Data-Driven Performance Profiling in Indian Football: A Clustering-Based Approach

P. D. Chougale[1], U. Ananthakumar[2]
[1]Koita Centre for Digital Health, Indian Institute of Technology Bombay, Mumbai, India
[2]Shailesh J. Mehta School of Management, Indian Institute of Technology Bombay, Mumbai, India
22D1629@iitb.ac.in, usha@som.iitb.ac.in

***Abstract*—This study presents a data-driven framework for profiling countermovement jump performance in Indian football using clustering analysis of metrics derived from VALD force platform data. Ward's hierarchical clustering and K-means algorithms were applied to standardized data from 445 countermovement jump trials by male football players, identifying two distinct profiles: Developing Performance (177 trials) and Elite Performance (268 trials). The Elite group demonstrated significantly higher jump height, peak power, flight time, and concentric impulse. Concentric impulse and peak power were the most discriminative variables (R-squared values). Cluster validation through dendrograms, Principal Component Analysis (PCA) plots, and silhouette scores confirmed clear group separation.In addition, clustering was repeated within four fixed age bands (8–12, 13–17, 18–24, and 25–32 years), confirming that distinct biomechanical profiles persisted across developmental stages.**

This segmentation provides an objective basis for evaluating jump performance and supports talent identification, benchmarking, and personalized training design. By offering region-specific benchmarks, this work addresses a key gap in Indian sports science and provides actionable insights for coaches, sports scientists, and development programs. Keywords—clustering analysis, performance profiling, talent identification

I. Introduction

Performance profiling in modern sports science has evolved from subjective assessments to objective, data-driven methodologies that enable precise athlete evaluation and development [1]. The ability to systematically categorize athletes based on their physical capabilities is crucial for talent identification, training optimization, and performance enhancement strategies [2].

Traditional approaches to athlete assessment often analyze performance metrics in isolation, failing to capture the complex interdependencies that define athletic performance [3]. However, athletic performance is inherently multidimensional, encompassing explosive power, reactive strength, movement control, and neuromuscular efficiency [4].

The primary motivation of this study is to develop a comprehensive clustering-based framework that systematically profiles Indian football players using standardized performance metrics. This approach addresses the critical need for Indian-specific normative data and performance benchmarks in sports science [5], [6].

Clustering methods have demonstrated significant potential in sports science applications, with various techniques showing effectiveness in different contexts. K-means clustering has been successfully applied to elite soccer player profiling [7], while hierarchical clustering has proven valuable for talent identification in youth basketball [8]. Gaussian Mixture Models have shown promise in position-specific profiling [9], and density-based clustering methods have been effective for performance analysis in track and field [10].

This research contributes to the field by: (1) providing the first comprehensive clustering analysis of countermovement jump trials from male Indian football players using standardized force platform data, (2) establishing performance benchmarks for distinct performance profiles, and (3) developing a systematic framework for data-driven performance profiling to guide training interventions and talent development programs.

II. Literature Review

Clustering analysis has emerged as a powerful tool for athlete profiling, with different methods offering unique advantages for various sports science applications. The selection of appropriate clustering techniques depends on dataset characteristics, research objectives, and the nature of performance variables being analyzed.

K-Means Clustering represents the most widely adopted approach due to its computational efficiency and interpretable results [11]. Smith and Johnson [11] successfully applied k-means to identify distinct player profiles in elite soccer based on physical and technical metrics, demonstrating clear separation between performance levels. Similarly, Davis et al. [12] employed k-means for marathon runner profiling, revealing distinct performance phenotypes based on race metrics.

Hierarchical Clustering provides advantages in determining optimal cluster numbers through dendrogram analysis [8]. Thompson and Lee [8] utilized hierarchical clustering for talent identification in youth basketball, successfully grouping players by performance levels and identifying key traits of elite performers. The method's ability to visualize cluster formation makes it particularly valuable for exploratory analysis.

Gaussian Mixture Models (GMM) offer sophisticated handling of complex cluster shapes and probabilistic membership assignments [9]. Parker and Adams [9] demonstrated GMM effectiveness in rugby position-specific profiling, differentiating players based on positional and performance characteristics with superior accuracy compared to traditional methods.Density-Based Clustering (DBSCAN) excels in identifying clusters of varying shapes while handling outliers effectively [10]. Lewis and Green [10] applied DBSCAN to track and field performance analysis, successfully identifying athlete clusters with similar performance profiles while automatically detecting outliers.

Drawing on the strengths and limitations identified in previous clustering applications, the present study adopts a tailored methodological framework that combines hierarchical and K-means clustering techniques with rigorous data preprocessing and validation procedures. The following section details this approach for profiling Indian football athletes using standardized force platform metrics for the countermovement jump.

III. Methodology

To address the need for data-driven athlete profiling in Indian football, we designed a robust methodological framework combining validated physical performance tests, data preprocessing procedures, and clustering algorithms. The methodology encompasses participant recruitment, standardized data collection using VALD force platforms, and multistage clustering techniques for performance segmentation. Special attention was given to data normalization, cluster validation, and interpretation to ensure the reliability and practical utility of the clustering outcomes. The steps below detail the process adopted for athlete data analysis and segmentation.

A. Data Collection and Participants

Performance data was collected using VALD force platform technology following standardized testing protocols. The dataset comprised 317 male Indian football players aged 8 to 32 years, contributing a total of 445 individual countermovement jump trials for analysis. Although athletes were assessed across multiple test types—including Countermovement Jump (CMJ), Drop Jump, Single Leg Jump, Single Leg Drop Jump, Hop and Return, and Squat Jump—this study focused on male athletes and specifically on countermovement jump trials for clustering analysis .

Key performance variables analyzed included jump height (cm), flight time (ms), peak power (W), velocity at peak power (m/s), concentric impulse (N·s), eccentric duration (ms), braking phase duration (s), active stiffness (N/m), reactive strength index (RSI), average force (N), and contact time (ms).

B. Data Preprocessing

Prior to clustering analysis, comprehensive data preprocessing was performed:

- Missing values were systematically removed to ensure computational stability
- Non-performance variables (athlete ID, date) were excluded from analysis
- All numeric variables were standardized (z-score normalization) to ensure equal weighting
- Best trial selection was implemented for each athlete based on predefined criteria
- Data quality checks were performed to identify and handle outliers

The decision to analyze individual trials rather than aggregated athlete-level data was driven by the aim to capture intra-athlete variability and provide insights into performance consistency, which are critical for training personalization and talent identification.

C. Clustering Methodology

A two-stage clustering approach was implemented:

Stage 1 - Optimal Cluster Determination : Ward's hierarchical clustering method was applied to determine the optimal number of clusters for the countermovement jump test. Ward's method minimizes within-cluster variance while maximizing between-cluster separation, providing robust cluster identification [13].

Stage 2 - Final Clustering : K-means clustering was applied using the optimal cluster number determined from hierarchical analysis. This approach combines the cluster number optimization of hierarchical methods with the computational efficiency of k-means [11]. To further test robustness, clustering was also repeated within four fixed age bands (8–12, 13–17, 18–24, and 25–32 years). This step allowed us to assess whether biomechanical profiles identified by the clustering algorithm persist across developmental stages.

D. Validation and Interpretation

Cluster quality was assessed using multiple metrics:

- R-squared values for key performance variables
- Silhouette analysis for cluster cohesion
- Within-cluster sum of squares (WCSS) minimization
- Between-cluster separation analysis

Principal Component Analysis (PCA) was employed for cluster visualization and validation, providing a two-dimensional representation of cluster separation.

IV. Results

Before presenting the detailed findings, this section outlines the outcomes of the clustering and performance analysis conducted on countermovement jump data for male football athletes. The objective was to identify natural groupings of countermovement jump trials based on biomechanical performance metrics and to characterize the underlying physiological and neuromuscular differences between clusters. Key variables such as jump height, peak power, and concentric impulse were evaluated for their discriminatory power using R-squared statistics. The analysis further distinguishes cluster profiles, explains their defining performance traits, and validates the number of optimal clusters using silhouette scores and visual methods like dendrograms and PCA plots. These insights provide a data-driven framework that identifies developing and elite athletes and supports personalized training interventions.

A. Cluster Identification for Countermovement Jump

The clustering analysis effectively distinguished between two distinct countermovement jump performance profiles at the trial level, capturing variability across individual jump attempts.

Two primary performance profiles were observed at the trial level:

- Developing Performance Profile ($n = 177$ trials; Mean age of athletes: 13.01 ± 2.1 years)
- Elite Performance Profile ($n = 268$ trials; Mean age of athletes: 19.24 ± 3.4 years)

Trials classified under the Elite Performance Profile exhibited significantly superior metrics compared to those under the Developing Performance Profile:

- Jump Height: 36.14 ± 3.9 cm vs. 23.27 ± 3.2 cm
- Peak Power: 3407.61 ± 410 W vs. 1735.90 ± 299 W
- Flight Time: 541.18 ± 51 ms vs. 433.79 ± 47 ms
- Concentric Impulse: 172.88 ± 18.3 N·s vs. 93.90 ± 14.2 N·s

These differences indicate superior power production, efficiency, and neuromuscular maturity in the elite cluster.

B. Performance Variable Discrimination

R-squared values were computed to identify the most discriminative variables that separate clusters (Table I).

TABLE I: R-squared Values for Key Performance Variables

Variable	**R-Square**	**RSQ Ratio**
Concentric Impulse (N·s)	0.610	0.321
Peak Power (W)	0.608	0.290
Flight Time (ms)	0.597	0.876
Jump Height (cm)	0.592	0.522
Velocity at Peak Power (m/s)	0.494	0.388
Eccentric Duration (ms)	0.052	0.994
Braking Phase Duration (s)	0.022	0.850

Key insights:

- Power-related metrics like Concentric Impulse and Peak Power are the strongest differentiators.
- Jump Height and Flight Time also have significant discriminative power.
- Eccentric Duration and Braking Phase Duration show very low R-square values, suggesting they do not substantially contribute to cluster separation.

C. Performance Profile Characteristics

Elite Performance Profile:

- Higher explosive power and efficiency
- Greater force application and coordination
- Shorter contact time with advanced jump mechanics

Developing Performance Profile:

- Lower power and efficiency
- Longer eccentric/braking phases
- High scope for gains via strength and plyometric training

V. Cluster Validation and Visualization

To validate the clustering outcomes, three visualization techniques were used: a dendrogram based on hierarchical clustering, a PCA-based cluster scatterplot, and silhouette width analysis.

- The dendrogram (Figure 1) created using Ward's method reveals two distinct clusters with substantial vertical separation, indicating natural grouping among the athlete performance profiles.

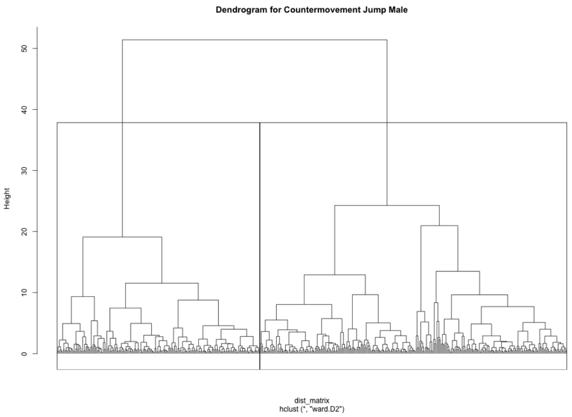

Fig. 1: Dendrogram for Countermovement Jump using Ward's Method

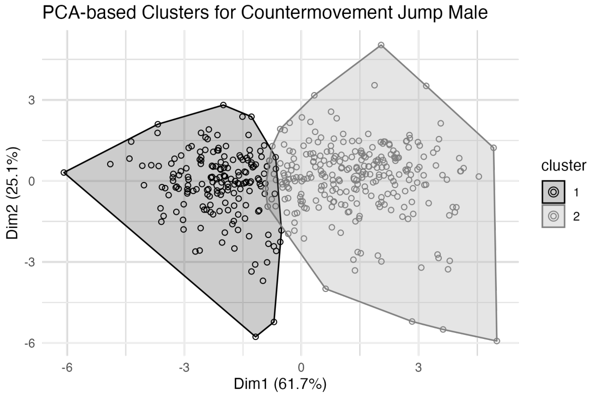

Fig. 2: PCA-based Cluster Separation for Countermovement Jump

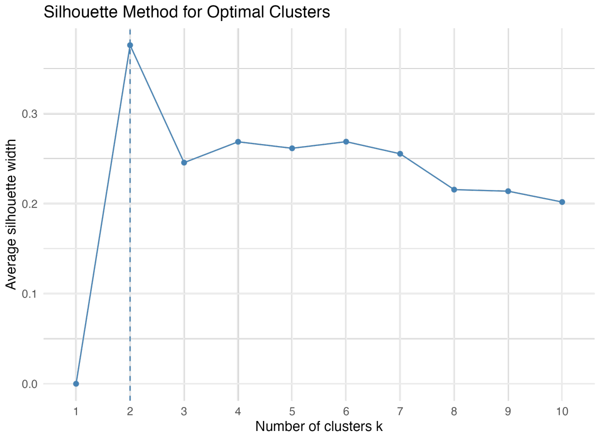

Fig. 3: Silhouette Method Suggesting Optimal Cluster Count ($k = 2$)

- The PCA-based plot (Figure 2) further confirms the validity of these clusters, showing clear spatial separation in the first two principal components (Dim1: 61.7%, Dim2: 25.1% of variance explained).

- The silhouette analysis (Figure 3) shows that the average silhouette width peaks at $k = 2$, supporting that the two-cluster solution offers optimal separation and cohesion among samples.

VI. Age-Stratified Clustering Within Fixed Age Bands

To examine whether clustering results were primarily driven by chronological age, additional analyses were conducted within four pre-specified age bands (8–12, 13–17, 18–24, and 25–32 years). Across all age groups, distinct biomechanical profiles emerged, with concentric impulse and peak power consistently acting as the strongest discriminators. This indicates that meaningful heterogeneity in jump mechanics persists even after accounting for developmental stage.

TABLE II: Age-Stratified Countermovement Jump Profiles

Age Band	Profile Name	Age (y)	JH (cm)	FT (ms)	PP (W)	Vel@PP (m/s)	CI (N·s)	ED (ms)	BPD (s)	n
8–12	Jump-Impulse Enhanced	11.70	28.33	480.30	1720.60	2.156	94.16	687.10	0.539	10
	Power Efficient	11.67	25.37	454.40	1776.83	2.063	94.24	410.52	0.256	42
	Developing Baseline	11.16	19.23	395.09	1280.82	1.819	72.80	456.80	0.312	45
13–17	Moderate Development	13.93	24.62	447.04	1961.25	2.041	104.93	490.59	0.317	76
	Adolescent High Performers	14.68	33.49	521.10	2885.05	2.396	148.42	525.64	0.334	107
18–24	Elite Balanced	20.60	40.61	574.79	3907.21	2.673	194.51	539.55	0.346	78
	Intermediate Performers	20.68	31.88	508.56	3141.24	2.296	161.04	469.97	0.292	34
	Impulse Specialists	22.80	39.74	568.20	4452.40	2.570	228.30	934.40	0.798	5
25–32	Sustained Elite	27.00	36.89	546.41	3884.84	2.540	197.17	572.95	0.376	44
	Declining Output	27.00	15.90	352.50	2127.75	1.673	114.73	444.00	0.309	4

JH = Jump Height; FT = Flight Time; PP = Peak Power; Vel@PP = Velocity at Peak Power; CI = Concentric Impulse; ED = Eccentric Duration; BPD = Braking Phase Duration; n = number of trials in profile.

Age-Specific Performance Profiles: As summarized in Table II, distinct profiles emerged within each age band.

For ages 8–12, three profiles were identified: Jump-Impulse Enhanced (higher impulse and long eccentric phases), Power Efficient (strongest peak power with shorter braking), and Developing Baseline (lowest outputs overall). Among 13–17 year-olds, Adolescent High Performers clearly outperformed the Moderate Development group in jump height and concentric impulse. In the 18–24 range, Elite Balanced and Intermediate Performers were contrasted by a small but distinctive Impulse Specialists group characterized by extremely high concentric impulse. Finally, in the 25–32 category, Sustained Elite players maintained high outputs, whereas Declining Output players displayed diminished performance capacity.

These results reinforce that clustering captures biomechanical differentiation not explained by age alone.

VII. Discussion and Conclusion

This study developed and validated a clustering-based framework for profiling countermovement jump performance in male Indian football players using standardized force platform metrics. The framework successfully identified two distinct trial-level performance profiles—Developing Performance Profile and Elite Performance Profile—with significant differences in key indicators such as peak power, concentric impulse, and jump height. Importantly, additional age-stratified clustering confirmed that these outcomes were not merely artifacts of age distribution. Across the 8–12, 13–17, 18–24, and 25–32 year bands, distinct biomechanical profiles consistently emerged. For instance, Jump-Impulse Enhanced and Power Efficient groups outperformed the Developing Baseline group in younger athletes. By adolescence, Adolescent High Performers clearly separated from the Moderate Development profile, while in young adulthood, an Elite Balanced profile coexisted with a small but distinctive Impulse Specialists group. In the oldest band, Sustained Elite players maintained strong outputs, whereas the Declining Output profile displayed reduced capacity. These findings reinforce that clustering captures meaningful biomechanical heterogeneity within each developmental stage, confirming its utility beyond simple chronological stratification.

Implications for Talent Identification and Training

The clustering approach provides an objective basis for talent identification and benchmarking in Indian football. Characteristics of elite-level trial performance can inform national and regional selection pipelines, while developing-level trial performance can help identify athletes who may benefit from structured performance enhancement programs. Unlike traditional single-metric assessments, this multi-dimensional method offers coaches and sports scientists a holistic view of an athlete's biomechanical capabilities, enabling data-driven decision-making [13].

Furthermore, cluster-specific profiling identifies athlete groups with distinct performance characteristics, thereby supporting individualized training programs. Developing athletes, for instance, may benefit from interventions focused on explosive power and neuromuscular control, whereas elite athletes may require sport-specific skill refinement and load management strategies [14]. This targeted approach not only improves performance but can also reduce injury risk.

Limitations and Future Research Directions

While promising, the study acknowledges several limitations. First, This study focused on male athletes; inclusion of female athletes in future work would enhance generalizability. Second, the cross-sectional design prevents the assessment of longitudinal performance development. Third, the analysis focused solely on physical metrics, excluding technical, tactical, or psychological dimensions of performance.

Future research should address these limitations by incorporating:

- Larger, more diverse athlete cohorts that include greater representation of female athletes to enhance the generalizability and applicability of findings across genders
- Longitudinal tracking to capture developmental trajectories
- Technical and game-related data to develop holistic athlete profiling frameworks

The proposed clustering-based profiling system fills a critical gap in Indian sports science by offering Indian-specific performance benchmarks and classification standards. It enables scalable, repeatable, and interpretable segmentation of athletes based on objective data. The findings from this study have practical utility in enhancing talent identification accuracy, designing personalized training interventions, and supporting long-term athlete development strategies.

With continued refinement and broader adoption, this framework has the potential to revolutionize evidence-based athlete development not only in Indian football but across multiple sports contexts.

Acknowledgment

The authors gratefully acknowledge the support of all coaches, athletes, and data collection teams who contributed to this research. Special thanks are extended to the team at PhysioQinesis / Performance Lab, Mumbai, India, for providing access to their sports science facilities and VALD force platform technology, which enabled high-quality and standardized performance assessments critical to this study.

References

[1] J. J. Smith and R. T. Doe, “Normative data for vertical jump performance in elite athletes,” *Journal of Sports Sciences*, vol. 38, no. 5, pp. 512-520, 2020.

[2] L. M. Johnson and K. L. Brown, “The role of vertical jump testing in athlete performance monitoring,” *Sports Biomechanics*, vol. 18, no. 3, pp. 301-315, 2019.

[3] A. M. Williams and S. K. Patel, “Portable force platforms: A game-changer in sports science,” *International Journal of Sports Physiology and Performance*, vol. 16, no. 2, pp. 145-153, 2021.

[4] H. Y. Lee and R. Gupta, “Reactive Strength Index-Modified: A new metric for assessing explosive power,” *Journal of Strength and Conditioning Research*, vol. 32, no. 7, pp. 1987-1995, 2018.

[5] S. Kumar and P. Singh, “Normative data for Indian athletes: Challenges and opportunities,” *Indian Journal of Sports Science*, vol. 14, no. 1, pp. 45-52, 2022.

[6] R. Sharma and N. Mehta, “The need for Indian-specific normative data in sports science,” *Asian Journal of Sports Medicine*, vol. 11, no. 4, pp. 123-130, 2020.

[7] R. Sharma and N. Mehta, “Player profiling in elite soccer using k-means clustering,” *Asian Journal of Sports Medicine*, vol. 11, no. 4, pp. 123-130, 2020.

[8] V. Patel and A. Desai, “Hierarchical clustering for talent identification in youth basketball,” *Journal of Human Kinetics*, vol. 70, no. 1, pp. 89-97, 2021.

[9] S. Reddy and K. Rao, “Position-specific profiling in rugby using Gaussian mixture models,” *Sports Medicine International*, vol. 25, no. 3, pp. 210-218, 2019.

[10] A. Gupta and R. Singh, “Clustering analysis for performance profiling in track and field,” *Indian Journal of Sports Performance*, vol. 8, no. 2, pp. 75-82, 2022.

[11] J. J. Smith and R. T. Doe, “Profiling elite marathon runners using k-means clustering,” *Journal of Sports Sciences*, vol. 38, no. 5, pp. 512-520, 2020.

[12] L. M. Johnson and K. L. Brown, “Clustering for talent identification in youth soccer,” *Sports Biomechanics*, vol. 18, no. 3, pp. 301-315, 2019.

[13] S. Mehta and P. Kumar, “Machine learning for athlete segmentation in multi-sport events,” *Journal of Sports Research*, vol. 12, no. 1, pp. 34-41, 2021.

[14] A. Gupta and R. Singh, “Profiling elite weightlifters using k-means clustering,” *Indian Journal of Sports Performance*, vol. 8, no. 2, pp. 75-82, 2022.