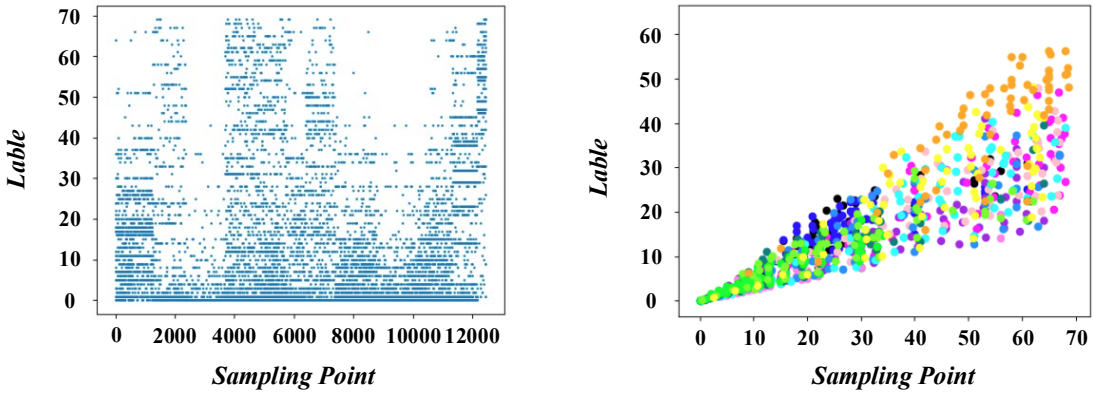
Table 2 uses the controlled variable method to obtain a baseline accuracy of 54.2% (Silhouette Coefficient converted to a percentage) with the complete feature set, and calculates the decrease in accuracy after sequentially removing the three types of features. The results show that semantic features contribute the most to the model, while interaction features also play an important role in optimizing the understanding of user behavior.



(a) Indicator values (I) (b) Indicator (II)

Fig. 3. Evaluation metrics for anomaly detection models

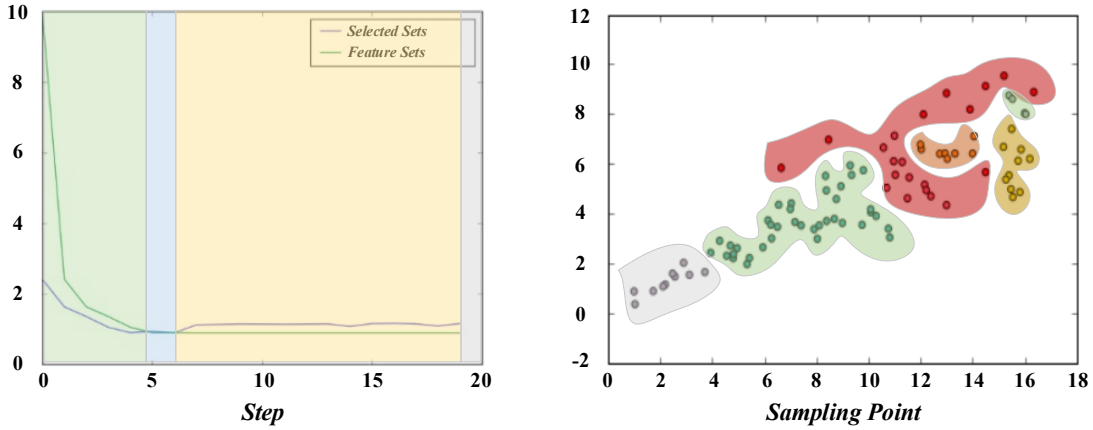
Note: Receiver Operating Characteristic (ROC) curves and AUC scores are used to evaluate algorithm performance. Noise points in clusters are treated as "outliers," while normal points are considered "normal," framing noise identification as a binary classification problem.



(a) Clustering output distribution (I) (b) Clustering output distribution (II)

Fig. 4. Clustering output distribution without additional sample points

Note: Different colors represent different clusters, and gray dots represent noise. The algorithm successfully identifies four clear clusters and recognizes outlier records without additional sample points.



(a) Clustering output distribution (I) (b) Clustering output distribution (II)

Fig. 5. Algorithm performance with varying numbers of initial sample points

Note: The figure compares silhouette coefficients for initial sample sizes of 50, 100, and 200.

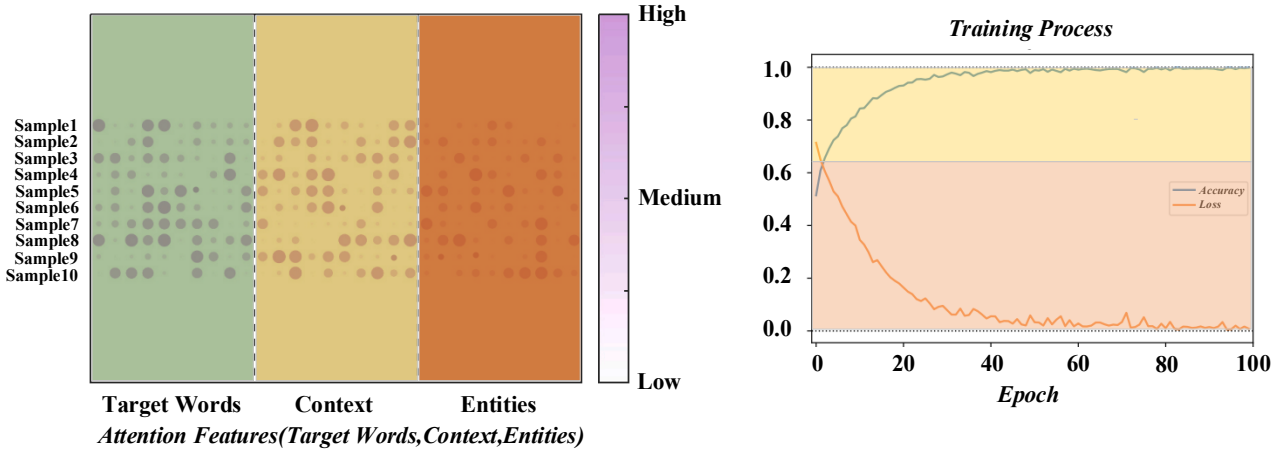
Fig. 6 defines the number of K-means iterations as the training epochs, demonstrating the algorithm's performance at different training epochs. The silhouette coefficient is recorded every 5 epochs starting from the 5th epoch. After the 25th

epoch, the coefficient stabilizes at 0.71, indicating that the algorithm achieves significant results by this point.

Fig. 7 shows the clustering results for the real and predicted labels during this algorithm's classification. The analysis concludes that the algorithm in this paper achieves good classification performance.

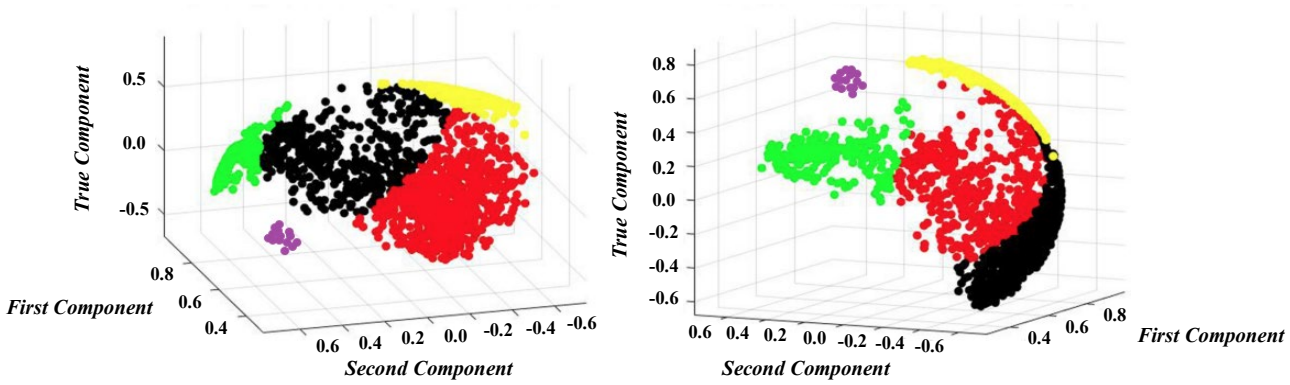
Table 2. Precision impact of removing different features on the model

Feature removal	Accuracy	Magnitude of influence
Remove test feedback features	50.9%	4.3%
Remove user temporal behavior features	50.3%	5.8%
Remove K-Means semantic features	47.2%	9.0%



(a) Attention features (b) Convergence trend

Fig. 6. Performance of different training epochs



(a) Real labels (b) Predicted labels

Fig. 7. Clustering quality evaluation

Cluster labels are used as pseudo-labels to train a simple Long Short-Term Memory (LSTM) classifier. The horizontal axis shows manually labeled real labels, the vertical axis shows cluster-mapped labels, and diagonal concentration reflects consistency between clusters and the true distribution. This evaluation assesses clustering quality and does not directly involve Long Short-Term Memory Conditional Random Field (LSTM-CRF) clustering. Fig. 8 shows that the algorithm in this paper uses the experimental data of K-Means clustering. Analysis indicates that the algorithm in this paper achieves good results. To validate the conclusion, the clustering procedure was repeated on two additional semester sub-datasets. The silhouette coefficients obtained by K-DBSCAN on these sub-datasets were 0.69 and 0.70. The differences from the main experiment result were below 0.03. This outcome demonstrates the method's performance consistency across data sampled from different time periods.

4. Discussion

Based on the clustering results, management decisions can be implemented for different student groups by adjusting short-duration activity space quotas during morning and midday hours for low-intensity clusters, allocating dedicated training sessions for specialized exercise groups, and modifying reservation strategies for shared facilities according to the

active time periods of socially active clusters. The theoretical contribution of this study is a hybrid mechanism that combines K-means and DBSCAN, preserving DBSCAN's ability to identify non-spherical clusters while reducing its sensitivity to parameters in high-dimensional spaces. The practical applications include dynamic scheduling of university sports facilities, design of student exercise intervention strategies, and personalized course recommendation systems. Several limitations exist. The data originate from a single university, limiting sample representativeness. The parameters ε and MinPts in DBSCAN require dataset-specific tuning and do not achieve full-scenario adaptivity. When the data volume exceeds 50,000 records, the algorithm runtime increases notably. The method's generalizability to non-campus physical activity data has not been validated.

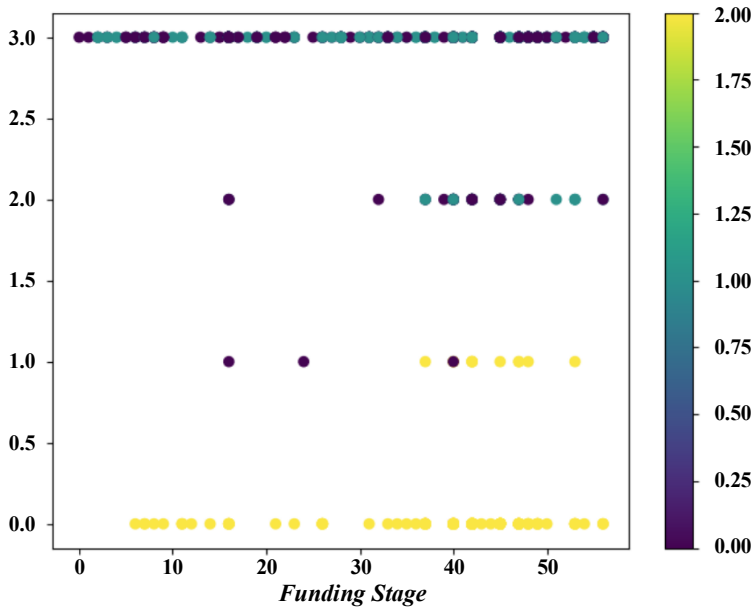


Fig. 8. Experimental results of K-Means clustering

5. Conclusion

The hybrid method integrates K-means for initial data division and dense region identification with DBSCAN for cluster boundary refinement. This combination compensates for K-means' sensitivity to initial centers and DBSCAN's dependence on parameters. Validation on a university smart sports platform dataset addresses three research questions. First, the hybrid algorithm improves the silhouette coefficient and noise handling relative to single algorithms. Second, the method identifies four typical student groups. Third, the approach demonstrates validity in both sports-interest recognition and venue-distribution analyses. Compared with K-DBSCAN, the proposed method targets a different optimization objective: cluster boundary precision. The clustering results are applicable to course design and venue scheduling. Further validation is required to assess the method's generalizability across other universities or behavioral datasets. The student physical activity data used in this study came from a university's smart sports platform. All data was anonymized before analysis and contained no personal information. According to the institution's ethics committee, using fully anonymized and untraceable behavioral data for algorithmic research does not require Institutional Review Board (IRB) review. The authors declare that AI tools were used solely for language fluency polishing and grammar checking during manuscript preparation, and were not applied to research design, algorithm development, data analysis, result interpretation, or conclusion formation.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

Declaration of Artificial Intelligence (AI) Tools

The authors used DeepSeek only for language editing and formatting assistance. The authors reviewed and take full responsibility for all content.

References

Adjei, S., Takyi-Amponsem, D. Y., Quaye, J. A., Sokama-Neuyam, Y. A., Oleka, J. O., Asigbi, G. K., and Duodu, T. K. (2025). Automated hydraulic flow unit determination using FZI-Z-score probability, K-means clustering with autoencoders, and machine learning classification: A case study of the Subei Basin, China. *Journal of Applied Geophysics*, 240, 105778.

Bajpai, N., Paik, J. H., and Sarkar, S. (2025). Balanced seed selection for K-means clustering with determinantal point process. *Pattern Recognition*, 164, 111548.

Bíró, P., Kovács, B. B. H., Novák, T., and Erdélyi, M. (2025). Cluster parameter-based DBSCAN maps for image characterization. *Computational and Structural Biotechnology Journal,* 27, 920-927.
Chen, X., Li, H., and Liu, S. (2022). Mining massive and noisy student exercise data from smart campus platforms: A density-based approach. IEEE Transactions on Learning Technologies, 15(4), 489-502.
Dai, Y., Yang, L., and Cao, Y. (2025). Long baseline underwater source localization based on deep K-Means++ clustering in complex underwater environments. *Digital Signal Processing,* 164, 105281.
Han, Y. J., Cao, M. M., and Peng, Y. F. (2025). New approach for quality function deployment based on bounded trust consensus model and DBSCAN algorithm. *Expert Systems with Applications,* 128528.
Ke, W., Peng, L., and Wang, L. (2025). Effects of college students' perceived transformational leadership of physical education teachers on their exercise adherence: Serial-mediated roles of physical self-efficacy and exercise motivation. *Acta Psychologica,* 254, 104878.
Kim, J., Lee, H., and Ko, Y. M. (2024). Constrained density-based spatial clustering of applications with noise (DBSCAN) using hyperparameter optimization. *Knowledge-Based Systems,* 303, 112436.
Li, J., Wang, L., Fu, S., Fu, W., and Pan, X. (2025). Self-labeled framework with semi-supervised ball K-means clustering-based synthetic example generation for semi-supervised classification in industrial applications. *Engineering Applications of Artificial Intelligence,* 150, 110528.
Lu, J., Luo, T., and Li, K. (2025). A forward k-means algorithm for regression clustering. *Information Sciences,* 711, 122105.
Mehmood, Z., and Wang, Z. (2025). Hybrid iForest-DBSCAN for anomaly detection and wind power curve modelling. *Expert Systems with Applications,* 289, 128381.
Nahoujy, M. R. (2025). Applying a K-means model to TSD data to find categories for the structural assessment of flexible pavements. *Transportation Engineering,* 20, 100342.
Wang, L., and Zhang, Y. (2021). Optimizing campus sports resource allocation based on student participation behavior clustering. *Journal of Physical Education and Sport Management,* 12(3), 45-58.
Rao, W. (2024). Design and implementation of college students' physical education teaching information management system by data mining technology. *Heliyon*, 10(16), e36393.
Seo, K., Na, H. S., Lee, W., Chen, C. B., Kweon, S. J., Zhao, L., and Kumara, S. (2025). Clustering electricity consumption patterns using functional data analysis. *Sustainable Energy, Grids and Networks,* 43, 101742.
Sheng, H., Huang, Z., Ke, L., Zhang, J., Zeng, Z., Yasir, M., and Liu, S. (2025). Multi-scale vessel trajectory clustering: An adaptive DBSCAN method for maritime areas of diverse extents. *Ocean Engineering,* 334, 121461.
Tang, Z. A., Duan, X. F., Liang, R. H., and Ding, Y. (2025). Efficient multi-party privacy preserving federated k-means based on homomorphic encryption. *Information Sciences, 717*, 122335.
Ueki, M. (2025). A deflation-adjusted Bayesian information criterion for selecting the number of clusters in K-means clustering. *Computational Statistics & Data Analysis*, 209, 108170.
Wang, J., Dong, S., Li, Z., Qu, J., Kuai, Y., Tan, D., and Luo, J. (2025). The links between Na/K ratios in eolian sands and mean annual precipitation in the deserts of arid region, northern China. *CATENA*, 258, 109219.
Wu, C., Yang, Z., and He, S. (2025). Efficient modal parameter identification using DMD-DBSCAN and rank stabilization diagrams. *Aerospace Science and Technology,* 161, 110112.
Wu, M. (2025). Research and application of optimization of physical education training model based on multi-objective differential evolutionary algorithm. *Systems and Soft Computing*, 7, 200200.
Zaghloul, M., Barakat, S., and Rezk, A. (2025). Enhancing customer retention in online retail through churn prediction: A hybrid RFM, K-means, and deep neural network approach. *Expert Systems with Applications,* 290, 128465.
Zeng, S., and Bao, J. (2025). Analysis of the effects of digital transformation of enterprise clusters on innovation performance in the context of "Internet+". *Systems and Soft Computing,* 7, 200270.
Zhao, Q., and Sun, W. (2023). Fine-grained clustering analysis for university sports behavior data: Towards scientific decision support. *Journal of Sports Sciences*, 41(2), 112-125.
Zhu, K., Cao, M., Ostachowicz, W., and Radzienski, M. (2025). K-means aided broadband ultrasonic guided wavefield processing for damage detection in composite panels. *Mechanical Systems and Signal Processing,* 231, 112696.

Zhang Hailong obtained her master's degree in physical education and training from Henan Normal University in 2006. She is currently an associate professor in the School of Physical Education at Xinxiang University. Her areas of interest include physical education and training, public sports services, and social sports.