Classification and Analysis of Collision and Contact Sports

Kyle Tranfaglia, Dustin O'Brien, Dylan Johnson

Department of Mathematics, Salisbury University

Data Science 490: Capstone Project

Dr. Kyle Teller

May 16, 2025

Abstract

Traditional sports classifications as "collision" or "contact" rely on subjective descriptors without neurocognitive validation, overlooking the impact of repeated head impacts (RHIs). This study established a data-driven framework using the Sports Cog 1 dataset to analyze relationships between concussion history, sports participation, and cognitive-behavioral factors. Our methodology included comprehensive data preprocessing, statistical analyses, principal component analysis (PCA) dimensionality reduction, and clustering techniques to identify distinct neurocognitive profiles. Results revealed an optimal two-cluster solution, validating the conventional binary classification system. Significant associations ($p < 0.05$, $p < 0.01$) emerged between concussion history and impaired sleep and attention measures. Sleep-related features aligned closer with traditional sport classifications, identifying football as a major collision sport, while attention-related features demonstrated strong clustering and effectively differentiated individuals with concussion history. This research provides an evidence-based approach to categorizing sports based on RHI exposure, with potential implications for safety protocols and athlete health policies. Future work will refine these classifications by incorporating neurological assessments (e.g., EEG data), expanding the dataset to include greater sport variety, more participants, comprehensive RHI data, and critically, examining sports positions that likely impact concussion risk and cognitive-behavioral outcomes.

Keywords: Concussion, Repeated Head Impact, Sport Classification, Neurocognitive Profiles, Clustering, Attention, Sleep, Cognitive Assessment, Behavioral Assessment, Athlete Health

Classification and Analysis of Collision and Contact Sports

Sports-related concussions have become a more discussed topic within the public health and athletic research industry due to short-term and long-term cognitive consequences. Traditionally, sports are classified into categories such as “collision,” “contact,” and sometimes “non-contact,” often relying on opinionated assessments of intensity and physicality. However, these classifications frequently lack data-driven evidence within neurological or cognitive outcomes. The widespread prevalence and complexity of concussions underscore the need for more evidence-based sport categorization systems. According to Nationwide Children’s Hospital (n.d.), concussions are a surprisingly common occurrence in sports, with high school athletes suffering thousands of concussions every year, most often in football, ice hockey, and soccer. While this resource provides public health guidance, clinical and neuroscience research redefine how concussion risk should be measured.

Hallock et al. (2023) emphasize the growing understanding of sport-related concussion as a complex condition, influenced by the science of head impacts and the consistent burden of exposure. Their work shows that, other than short-term symptoms, concussions can produce constant neurocognitive dysfunction, particularly in areas such as attention, memory, and emotional control. These effects may worsen over time, especially in athletes with repeated head impacts (RHIs), regardless of whether each incident meets the criteria to be considered a concussion. Supporting this, Montenigro et al. (2017) demonstrated that RHI exposure, not just diagnosed concussion, predicts later-life impairments in executive function, emotional well-being, and overall cognition. Their study of former football players suggests that conventional risk classifications, which tend to focus solely on visible collisions, may undermine the neurocognitive dangers present in different sports. The implications of this are profound;

even sports perceived as moderately risky may carry significant long-term consequences depending on position, frequency of play, and exposure.

Adding to this reevaluation, Oldham et al. (2024) explored how sport type influences injury recurrence after a concussion. Their findings revealed that collision sports significantly increase the likelihood of subsequent injuries compared to contact or non-contact sports, reinforcing the need for a more nuanced and data-driven classification system that accounts for both acute injury patterns and long-term vulnerability. Wellm et al. (2024) directly challenge the assumption that certain sports, such as basketball, are "contactless." Through objective quantification of in-game physical interactions, they reveal frequent and often overlooked contact events that blur the lines between traditional classifications. Their research underscores how conventional categorizations fail to capture the true nature of physical interactions in many sports, potentially leading to an underestimation of concussion and RHI risk.

A recent study showed that 91.7% of football players had Chronic Traumatic Encephalopathy (CTE) of the 376 studied. The study also acknowledges potential failures in the ability to properly represent brain injuries due to potential influence due to selection bias among represented players (Boston University Chobanian Avedisian School of Medicine, 2023). Simulations have similarly been conducted, potentially removing such bias as presented in Dubas et al. (2020), which proved unreliable in predicting post-injury. These studies show a clear need for new model creation and improvement to allow for a clearer impact of sports.

It is worth noting that studies such as Marsh and Kleitman (2003) highlight a more positive perspective on the role of sports in student development. Their research suggests that students who regularly participate in sports tend to exhibit higher levels of school commitment compared to those who do not, indicating that environmental and behavioral factors can influence psychological outcomes. This finding stands in contrast to the conclusions drawn by Sullivan and Riccio (2010), who reported that traumatic brain injury (TBI) negatively affects adolescent cognitive function across all settings. Similarly, Berisha et al. (2017) observed speech performance issues in NFL players diagnosed with CTE, reinforcing the broader cognitive impacts of brain trauma in sports. These contrasting findings suggest that while sports participation may foster positive engagement for some, the risks associated with injury, particularly head trauma, can lead to adverse long-term outcomes.

Taken together, these findings highlight a major gap in how sports are currently categorized and understood in terms of concussion risk and cognitive outcomes. The literature reveals that traditional classification systems rely heavily on subjective descriptors rather than empirical evidence linking sport participation to specific neurocognitive outcomes. This gap hampers efforts to develop targeted prevention strategies, appropriate safety protocols, and effective health policies for athletes across different sports. This research project aims to bridge that gap by applying statistical clustering techniques to the Sports Cog 1 dataset.

Background

The conventional categorization of sports into "collision" and "contact" disciplines has predominantly relied on subjective evaluations and broad descriptors rather than evidence-based criteria. Collision sports are typically characterized by athletes intentionally striking or colliding with each other or the playing surface as an intrinsic aspect of gameplay, exemplified by sports such as football, rugby, and ice hockey. Conversely, contact sports involve more frequent but less forceful physical interactions that are not central to the sport's fundamental mechanics, as seen in basketball and soccer. These classifications significantly influence policies concerning youth participation, safety regulations, equipment standards, and medical guidelines.

However, the applied classifications exhibit considerable variability among medical professionals, athletic associations, and sports governing bodies. This inconsistency poses challenges for comparing injury rates, assessing risk factors, and implementing standardized safety measures across diverse sporting activities. The absence of objective criteria for categorization has resulted in disparate classifications depending on the source, leading to potential confusion and the implementation of suboptimal risk management strategies. The current classification system inadequately accounts for measurable impacts on cognitive function and brain health, particularly those associated with RHIs, defined as the cumulative effect of sub-concussive impacts experienced over time, which has been increasingly implicated in structural brain alterations and long-term neurodegenerative conditions, including CTE. Despite growing evidence of these associations, current sports classifications rarely incorporate neurocognitive markers that could provide more objective measures of impact exposure.

In contrast to these subjective definitions, our research employs a data-driven approach to refine and potentially redefine these classifications. By analyzing patterns in cognitive and behavioral data from athletes with varied sports participation histories, we aim to identify statistically significant differences that may more accurately delineate collision and contact sports based on empirically measured outcomes. Specifically, our analysis focuses on sleep and attention-related issues in athletes with and without a history of concussion, as these domains have demonstrated sensitivity to both diagnosed concussions and sub-concussive impacts.

The objective of this project is to develop a more empirical foundation for classifying sports based on their associated cognitive and behavioral profiles. This approach offers valuable insights into how different sports may affect brain function and athlete well-being beyond traditional classification methods. By establishing a more objective classification system, we can better inform safety protocols, guide policy decisions, and ultimately enhance protection for athletes across all levels of competition.

Research Questions

By analyzing cognitive, behavioral, and concussion data from 332 online survey participants gathered from a paid online survey collection organized by the Psychology department at Salisbury University, this study aims to answer the following questions: How can objective, data-driven classification systems distinguish between "collision" and "contact" sports categories? Can a two-cluster system effectively distinguish sport types? Can clustering techniques effectively identify distinct sport groupings based on athletes' behavioral and cognitive performance data? Where do inconsistencies arise when comparing traditional sports classifications to data-driven approaches? Which machine learning approaches most effectively differentiate between collision and contact sports based on athlete data? To what extent do behavioral and cognitive performance metrics distinguish athletes in collision sports from those in contact sports? The results could enhance our understanding of sports categorization beyond traditional subjective methods, potentially informing safer athletic policies and more targeted interventions for athletes based on evidence-driven sport classifications rather than conventional designations.

Methodology

Data Source and Collection

This study utilized the Sports Cog 1 dataset, comprising head injury history, cognitive task performance, and behavioral data from athletes and non-athletes between the ages of 18 and 25. The dataset included survey responses on sports participation, concussion history, sleep patterns, attention behaviors, pain experiences, motivation metrics, and demographic information. The starting dataset contained over 360 participants, but was condensed to 332 participants due to failure to meet requirements.

Data Preprocessing and Cleaning

The initial data preprocessing involved systematic filtering to ensure data quality and participant eligibility. Participants were excluded from analysis based on several criteria: those who withheld consent for data utilization in research, individuals outside the target age range of 18-25 years, participants who failed to pass at least two of three attention check questions embedded within the survey, and those with substantially incomplete or invalid responses. Additionally, personally identifiable information, including participant names, contact details, and other identification markers, was removed to maintain confidentiality and comply with ethical research standards.

The dataset required extensive cleaning to standardize responses and ensure consistency for analytical procedures. Several variables required standardization due to their free-text format. Sport status responses containing information regarding sport participation, duration, and positions were extracted and standardized using regular expression patterns and text normalization techniques. Sleep schedule information, including bedtime and wake-up times, was converted to a consistent 24-hour format, allowing for calculation of sleep duration metrics. Concussion history responses detailing the number of concussions and associated sports were parsed and normalized for quantitative analysis.

Various text cleaning operations were performed, such as response stripping to remove extraneous characters and whitespace, text normalization to standardize capitalization and terminology, and regular expression replacements to extract consistent information from variable response formats. Time conversions were applied to standardize reported sleep schedules, and string concatenation was used to create compound variables where appropriate. Natural language processing techniques were employed using the OpenAI API via LangChain to interpret and categorize responses for particularly complex or inconsistent responses.

To prepare the data for statistical analysis and clustering, numerical variables were normalized to ensure comparability across different measurement scales. Categorical variables were encoded using appropriate schemes: binary variables were encoded as 0/1, ordinal variables were encoded to preserve their inherent ordering, and nominal variables with multiple categories were encoded using one-hot encoding techniques.

Exploratory Data Analysis

Comprehensive exploratory data analysis was conducted to understand the characteristics and distributions within the dataset. Sports participation analysis was performed across more than 20 different sports categories, and concussion prevalence analysis revealed that more than 80% of participants reported no history of concussions (Figure 1). Distribution analysis of key variables included central tendency measures such as mean, median, and mode; dispersion measures including standard deviation, variance, and range; and frequency distributions of categorical responses. Systematic comparisons were performed to identify patterns and differences between athletes and non-athletes, between participants with and without reported concussion history, and across different sport categories to identify sport-specific patterns.

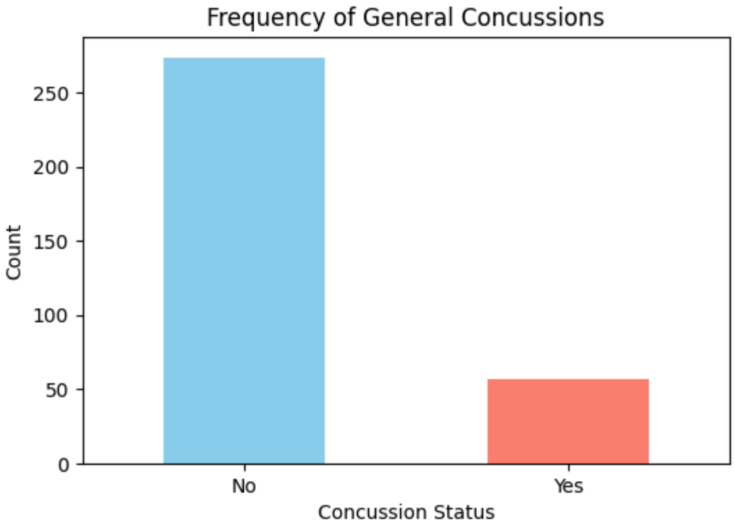


Figure 1: A bar graph showing the number of participants who had a concussion

In-depth analysis was conducted for key cognitive-behavioral domains. A major domain was sleep behavior metrics, which included sleep duration analysis (average hours per night), sleep quality indicators, sleep schedule consistency, and sleep disturbance frequency. Another significant domain was Attention performance metrics, encompassing sustained attention capabilities, attention disruption frequency, task switching performance, and focus maintenance metrics. Two smaller domains included pain, which focused on pain regularity, and motivation issues impacting daily tasks.

Preliminary visualization of sport distribution within our dataset revealed substantial variation in representation across sporting activities (Figure 2a). Soccer was the most prevalent sport (24.3%), followed by basketball (11.3%), football (9.6%), tennis (9.6%), and volleyball (7.9%). This distribution provided a robust sample of traditionally classified contact and collision sports, allowing for meaningful cross-sport comparisons. When examining concussion distribution across sports (Figure 2b), several noteworthy patterns emerged that inform our understanding of relative concussion risk. Despite soccer comprising 24.3% of the athlete population, it accounted for a proportional 23.8% of reported concussions. However, football demonstrated a disproportionately high concussion burden, representing 21.4% of all concussions despite comprising only 9.6% of athletes in the dataset. This more than two-fold overrepresentation in concussion prevalence aligns with existing literature on football's elevated concussion risk as a collision sport. Even more striking were the findings related to less common sporting activities: motocross represented 11.9% of concussions despite minimal representation in the overall dataset (notably one person), while combat sports (taekwondo/MMA/kickboxing) accounted for 7.1% of concussions despite representing less than 2% of the athlete population. These disproportionate rates provided initial evidence for sport-specific concussion risk that traditional classification schemes might not adequately capture, suggesting the value of our data-driven approach to sport categorization based on neurocognitive impact.

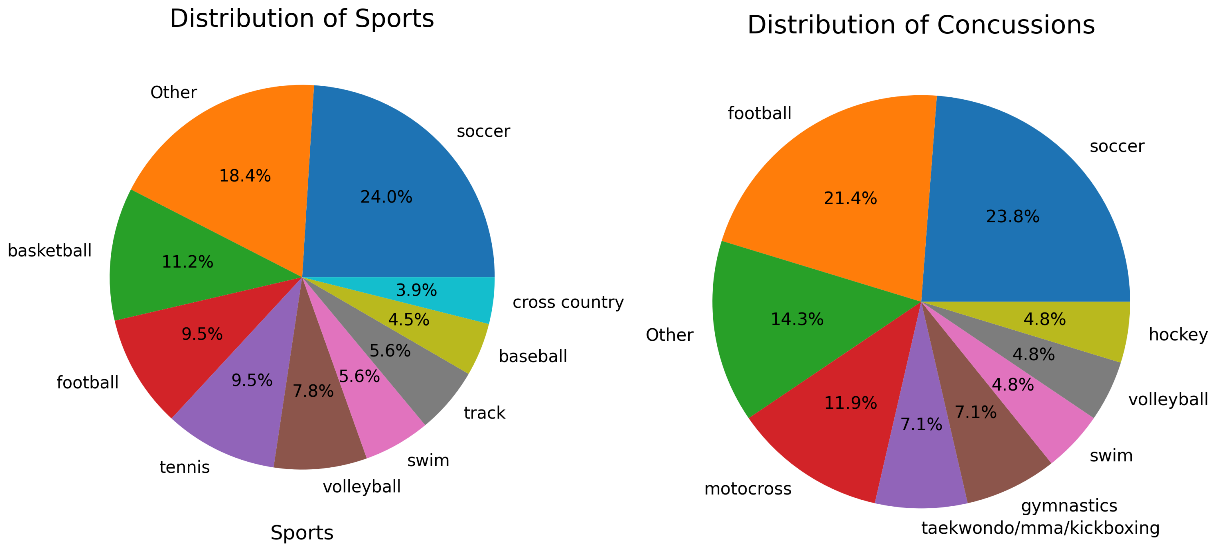


Figure 2 (a, b): Distribution of sports types and distribution of concussions by sport

Sleep metrics demonstrated revealing patterns that further informed our classification approach. The overall distribution of average sleep duration showed a relatively normal distribution centered around 7-8 hours per night, with the majority of participants (approximately

70%) reporting between 6-8 hours of sleep (Figure 3). However, consistent and statistically significant differences emerged when examining sleep quality by concussion status (Figure 4). Among the general population, participants with concussion history reported markedly worse sleep quality, with 40% reporting "fairly bad" or "very bad" sleep compared to 22% of those without concussion history (Figure 4c). This pattern was even more pronounced among athletes specifically, where 45% of concussed athletes reported "fairly bad" or "very bad" sleep quality compared to 18% of non-concussed athletes (Figure 4a). Furthermore, comparing the distribution of sleep quality for non-athletes by concussion status revealed interesting differences in sleep quality patterns, as surprisingly, 90% of non-athletes with concussions reported “fairly good” sleep quality compared to 58% of non-concussed participants reporting “fairly good” sleep quality (Figure 4b). These preliminary findings suggested that sleep metrics could serve as a valuable dimension for evidence-based sport classification, with particular sensitivity to concussion history that might reflect accumulated neurological impacts from RHI.

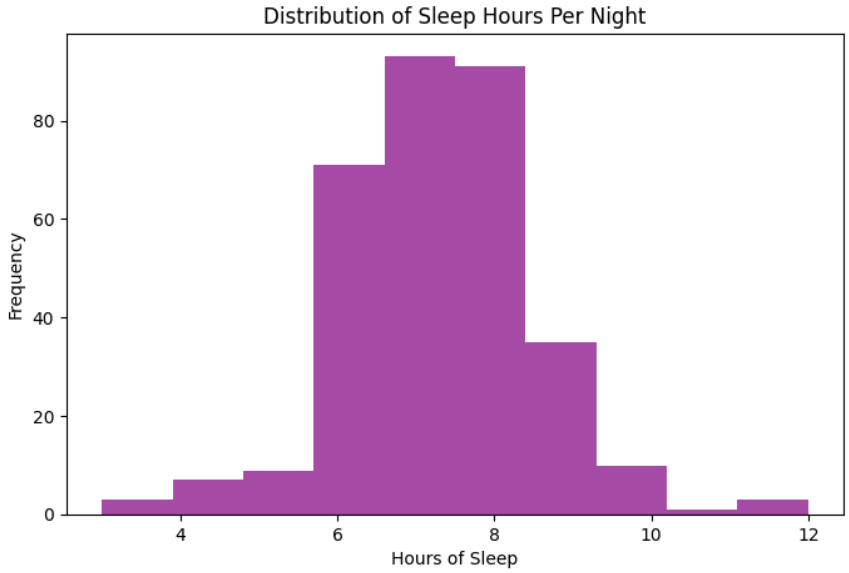


Figure 3: Distribution of average hours of sleep per night

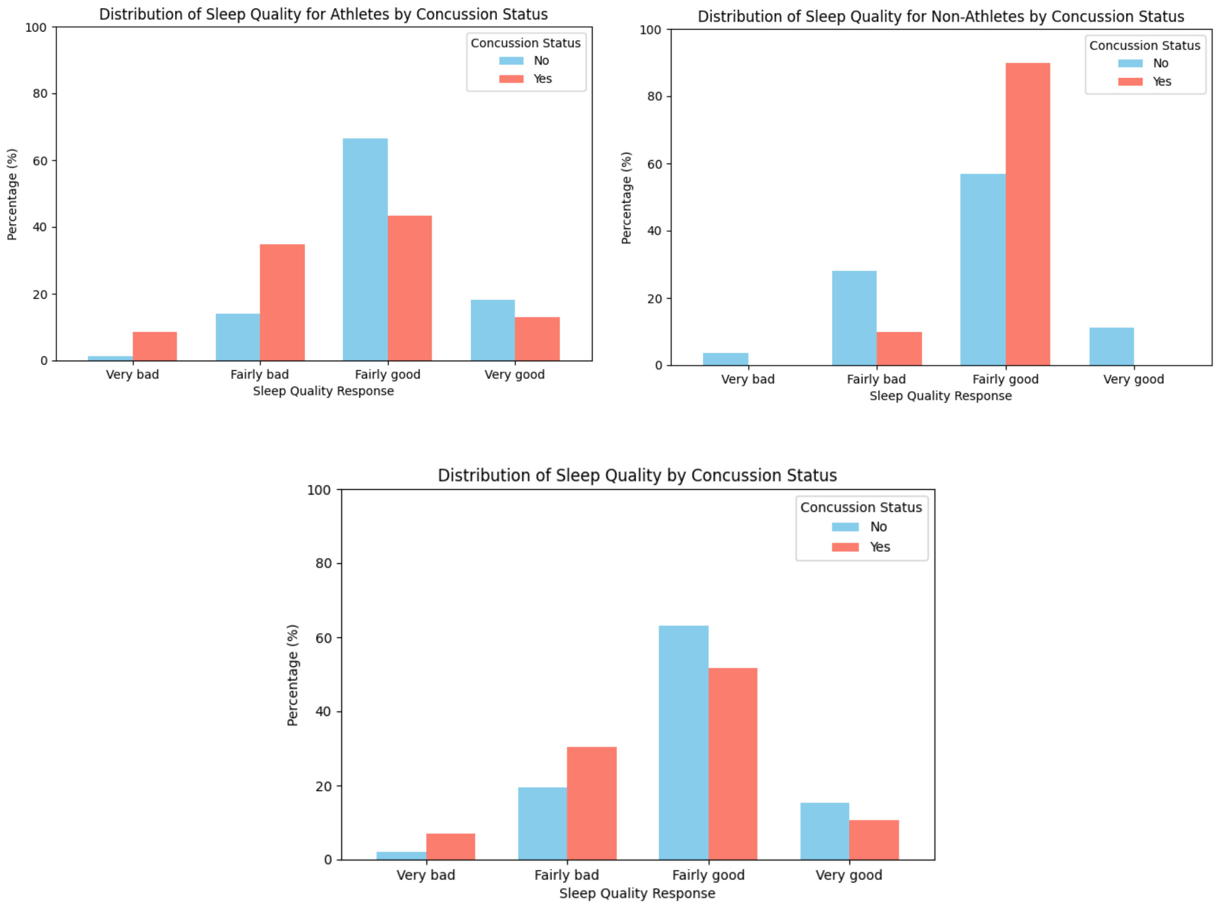


Figure 4 (a-c): Distribution of sleep quality responses by concussion status

Attention metrics, operationalized as self-reported concentration issues, demonstrated equally compelling patterns that reinforced and complemented the sleep findings. There are consistent differences in concentration difficulties between participants with and without a concussion history (Figure 5). In the overall population, individuals with concussion history were nearly four times more likely to report "always" having concentration issues (20% vs. 5%) and significantly more likely to report "often" having concentration problems (33% vs. 23%) (Figure 5c). When examining athletes specifically, this pattern persisted with 21% of concussed athletes reporting "always" having concentration difficulties compared to just 2% of non-concussed athletes (Figure 5a). Similarly, non-athletes with concussion history showed heightened concentration difficulties, with 40% reporting "often" having concentration issues versus 25% of those without concussions (Figure 5b). These consistent patterns across population subgroups provided compelling evidence that attention metrics could serve as sensitive indicators of neurocognitive impact from concussion exposure. Furthermore, the magnitude of these differences suggested that attention metrics might provide even stronger discriminatory power than sleep metrics for developing an evidence-based sport classification framework. The combination of sleep and attention metrics thus emerged as particularly promising dimensions for our clustering approach to sport classification based on neurocognitive profiles.

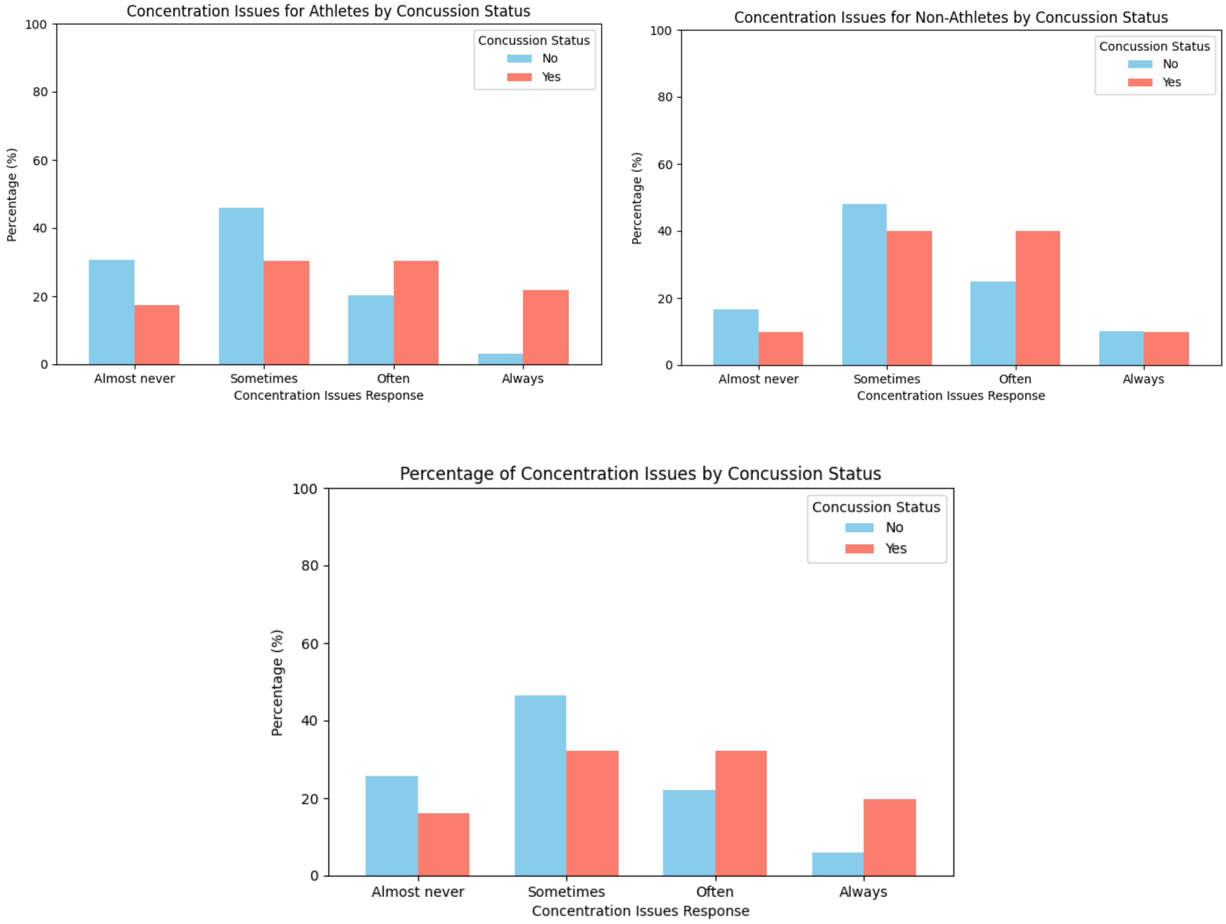


Figure 5 (a-c): Distribution of concentration issue responses by concussion status

Statistical Analysis and Feature Selection

To identify statistically significant relationships between concussion history, sports participation, and cognitive-behavioral factors, we employed proportion tests to compare the distributions of feature responses between the concussion and non-concussion groups. Statistically significant differences were identified at alpha levels of $p < 0.05$ and $p < 0.01$. Correlation analyses were conducted to examine associations between concussion history and cognitive-behavioral metrics, sport participation patterns and cognitive-behavioral outcomes, as well as between specific sports and concussion prevalence. Group differences were further explored using appropriate parametric and non-parametric statistical tests, selected based on the distributional characteristics of the data. Given that the majority of the features were categorical with multiple response levels, the Two-Sample Proportion Test was chosen as the most suitable method to determine if response distributions differed significantly between individuals with and without a history of concussion. This test allows for a direct comparison of the proportion of participants selecting each response level across the two groups, enabling an evaluation of whether specific response patterns were comparatively more prevalent in one group.

Before conducting the proportion tests, we verified the underlying assumptions necessary for the validity of the results. Specifically, the two-sample proportion test assumes that each cell within the comparison, representing each response level within each group, has a sufficiently large sample size to justify the use of a normal approximation. A commonly accepted guideline is a minimum of five responses per category within each group. This condition was met for all categorical features under consideration. With this assumption satisfied, we proceeded to conduct the tests for each feature-response pair. The resulting p-values were then used to assess statistical significance and identify features exhibiting meaningful differences between the concussion and non-concussion groups. To reduce the dimensionality of the feature space while retaining meaningful variance, address potential multicollinearity among related variables, and determine the optimal number of components based on explained variance criteria, Principal Component Analysis (PCA) was implemented.

Clustering Analysis and Comparison

Clustering analysis employed an iterative feature selection approach. The initial feature set included all preprocessed cognitive-behavioral metrics. A significance-based feature set was created from features demonstrating statistically significant associations ($p < 0.05$ and $p < 0.01$) with concussion history. Domain-specific feature sets were also developed for sleep-specific and attention-specific features, which were used to cluster participants by domain-specific features, given a significance threshold. Multiple clustering approaches were evaluated, including K-means clustering with varying numbers of clusters ($k = 2$ to $k = 5$), hierarchical clustering using different linkage methods and distance metrics, and density-based clustering to identify naturally emerging clusters. The validity and optimal number of clusters were assessed through silhouette scoring to evaluate cluster separation and cohesion, the elbow method to identify the optimal number of clusters, and feature distribution inspection across identified clusters.

The final analytical stage involved comparing traditionally defined sport classifications with our data-driven clustering results. Each sport was mapped to its conventional classification (collision or contact). Distribution of sports across identified clusters was analyzed, and concussion prevalence across both traditional classifications and data-driven clusters was examined. Assessment of which cognitive-behavioral features most strongly differentiated the clusters revealed that sleep-related features aligned most closely with traditional sport classifications (although it still presented minor differences), while attention-related features provided novel discriminatory power that differed from conventional categorizations.

Software and Tools

All data preprocessing, analysis, and visualization were conducted using Python 3.9 with an extensive suite of specialized libraries. Data manipulation and cleaning relied primarily on pandas for structured data operations and NumPy for numerical computations, while datetime, dateutil, and re/regex facilitated time-based calculations and pattern matching for text standardization. Statistical analyses employed SciPy and the statistical programming language R for hypothesis testing and distribution analysis.

The machine learning pipeline incorporated multiple components from scikit-learn: data preprocessing utilized StandardScaler and RobustScaler for feature normalization, SimpleImputer for handling missing values, and PCA for dimensionality reduction. Cluster analysis implemented several algorithms, including DBSCAN, AgglomerativeClustering, and most successfully, KMeans with comprehensive evaluation metrics (silhouette_score, calinski_harabasz_score, davies_bouldin_score). A random forest classifier was employed for additional validation and feature importance assessment. Dimensionality reduction for visualization purposes was performed using t-SNE (TSNE) to represent high-dimensional relationships in two-dimensional space.

A combination of Matplotlib and Seaborn enabled visualization of data distributions, relationships, and analytical results, as well as standard statistical plots and customized visualizations of complex relationships. The Counter package facilitated frequency analysis, particularly for categorical data distributions. Finally, text processing relied on regular expression libraries, and complex text normalization tasks were accomplished through LangChain with OpenAI API integration and JSON formatting.

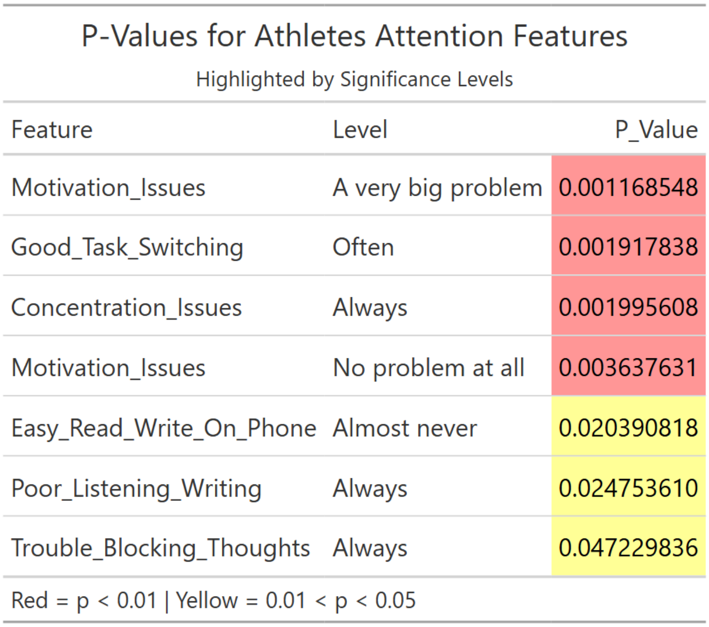
Results

Feature Relevance

Our statistical analysis revealed clear distinctions between athletes with and without concussion history across both sleep and attention domains. To establish reliable correlations between RHI and cognitive-behavioral outcomes, we extracted relevant features using two-proportion statistical testing with significance thresholds of 0.01 and 0.05 (Figure 6).

Statistical testing identified specific response levels within cognitive-behavioral features that significantly differentiated athletes with a concussion history from those without. Four attention features showed robust significance at the $p < 0.01$ threshold (Figure 6a): motivation issues (both "A very big problem" and "No problem at all" response levels), successful task-switching capabilities ("Often" response), and concentration difficulties ("Always" response). Additionally, three attention features met the $p < 0.05$ threshold: difficulty reading/writing while on the phone, poor listening/writing skills, and trouble blocking unwanted thoughts. The statistical strength of these relationships suggests that attention mechanisms may be particularly vulnerable to RHI exposure.

The analysis of sleep-related measures yielded similarly compelling results, with nine specific sleep features emerging as significant at the $p < 0.05$ threshold (Figure 6b). Six features demonstrated high statistical significance ($p < 0.01$): inability to sleep (level 3), frequency of bad dreams (level 3), sleep medication usage (level 0), difficulty staying awake (levels 0 and 2), and overall sleep quality (levels 1 and 2). These findings indicate that athletes with a concussion history reported significantly different patterns of sleep disturbance compared to those without such history, with particular emphasis on medication usage, dream disturbances, and sleep maintenance difficulties.

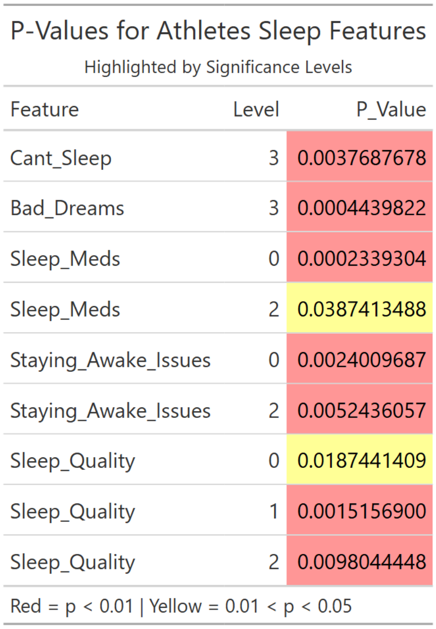


P-Values for Athletes Attention Features

Highlighted by Significance Levels

Feature	Level	P_Value
Motivation_Issues	A very big problem	0.001168548
Good_Task_Switching	Often	0.001917838
Concentration_Issues	Always	0.001995608
Motivation_Issues	No problem at all	0.003637631
Easy_Read_Write_On_Phone	Almost never	0.020390818
Poor_Listening_Writing	Always	0.024753610
Trouble_Blocking_Thoughts	Always	0.047229836

Red = p < 0.01 | Yellow = 0.01 < p < 0.05



P-Values for Athletes Sleep Features

Highlighted by Significance Levels

Feature	Level	P_Value
Cant_Sleep	3	0.0037687678
Bad_Dreams	3	0.0004439822
Sleep_Meds	0	0.0002339304
Sleep_Meds	2	0.0387413488
Staying_Awake_Issues	0	0.0024009687
Staying_Awake_Issues	2	0.0052436057
Sleep_Quality	0	0.0187441409
Sleep_Quality	1	0.0015156900
Sleep_Quality	2	0.0098044448

Red = p < 0.01 | Yellow = 0.01 < p < 0.05

Figure 6 (a, b): Statistically significant features by p-value analysis for attention and sleep

This statistical approach substantially reduced our feature space from over 150 initial features (including over 100 attention and sleep-related measures) to a more focused set of 16 significant features (9 sleep-related and 7 attention-related). The feature reduction process was critical for subsequent clustering analysis, as it eliminated potential noise variables and allowed us to concentrate on the most relevant cognitive-behavioral indicators associated with RHI exposure. The statistical significance of these specific sleep and attention features establishes a clear connection between concussion history and cognitive-behavioral outcomes in our sample.

Optimal Clusters

To determine the optimal number of clusters for our analysis, we applied three distinct cluster evaluation metrics to the dataset after dimensionality reduction via PCA on all features determined significant at $p < 0.05$ (Figure 7). The Silhouette score, which measures how similar an object is to its own cluster compared to other clusters, demonstrated its highest value (0.35) at k=2, with a steady decline as the number of clusters increased (Figure 7a). This suggests that the two-cluster solution provides the most distinct and well-separated groupings. Similarly, the Calinski-Harabasz index, which evaluates cluster separation based on the ratio of between-cluster variance to within-cluster variance, reached its maximum value (147) at k=2, decreasing substantially with additional clusters (Figure 7b). This further supports the two-cluster solution as providing the most distinct separation in our dataset. The Davies-Bouldin index, where lower values indicate better clustering, achieved its minimum value (1.12) at k=2, with progressively increasing values for higher cluster counts, again confirming the optimal nature of the two-cluster solution (Figure 7c).

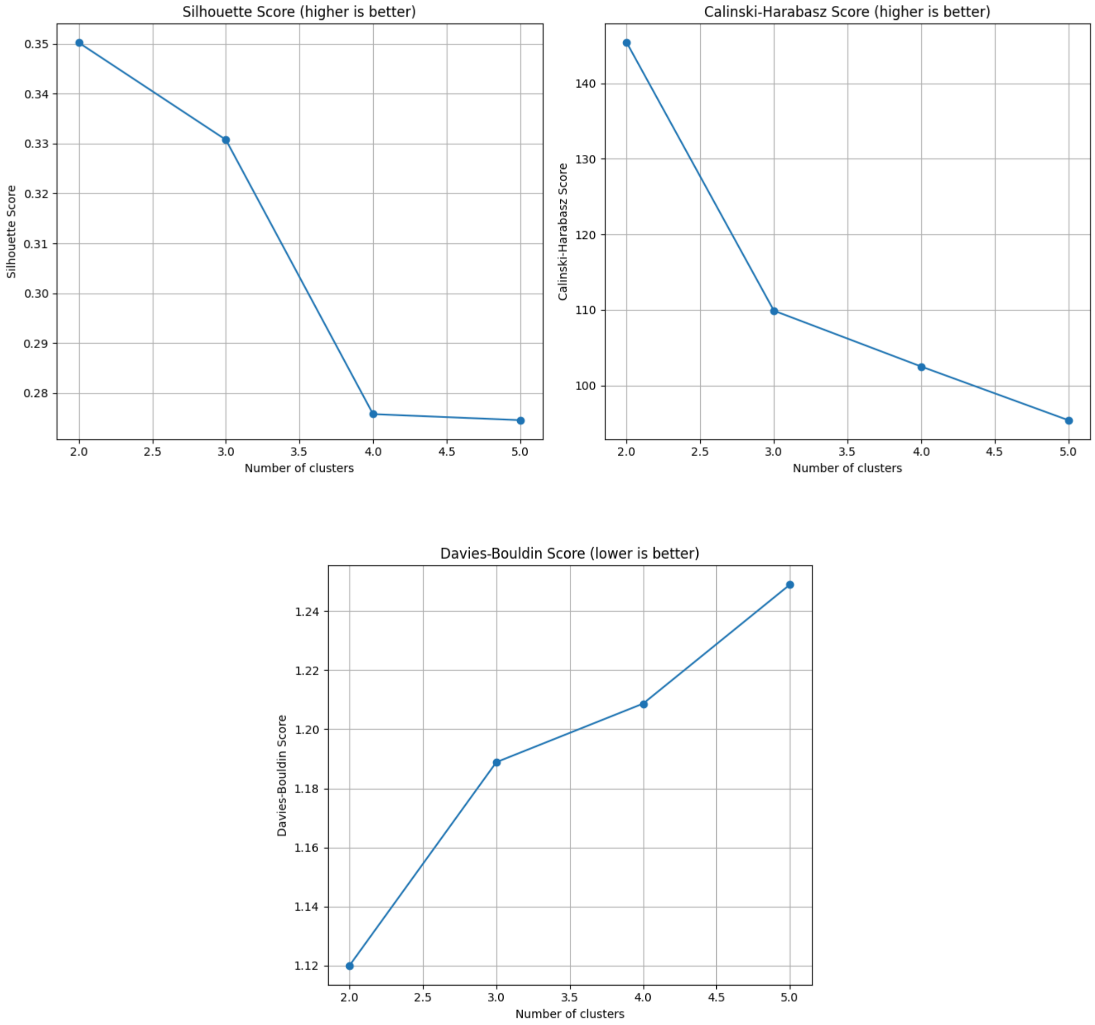


Figure 7 (a-c): Clustering metrics - optimal number of clusters by evaluation

The consistency across all three evaluation metrics provides strong evidence for a natural binary structure within our data. This finding is particularly noteworthy as it empirically validates the current binary categorization of sports into collision and contact classifications based solely on cognitive and behavioral data, without prior assumptions about sport types. While the values of the evaluation metrics indicate moderate rather than strong cluster separation, their unanimous agreement on k=2 as the optimal solution strengthens our confidence in this result. The convergence of multiple evaluation approaches on the same solution is especially meaningful given that these metrics evaluate clustering quality using different mathematical principles.

PCA Clustering

After determining that two clusters represented the optimal solution for our dataset, we proceeded to visualize these clusters using PCA under different feature selection approaches (Figure 8). The PCA visualization using only the 16 statistically significant features ($p < 0.05$) reveals remarkably clear separation between the two clusters, with distinct boundaries and minimal overlap (Figure 8c). The first principal component appears to effectively differentiate the clusters, suggesting that the variance captured by this component strongly correlates with the underlying distinction between collision and contact sport participation patterns. The well-defined clusters validate our feature selection approach and indicate that the identified cognitive-behavioral features effectively capture meaningful differences between athlete groups.

To evaluate the specific contribution of our feature selection method, we conducted a comparison analysis using 16 randomly selected features (Figure 8b). This visualization demonstrates substantially poorer cluster separation, with considerable overlap and scattered distribution of data points. The random features failed to produce coherent groupings, resulting in clusters with ambiguous boundaries and limited discriminative power. This stark contrast with the significant-feature clustering reinforces the importance of our statistical approach to feature selection, confirming that the cognitive-behavioral patterns we identified are genuinely meaningful rather than artifacts of dimensionality reduction.

We also visualized clustering results using all available features for a comprehensive assessment, which similarly showed poor separation between clusters (Figure 8c). Despite having access to all potential discriminative information, the inclusion of non-significant features introduced noise that obscured the underlying structure of the data. The clusters appear more intermixed and less distinct compared to the significant-feature approach, demonstrating that feature quantity does not improve clustering quality in this analysis.

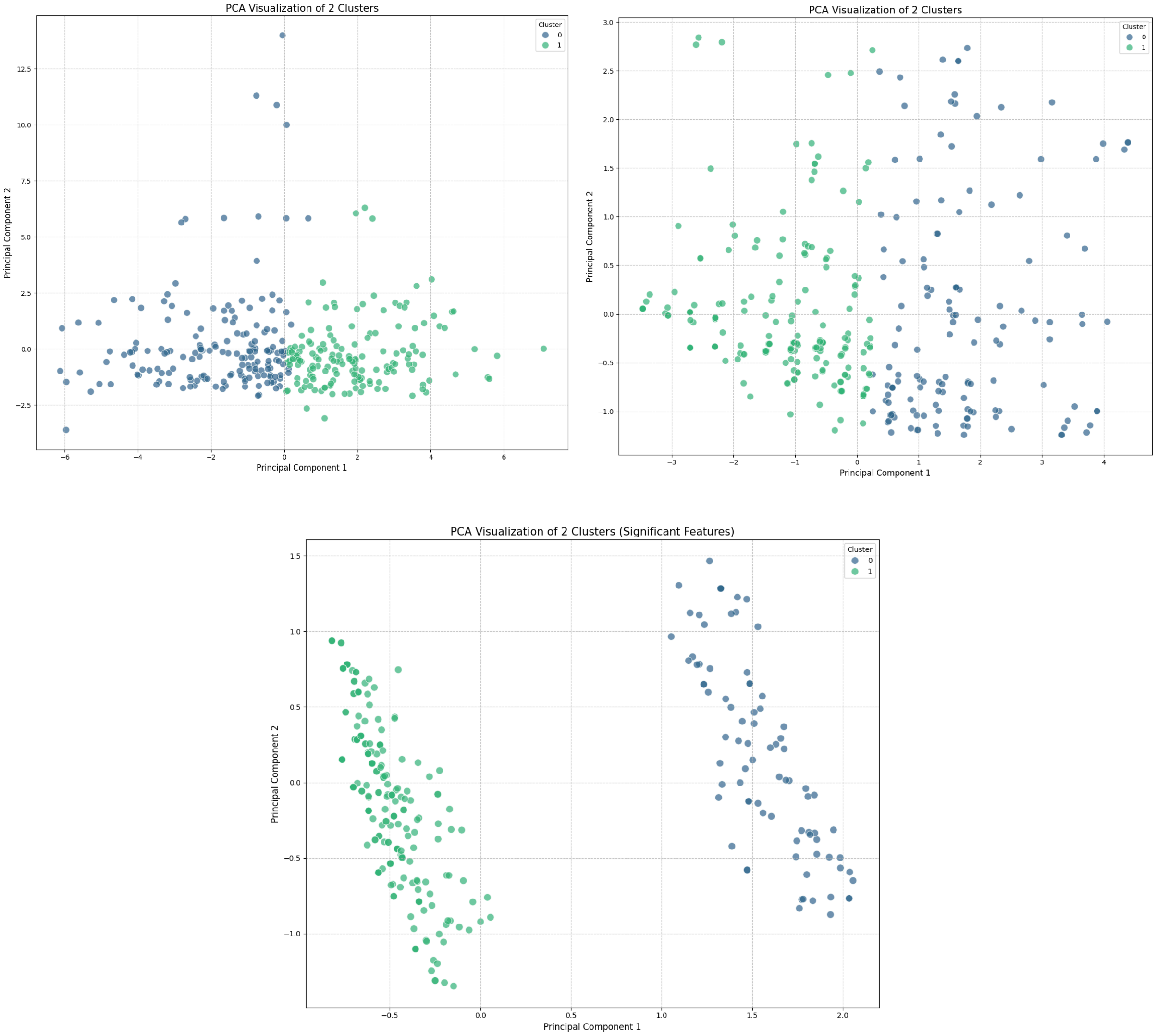


Figure 8 (a-c): All features, 16 random features, and all significant features (16) PCA clustering

Clustering by Concussions

Following our cluster validation and PCA visualization analyses, we examined how the feature domain clusters (attention only and sleep only) related to participant concussion history.

This analysis was crucial for understanding whether the cognitive-behavioral patterns we identified could effectively distinguish between participants with different exposure levels to RHI. We first analyzed clustering efficacy using the entire dataset (including both athletes and non-athletes) with attention and sleep features at the $p < 0.05$ significance threshold (Figure 9). The attention-based clustering demonstrated superior differentiation of concussion history, with one cluster showing markedly higher average number of concussions compared to the other cluster. While sleep-based features also yielded distinguishable clusters, the separation was less pronounced than that achieved with attention features. The combined feature set did not substantially improve the differentiation between clusters in terms of average concussion history, indicating potential redundancy or interference when both feature sets were used together.

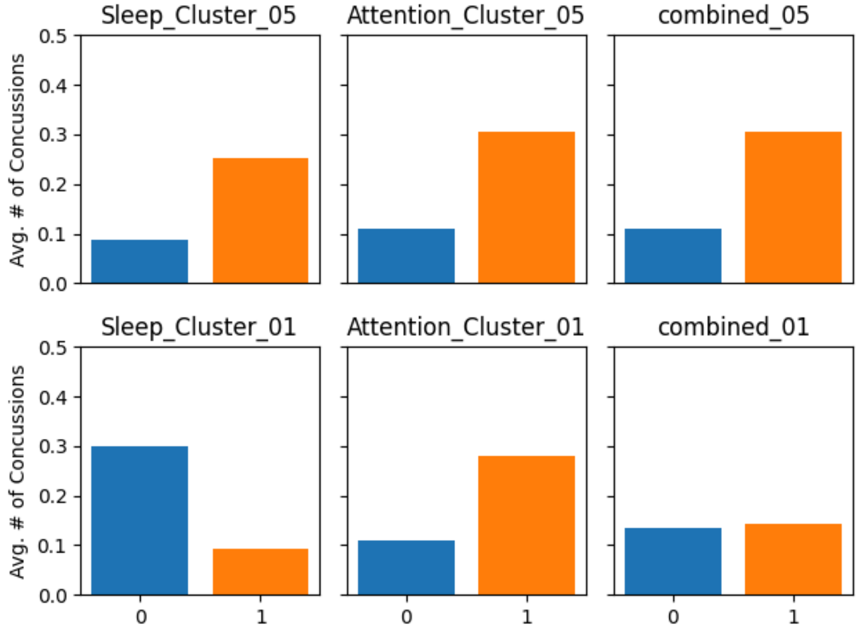


Figure 9 (a-f): Average number of concussions by cluster

To more rigorously evaluate these clustering approaches, we calculated silhouette scores without applying PCA dimensionality reduction (Figure 10). These scores confirmed that attention-based features produced significantly more cohesive and well-separated clusters compared to sleep-based features. The attention-based clustering achieved a silhouette score of approximately 0.81 and 0.54, indicating robust cluster definition, while sleep-based clustering yielded substantially lower scores (between 0.17 and 0.27). The combined feature approach showed intermediate performance (scores between 0.36 and 0.77), further supporting our observation that attention features were the primary drivers of effective clustering. The decline in performance when using combined features compared to attention features alone may reflect increased feature space sparsity, which often challenges clustering algorithms.

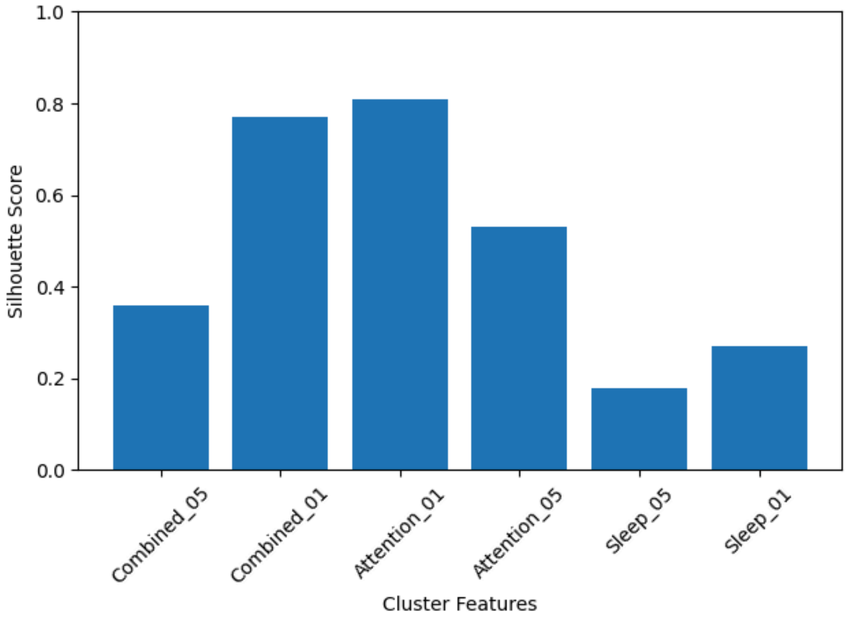


Figure 10: Domain-specific silhouette scores by cluster

When we refined our analysis to focus exclusively on athletes, removing potential noise introduced by non-athlete data, we observed some notable shifts in clustering patterns (Figure 11). Using features significant at $p < 0.01$, sleep-based clustering displayed a greater difference between clusters (0.15) compared to attention-based features, which showed equivalent concussion averages between clusters, contradicting our earlier findings. This unexpected reversal suggests that the relationship between cognitive-behavioral patterns and concussion history may be more complex than initially apparent and potentially influenced by sport-specific factors. With the refined athlete-only dataset, the combined feature approach at $p < 0.01$ revealed a meaningful difference (0.1) between clusters in terms of concussion history, while the $p < 0.05$ threshold continued to show patterns similar to those observed with attention features alone.

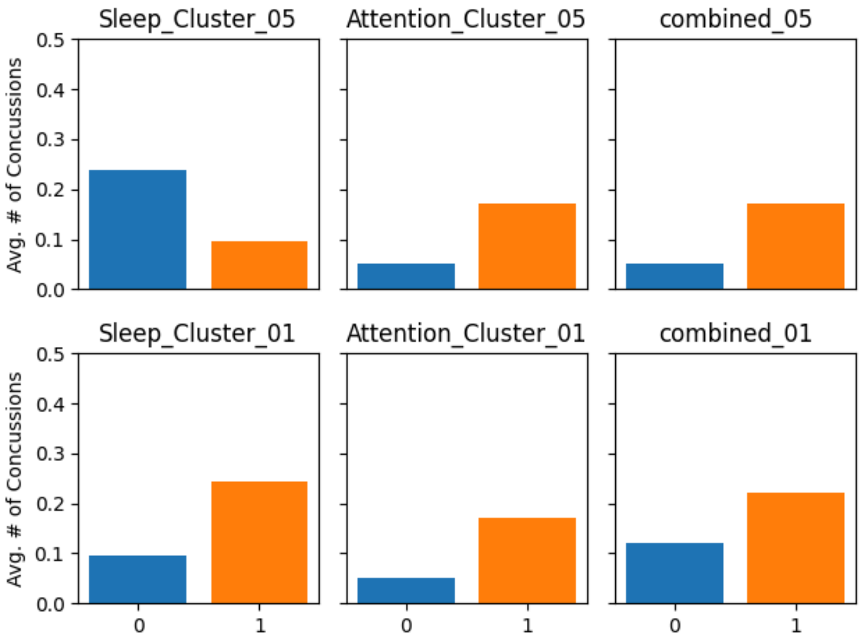


Figure 11 (a-f): Average number of concussions by cluster for athletes only

These findings demonstrate that the relationship between cognitive-behavioral features and concussion history is nuanced and may vary depending on the population analyzed and the specific features selected. While attention features generally provided stronger clustering when considering the entire participant pool, sleep features became more discriminative when focusing exclusively on athletes. This suggests that different cognitive domains may be differentially affected by RHI exposure in athletes compared to non-athletes, highlighting the importance of targeted feature selection when developing classification models for sport categories.

Clusters by Sports

The final phase of our analysis examined how different sports were distributed across the two clusters identified through our various feature selection approaches. When clustering was performed using sleep-related features at the p < .05 significance threshold, we observed that football players were distributed across both clusters, with 10 players assigned to the "collision" cluster and 12 players to the "contact" cluster (Figures 12a and 12b). This nearly even split contradicts the traditional classification of football as a definitive collision sport. Soccer, conventionally categorized as a contact sport, showed a similar pattern with 9 players in the collision cluster but 38 in the contact cluster. The sleep-based clustering also placed a substantial number of non-athletes (40) in the collision cluster, suggesting that sleep disturbances associated with collision sports may also appear in the general population for reasons unrelated to sports participation.

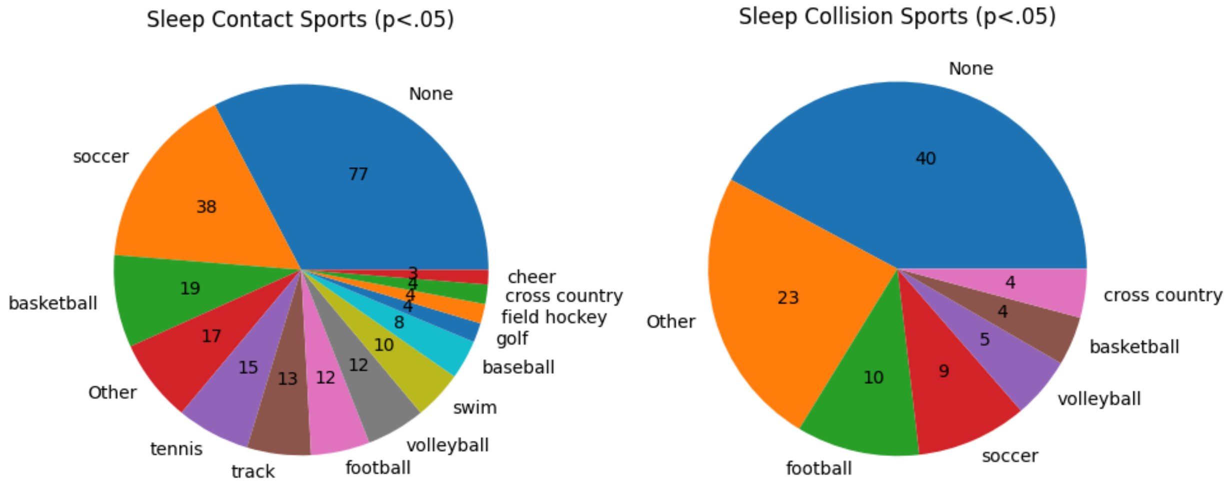


Figure 12 (a, b): Contact and collision clusters by sleep features (p < .05)

When we applied more stringent feature selection (p < .01) to sleep-related clustering, the classification became more aligned with traditional expectations (Figures 13a and 13b). Football showed a stronger presence in the collision cluster (12 players versus 9 in contact), supporting its conventional classification as a collision sport. However, significant discrepancies remained, with a substantial number of non-athletes (18) still classified in the collision cluster, and 11 soccer players appearing in the collision group rather than exclusively in the contact group. Basketball and volleyball showed similar distributions across both clusters, suggesting that the cognitive-behavioral patterns associated with these sports do not fit well into the binary classification system.

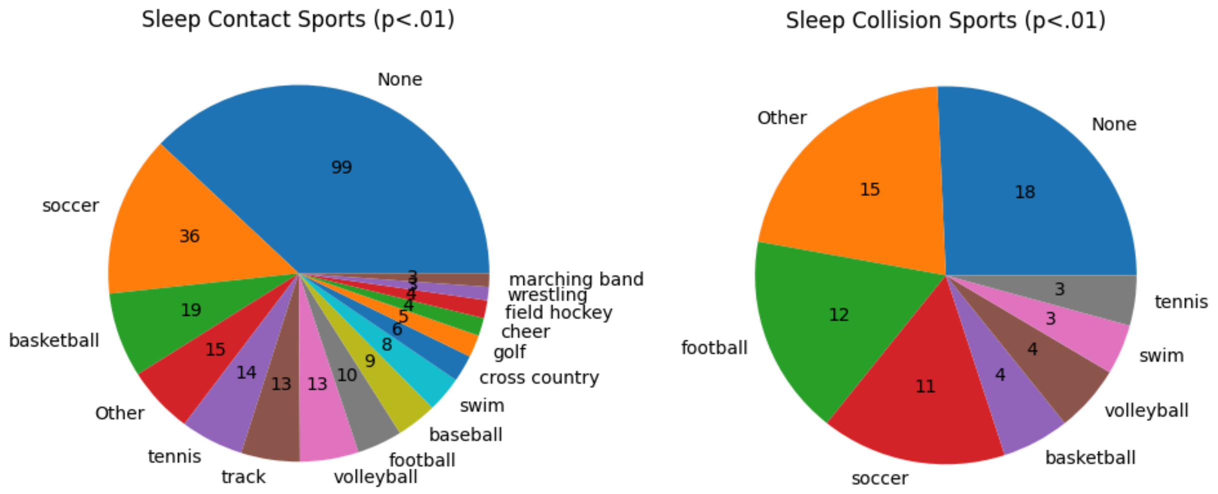


Figure 13 (a, b): Contact and collision clusters by sleep features (p < .01)

Attention-based clustering yielded even more unexpected distributions that significantly diverged from conventional sport classifications (Figures 14a and 14b). Most notably, football, universally considered a collision sport, was predominantly classified in the contact cluster (20 players), with only 2 players appearing in the collision cluster. Soccer showed a similar pattern with 41 players in the contact cluster and 6 in the collision cluster. Swimming and tennis, typically considered non-contact sports, had comparable representation in the collision cluster to football. These findings suggest that the cognitive-behavioral patterns related to attention may not align with the physical contact intensity traditionally used to classify sports.

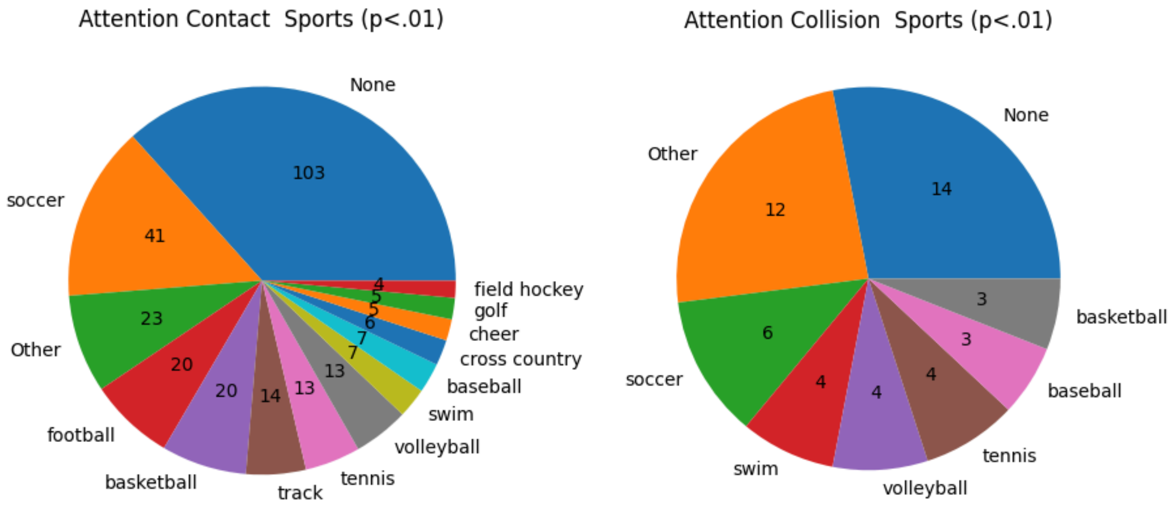


Figure 14 (a, b): Contact and collision clusters by attention features (p < .01)

The discrepancies between our data-driven classifications and traditional sport categories underscore several key considerations. First, they suggest that the cognitive and behavioral effects of sports participation may not align directly with levels of physical contact. Second, they highlight the potential significance of individual variability in cognitive-behavioral patterns, which may rival or exceed sport-specific influences. Third, they point to the likelihood that position-specific differences within sports, such as the contrast between a football kicker and a lineman, were obscured in our analysis, despite potentially vast differences in head impact exposure. Overall, these findings illustrate the complexity of classifying sports based on cognitive-behavioral outcomes rather than physical attributes. Although sleep-related features showed slightly better alignment with conventional sport categories, particularly for football, neither feature set produced clusters that clearly mapped onto traditional binary classifications.

Discussion

Methodological Limitations

While our statistical analysis reveals meaningful patterns, several methodological limitations must be acknowledged. The most prominent constraint is the underrepresentation of collision sports in our dataset. Football served as the primary reference point for evaluating model performance in classifying collision sports; however, this reliance on a single sport, particularly one with relatively few participants compared to contact sports, introduces uncertainty regarding the robustness and generalizability of our classification models. Overall, the small sample size limits our ability to confidently validate the accuracy of our clustering approaches for collision sports.

Additionally, our interpretations hinge on assumptions inherent to the clustering process. The analysis presumes that clusters form based on differences in RHI exposure across sports. However, a more nuanced view suggests these clusters may not neatly categorize sports as "collision" or "contact," but rather capture cognitive-behavioral patterns associated with higher levels of sports participation and potential RHI exposure. This interpretation calls for caution in overgeneralizing the meaning of cluster membership.

Another significant challenge arises from the influence of non-athlete data on classification outcomes. As illustrated in Figure 11, the inclusion of non-athletes introduces potential confounding factors and biases, complicating interpretations. Even when non-athletes are excluded, complex correlations persist, suggesting that the relationship between cognitive-behavioral profiles and sports participation extends beyond simplistic models of RHI exposure, such as concussion history.

Of particular concern is the observed inconsistency across clustering methods. As shown in Figures 12 and 13, small methodological changes, such as applying stricter thresholds for feature selection, can dramatically shift classification outcomes. For instance, this was observed with football as changing the threshold for feature significance reversed the categorization. This instability highlights the importance of precise feature selection, robust modeling techniques, and larger datasets in developing reliable sports classification systems. Future research should prioritize rigorous feature selection protocols, expanded data collection efforts, and the use of psychologically validated metrics to enhance classification consistency and reliability.

Validation of Binary Classification System

Despite these limitations, our findings provide empirical support for the prevailing binary classification system in sports psychology. The PCA-based cluster evaluation metrics in Figure 7 indicate that a two-cluster solution offers the best fit for the data. This result aligns with the established psychological consensus that sports can be broadly classified based on physical contact intensity. Figure 8 further reinforces this conclusion, showing that clusters formed using concussion-relevant features display clear separation. These results validate both the existing classification framework and our clustering approach, affirming that cognitive-behavioral measures can effectively reflect differences in RHI exposure between sport types.

Comparative Performance of Sleep and Attention Features

One of the key findings of our analysis is the superior performance of sleep-related features over attention-related features in certain contexts. As demonstrated in Figure 6, sleep features showed stronger associations with concussion history than attention features. This distinction became more pronounced when non-athletes were excluded, with sleep-based clustering revealing a higher mean concussion frequency in the “collision” cluster. Furthermore, Figure 13 indicates that classifications derived from sleep features are more closely aligned with traditional expectations, particularly in categorizing football as a collision sport.

That said, attention-related features exhibited advantages in other areas. They achieved higher silhouette scores and, when non-athletes were included, showed greater average concussion frequency in the collision cluster. These contrasting results underscore the complex and multifaceted relationship between cognitive domains and RHI exposure. They also highlight the need for a broader examination of other potentially relevant cognitive domains, such as memory, mood, and fatigue, that may enhance classification models and improve sensitivity to RHI-related patterns. Finally, the current attention metrics contain overlapping constructs, suggesting that a more diverse and refined feature set could enhance classification precision.

Future Work

Our findings indicate meaningful connections between sports participation, concussion history, and cognitive-behavioral features, particularly in sleep and attention measures. However, the clusters identified do not overwhelmingly align with current contact and collision sport classifications. While the collision cluster contains a majority of football players, it is not an overwhelming majority. Similarly, sports like basketball and soccer show distribution across both clusters rather than clear alignment with a single category. These observations suggest a more nuanced approach is necessary.

A promising direction for future research would be to explore how specific positions within sports impact RHI exposure and corresponding cognitive-behavioral outcomes. If subsequent research included position-specific data alongside sport participation, we could establish more precise connections between RHI, cognitive-behavioral features, and sport participation. This approach might ultimately lead us to reconsider the current classification schema altogether, replacing broad sport categories with position-based analytics that better reflect actual impact exposure. This refinement makes intuitive sense—players such as kickers, punters, long snappers, and safeties in football experience significantly different levels of contact than other positions. Similar positional variations exist across many sports. Therefore, incorporating player positions into our analysis would substantially expand this work's value and provide a clearer delineation between risk profiles among athletes.

The integration of neurological assessments, particularly electroencephalogram (EEG) data, could further characterize cognitive-behavioral differences among athletes across various sports and positions. EEG measurements would provide valuable insights into the cognitive impact of different sports, the effects of RHI on brain function, and the connection to cognitive-behavioral characteristics. Self-reported cognitive-behavioral features have inherent limitations; adding objective, continuous data collection methods like EEG would enable more comprehensive cognitive analysis. This approach might reveal underlying neurological patterns that differ from self-reported attributes, potentially identifying more meaningful similarities or differences between athletes who otherwise report similar cognitive-behavioral experiences.

Future research should significantly expand the dataset to include greater sport variety (especially collision sports beyond football), more participants, comprehensive RHI data, and, critically, sports positions that likely influence concussion risk and cognitive-behavioral outcomes. Our current dataset has several limitations: its relatively small size, potential validity concerns due to the nationwide paid online survey methodology, limited diversity in collision sports (requiring more representation from ice hockey, lacrosse, combat sports, etc.), and limited information on RHI exposure. An ideal future dataset would include several thousand participants across 40+ sports with better representation of collision sports, detailed position information, and more comprehensive RHI data, such as vision issues, migraine history, dizziness episodes, medical consultations for head impacts, and suspected concussion frequency. Combined with the aforementioned EEG data, such comprehensive information would elevate our understanding of how different sports and positions affect cognition and athlete safety.

References

Alosco, M. L. (2018). Age of first exposure to tackle football and chronic traumatic encephalopathy. *Annals of Neurology, 83*(5), 886–901. https://doi.org/10.1002/ana.25245

Berisha, V., Wang, S., LaCross, A., Liss, J., & Garcia-Filion, P. (2017). Longitudinal changes in linguistic complexity among professional football players. *Brain and Language, 169*, 57–63. https://doi.org/10.1016/j.bandl.2017.02.003

Boston University Chobanian Avedisian School of Medicine. (2023). *Researchers find CTE in 345 of 376 former NFL players studied.* https://tinyurl.com/BUCASM

Dubas, R. L., Teel, E. F., Kay, M. C., Ryan, E. D., Petschauer, M., & Register-Malik, J. K. (2020). Comparison of concussion sideline screening measures across varying exertion levels within simulated games. *Journal of Sport Rehabilitation, 30*(1), 90–96. https://doi.org/10.1123/jsr.2019-0307

Hallock, H., Mantwill, M., Vajkoczy, P., Wolfarth, B., Reinsberger, C., Lampit, A., & Finke, C. (2023). Sport-related concussion. *Neurology: Clinical Practice, 13*(2), e200123. https://doi.org/10.1212/CPJ.0000000000200123

Marsh, H. W., & Kleitman, S. (2003). School athletic participation: Mostly gain with little pain. *Journal of Sport & Exercise Psychology, 25*(2), 205–228.

Montenigro, P. H., Alosco, M. L., Martin, B. M., Daneshvar, D. H., Mez, J., Chaisson, C. E., Nowinski, C. J., Au, R., McKee, A. C., Cantu, R. C., McClean, M. D., Stern, R. A., & Tripodis, Y. (2017). Cumulative head impact exposure predicts later-life depression, apathy, executive dysfunction, and cognitive impairment in former high school and college football players. *Journal of Neurotrauma, 34*(2), 328–340. https://doi.org/10.1089/neu.2016.4413

Nationwide Children's Hospital. (n.d.). *Concussion clinic.* Retrieved May 13, 2025, from https://www.nationwidechildrens.org/specialties/concussion-clinic

Oldham, J. R., Bowman, T. G., Walton, S. R., Beidler, E., Campbell, T. R., Smetana, R. M., Munce, T. A., Larson, M. J., Cullum, C. M., Bushaw, M. A., Rosenblum, D. J., Cifu, D. X., & Resch, J. E. (n.d.). Sport type and risk of subsequent injury in collegiate athletes following concussion: A LIMBIC MATARS Consortium investigation. *Brain Injury*, 1–9. https://doi.org/10.1080/02699052.2024.2310782

STAT 800. (n.d.). *5.5—Hypothesis testing for two-sample proportions.* Penn State Eberly College of Science. Retrieved May 13, 2025, from https://online.stat.psu.edu/stat800/lesson/5/5.5

Sullivan, J. R., & Riccio, C. A. (2010). Language functioning and deficits following pediatric traumatic brain injury. *Applied Neuropsychology, 17*(2), 93–98. https://doi.org/10.1080/09084281003708852

Wellm, D., Jäger, J., & Zentgraf, K. (2024). Dismissing the idea that basketball is a “contactless” sport: Quantifying contacts during professional gameplay. *Frontiers in Sports and Active Living, 6*. https://doi.org/10.3389/fspor.2024.1419088