IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Athlete Statistics Visualization and Prediction

A. Patel [AU2140118][1], A. Prajapati [AU2140090][2], B. Shah [AU2140088][3], and S. Virani [AU2140139][4]

[1,2,3,4]are with School of Engineering and Applied Science, Ahmedabad University, Gujarat, India
(email: arpit.p@ahduni.edu.in, aryan.p3@ahduni.edu.in, buland.s@ahduni.edu.in, sujal.v@ahduni.edu.in).

ABSTRACT This project delves into the realm of sports data analytics with a focus on the performance analysis of a Division-1 basketball team during the tumultuous period of the COVID-19 pandemic. Leveraging a multi-modal dataset encompassing physical, physiological, and performance metrics, our endeavor aims to extract actionable insights to enhance athlete performance and mitigate injury risks. The methodology employed involves rigorous data preprocessing, including MICE imputation and model-fit based weighing for feature engineering, to ensure the suitability of the dataset for predictive modeling. While initial attempts to address class imbalance through SMOTE trials faced challenges due to the continuous nature of modalities, alternative approaches were explored. Predictive modeling using Random Forest, XGBoost, and ensemble techniques revealed promising accuracies upon training with data aggregated from multiple athletes across the 2021-2022 seasons. Our iterative approach underscores the dynamic nature of sports analytics, where challenges encountered pave the way for innovation and refinement, ultimately contributing to a deeper understanding of athlete performance dynamics and informed decision-making in sports.

INDEX TERMS Basketball, collegiate athletes, data mining, sports analytics, machine learning, game performance, injury prediction

I. INTRODUCTION

The burgeoning field of sports data analytics continues to gain traction, with its applications extending to collegiate sports and professional leagues alike, including Esports. Recognizing the profound impact of deriving insights into athletic performance, teams are increasingly turning to data-driven methodologies to enhance game outcomes and mitigate injury risks. This interdisciplinary endeavor brings together athletes, coaches, exercise scientists, engineers, and data analysts, underscoring the collaborative nature of sports analytics.

The focal point of this study is the analysis of a Division-1 basketball team's season, marked notably by the disruptions caused by the COVID-19 pandemic. Amidst these challenges, our project embarks on the comprehensive examination of various facets of athlete performance. By delving into daily sleep patterns, training regimens, cardiac rhythm, jump metrics, and cognitive patterns, we aim to unravel the intricate dynamics influencing athletic performance.

Building upon prior groundwork, our project undertakes a strategic approach to harnessing machine learning techniques for predictive modeling. Preceding the mid semester milestone, significant progress has been made in data cleaning and preprocessing, culminating in the adoption of MICE imputation to address missing values. Moving forward, our roadmap entails implementing SMOTE for data balancing and leveraging Random Forest and XGBoost algorithms for predictive analytics across five modalities: sleep, training, cardiac rhythm, jump, and cognitive performance.

By integrating objective data streams with subjective training feedback and survey responses, our initiative aspires to furnish actionable insights for coaches and athletes alike. The ultimate goal is to empower decision-making processes, enabling teams to optimize performance strategies and minimize injury susceptibility in a dynamic and demanding sporting landscape.

II. METHODOLOGY

Following the mid semester milestone, our project progressed into the implementation phase, focusing on the execution of planned tasks aimed at enhancing the predictive capabilities of our machine learning models. The methodology employed during this phase can be summarized as follows:

A. *Model-Fit Based Weighing for Feature Engineering:*

Before proceeding with the SMOTE trials, we needed to create 5 columns for each of the 5 modalities mentioned in the project problem statement so as to provide columns that need data balancing for the SMOTE algorithm; this was a necessary step as these 5 columns would be further used for prediction these modalities. In order to create columns for each modality we took under consideration the best parameters that influenced a particular modality and assigned weights to each one of them. Further, using existing information from empirical results and research findings and news, we adjusted the values of the weights of each parameter and from that derived functions for each of the modality columns. We then generated these modality columns that represented data values representing scores over a range, indicating the continuous nature of the distribution.

Sleep	Training	Cardiac Rhythm	Jump	Cognitive
Hours.in.Bed	Training.load.score	RHR (Resting Heart Rate)	Jump.Height	Recovery
Hours.of.Sleep	Cardio.load	HR.min..bpm.	Peak.Power	Sleep.Score
Sleep.Need	Duration	HR.avg..bpm.	Peak.Power.BM	Total.distance..m.
Sleep.Efficiency....	HR.min..bpm.	HR.max..bpm.	RSI (Relative Strength Index)	Distance...min..m.min.
Sleep.Disturbances	HR.avg..bpm.			Maximum.speed..km.h.
Latency..min.	HR.max..bpm.			Average.speed..km.h.
Total.Cycle.Sleep.Time..hours.	RT.Volume.Load			Game.Score
Restorative.Sleep..hours.				
Sleep.Consistency				
Respiratory.Rate				

FIGURE 1. **The best parameters for each of the 5 modalities**

Procedure Instance (for suppose the modality of 'Jump'):

Let's assume a linear combination of these features with some weights:

Jump = (w1 x Jump.Height) + (w2 x Peak.Power) + (w3 x Peak.Power.BM) + (w4 x RSI)

Here, w1, w2, w3, and w4 are weights that determine the importance of each feature in predicting jump performance. These weights can be estimated through empirical analysis or expert knowledge.

Adjusting the weights of the features requires understanding how each feature may contribute to each modality. Thus, it's essential to conduct thorough research or consult domain experts for more accurate weighting. Following are key notes from the sample adjustment of weights process for the modality of 'Jump':

- We updated the weights based on their potential impact on jump performance, considering factors such as biomechanics, power output, body composition, and strength.

- The weights are now distributed differently, with higher importance given to features that are believed to have a stronger influence on jump performance.

- It's crucial to note that these weights are based on empirical evidence from research studies and general

insights from sports science literature and studies related to physical activity and exercise performance.

Weights Validation:
Some general insights based on common trends observed in sports science literature and studies related to physical activity and exercise performance:

1. Jump.Height:
- Research suggests that jump height is a fundamental measure of explosive power and athletic performance. Studies have shown that training interventions aimed at improving jump height can lead to enhancements in various athletic abilities, including speed, agility, and overall sports performance. Therefore, assigning a relatively higher weight to Jump.Height (e.g., 0.45) reflects its critical role in assessing an athlete's explosive power and vertical leap, which are essential for success in sports such as basketball.

2. Peak.Power and Peak.Power.BM:
- Peak power output and power-to-body mass ratio are indicators of an athlete's ability to generate force quickly, which is crucial for explosive movements like jumping. Studies in sports science literature have consistently highlighted the importance of power-based training in improving athletic performance and jump height. Assigning relatively high weights to Peak.Power and Peak.Power.BM (e.g., 0.30 and 0.15, respectively) acknowledges their significant contribution to jump performance and reflects their importance in evaluating an athlete's power capabilities.

3. RSI (Relative Strength Index):
- The relative strength index reflects an athlete's strength relative to their body mass and can provide insights into their overall physical readiness and neuromuscular efficiency. Research has shown that strength training and neuromuscular conditioning are essential components of jump performance and athletic development. Therefore, assigning a moderate weight to RSI (e.g., 0.10) acknowledges its role in assessing an athlete's strength and neuromuscular function, which are important determinants of jump performance.

These adjustments are based on empirical evidence and general insights from sports science literature, highlighting the importance of explosive power, power-to-body mass ratio, and neuromuscular efficiency in predicting jump performance. It's essential to validate these adjustments with empirical data and expert consultation to ensure their relevance and accuracy in the specific context of basketball and individual athlete characteristics. Conducting controlled experiments or longitudinal studies within basketball teams can provide further insights into the relationships between these features and jump performance in athletes.

Similar approach and steps for validations for remaining 4 modalities are as follows:

1. Sleep Modality:
- Review studies on athlete sleep patterns and sleep quality assessment methods to validate the adjusted weights.
- Consult sleep experts or sports scientists to assess the weighting scheme's consistency with established sleep quality indicators in athletes.
- Compare the predicted sleep scores with self-reported sleep quality or objective sleep measures (e.g., actigraphy) from basketball players to evaluate the weighting scheme's effectiveness in predicting sleep quality.

2. Training Modality:
- Evaluate the adjusted weights based on research findings on training load monitoring and performance assessment in athletes.
- Consult sports scientists or coaches to determine the weighting scheme's alignment with established training principles and athlete performance metrics.
- Compare the predicted training intensity scores with recorded training loads or perceived exertion ratings from basketball players to assess the weighting scheme's validity in quantifying training intensity.

3. Cardiac Rhythm Modality:
- Validate the adjusted weights by reviewing research on heart rate variability and cardiovascular health in athletes.
- Seek feedback from sports scientists or medical professionals specializing in sports cardiology to evaluate the weighting scheme's relevance to cardiac rhythm assessment in athletes.
- Compare the predicted cardiac rhythm scores with heart rate data collected during training or competition to determine the weighting scheme's accuracy in predicting cardiac rhythm variations.

4. Cognitive Modality:
- Assess the adjusted weights based on research on exercise effects on cognitive function and performance assessment in athletes.
- Consult cognitive scientists or sports psychologists to evaluate the weighting scheme's consistency with established cognitive performance indicators in athletes.
- Administer cognitive tests or assessments to basketball players and compare the predicted cognitive performance scores with test results to validate the weighting scheme's effectiveness in quantifying cognitive performance.

By following these validation steps and considering feedback from relevant experts, we can assess the reliability and accuracy of the adjusted weights for each modality in

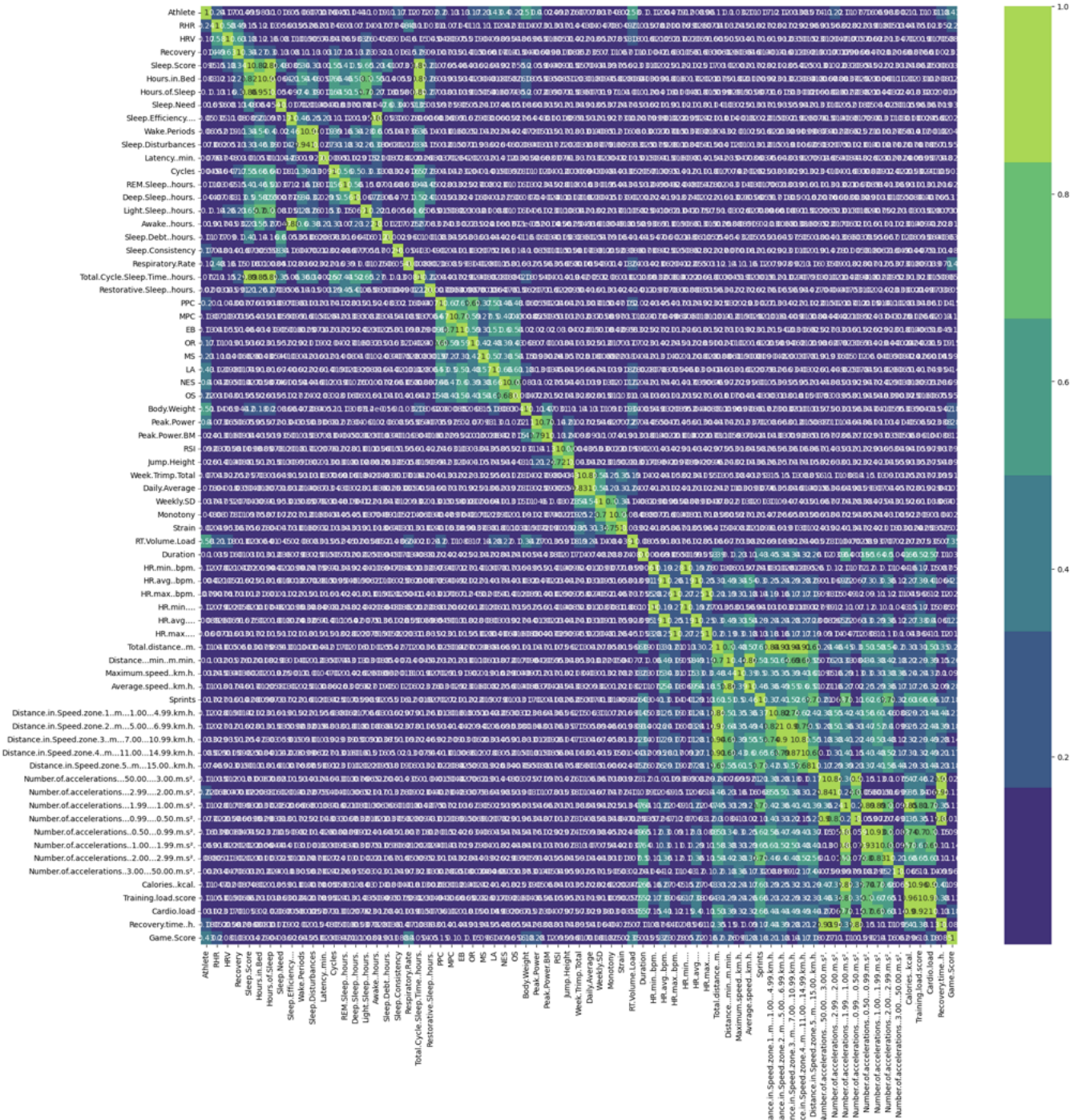


FIGURE 2. **The heatmap showcases the correlation between variables in the dataset. Each square denotes the correlation coefficient, with colors ranging from green to yellow based on the "viridis" color palette. Annotated values display correlation coefficients, aiding in understanding variable relationships.**

predicting sleep quality, training intensity, cardiac rhythm, and cognitive performance in basketball players.

B. *SMOTE Trials for Data Balancing:*

Recognizing the significance of balanced datasets in mitigating class imbalance issues, we integrated Synthetic Minority Over-sampling Technique (SMOTE) into our

workflow. SMOTE facilitates the generation of synthetic samples for minority classes, thereby rebalancing the dataset and mitigating the adverse effects of class imbalance on model performance. However, while conducting the SMOTE trials, based on our observations and lack of pre-sights, we had to drop down the idea of utilizing SMOTE since it further hindered the path of how our predictive statistics would be visualized.

C. *Prediction Using Models Boosted By Multiple Decision Trees:*

Looking ahead, the project's trajectory converges on modeling and data prediction, with a keen focus on evaluating collected data against game statistics. The essential aim for post-mid semester project work remained in the use of methodology that internally supports multiple boosted decision trees as the base models; models satisfying these are XGBoost and Random Forest. Random Forest is known for providing interpretability; it is easier to compute contribution of features to decision since Random Forest makes use of change in accuracy to find feature importance when the feature in consideration is excluded from decision.

After dropping the SMOTE implementation, we added the date column into the dataset back since we decided to split the training and testing data based on the years: training data would consist of the years 2021-2022 while the test data would contain information from the year 2023. This essentially partitioned the dataset into a 80%-20% ratio. We then made use of five modeling techniques for prediction: Random Forest, XGBoost and Ensemble XGBoost + Random Forest, Stacking | GBM + XGBoost and Ensemble XGBoost + Random Forest + Gradient Boost.

We trained the algorithms with the data of all athletes for the years 2021-22 and for the predictive analysis, we predicted the individual modalities of a specified athlete. This filtering in the dataset was performed on the basis of the identification values (data type: integer) of the "Athlete" column. To study the model accuracy, we calculated the following evaluation parameters: MSE (Mean Squared Error), MAE (Mean Absolute Error) and R2 Score (R-Squared Score).

III. DISCUSSION

The journey from conceptualization to implementation revealed several key insights and challenges inherent in the domain of sports data analytics. Our discussion encapsulates reflections on the methodologies employed, lessons learned, and avenues for future exploration within the project scope.

A. *Data Preprocessing Complexities*

Initial assumptions regarding the complexity of predictability testing and model training were challenged as we delved into the intricacies of data preprocessing. The significance of first-hand data handling emerged prominently, underscoring its pivotal role in shaping the efficacy of subsequent analysis and predictive modeling endeavors. Notably, the adoption of advanced techniques such as MICE imputation for handling missing data underscored the importance of leveraging machine learning algorithms iteratively throughout the data preprocessing phase.

B. *Addressing Class Imbalance*

The endeavor to mitigate class imbalance issues through SMOTE trials unveiled challenges and insights regarding oversampling techniques. While SMOTE initially seemed promising, its implementation revealed limitations in effectively addressing class imbalance, prompting the exploration of alternative approaches such as ENN. The integration of Edited Nearest Neighbor technique sought to refine the oversampled dataset by selectively removing majority class instances with neighbors belonging to different classes, thereby enhancing model robustness and performance. However, another hindrance lay in the way; Based on references from the suggested research papers, our developed initial plan of implementing SMOTE had to be called off midway due to the fact that the modality columns on which we were applying data balancing were of continuous type and SMOTE supports minority class synthesis of only those parameters that are classified and not continuous. Since we needed to predict the individual scores (data type: integer) for different athletes we decided to abort the SMOTE trials.

C. *Prediction Challenges*

When we first commenced the prediction phase, we decided to train the models with a dataset of a specific athlete only, for the years 2021-22 and thus for testing, we utilized the dataset of the same athlete for the year 2023. However, the accuracy scores were not promising and thus we decided to train the models on the dataset of all athletes for the years 2021-22 in order to split the data in 80%-20% form so as to increase the accuracy of the predictions. The evaluation parameters then displayed results denoting notable good levels of accuracy of the predicted scores with respect to the actual scores of the modalities from the testing dataset.

In essence, the discussion illuminates the iterative nature of the project's progression, wherein challenges encountered along the way serve as springboards for innovation and refinement. As we navigate the evolving landscape of sports data analytics, each lesson learned and methodology explored contributes to a richer understanding of athlete performance dynamics and the potential for

data-driven insights to drive informed decision-making in the sporting arena.

IV. CONCLUSION

The culmination of our project underscores the transformative potential of data analysis and predictive modeling in revolutionizing sports management practices. By harnessing the power of machine learning algorithms, we have embarked on a journey to unlock valuable insights into player performance dynamics, with far-reaching implications for roster composition, gameplay strategies, and injury prevention strategies.

Unlocking Player Performance Insights:
Through rigorous data analysis and modeling, we have unraveled invaluable insights into player efficiency and performance metrics. These insights serve as a cornerstone for informed decision-making, enabling management to tailor formations and strategies based on individual player capabilities and tendencies. Moreover, the ability to identify and address weaknesses through targeted training regimens enhances player development and overall team performance.

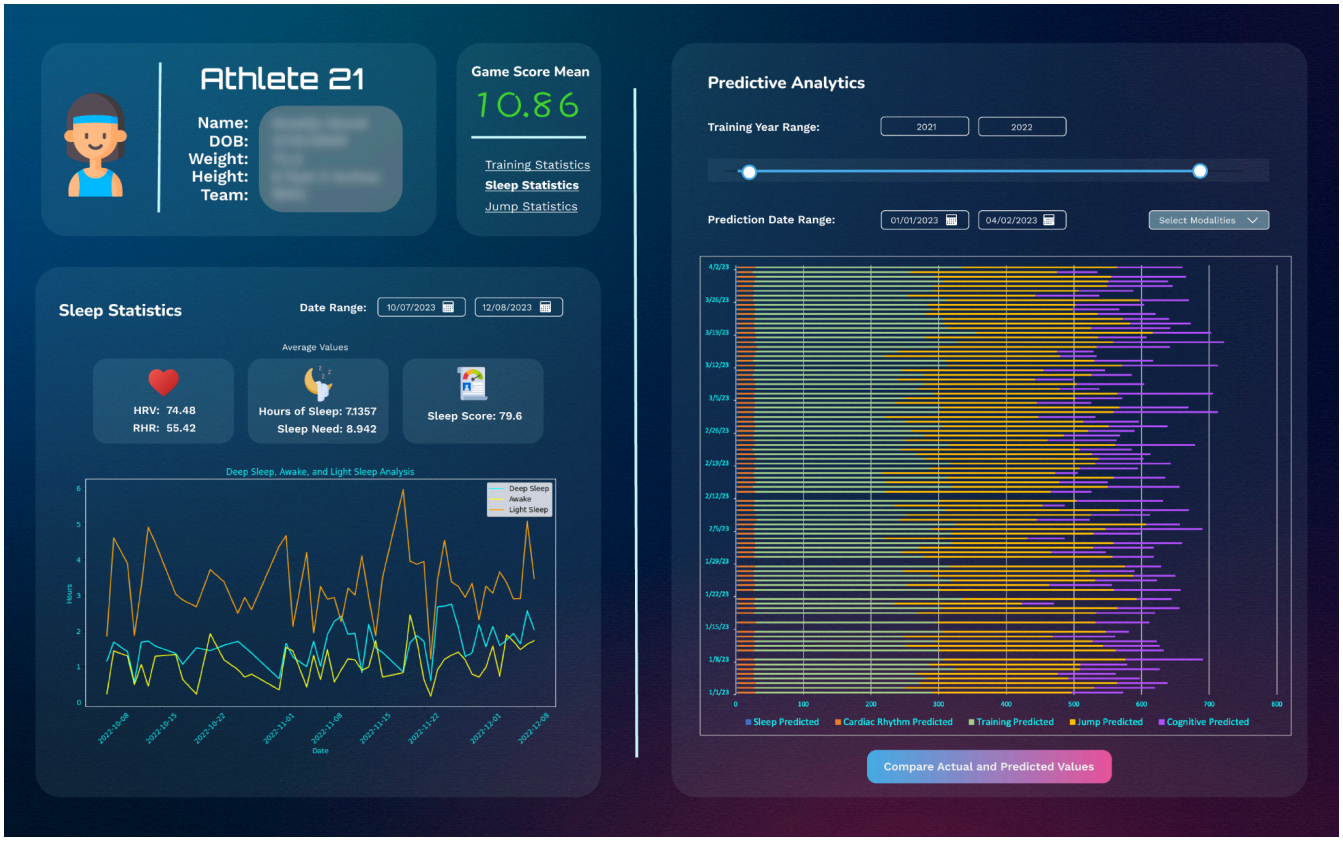
Enhancing Injury Prediction and Prevention:
A pivotal outcome of our endeavor lies in the realm of injury prediction and prevention. By leveraging machine learning techniques, we have endeavored to anticipate and mitigate potential injury risks, thereby safeguarding player well-being and minimizing disruptions to team dynamics. The proactive identification of injury-prone athletes empowers management to implement preemptive measures and interventions, fostering a culture of player welfare and longevity.

Empowering Data-Driven Decision-Making:
Central to our project's ethos is the ethos of data-driven decision-making. The insights gleaned from our analyses serve as a compass for strategic planning and execution, empowering coaches, and management to optimize roster compositions, devise effective gameplay strategies, and manage player flow during games. By aligning decisions with empirical evidence and predictive models, teams can gain a competitive edge and maximize performance outcomes on and off the field.

In conclusion, our project exemplifies the symbiotic relationship between data analytics and sports management, wherein the fusion of empirical insights and predictive modeling techniques catalyzes innovation and excellence in athletic endeavors. As we continue to traverse the ever-evolving landscape of sports analytics, the lessons learned and methodologies pioneered pave the way for a future where data-driven insights drive sustainable success and elevate the performance paradigm in the realm of sports management.

Models	Evaluation Score	Sleep	Training	Cardiac Rhythm	Jump	Cognitive
Random Forest	*MSE (Mean Squared Error)*	0.00138446371244694	2.00581065809911	0.072711517441861	0.0287257322049844	0.297774226067874
	MAE (Mean Absolute Error)	0.029545170872093	1.16528289036543	0.193686046511628	0.10806355813955	0.418128073089713
	R- Squared (R^2) Score	0.912573715234629	0.998079667588118	0.960762749029418	0.999960501291569	0.999529325542541
XGBoost Regression	*MSE (Mean Squared Error)*	0.00048003088448091	1.92848881053008	0.0406278638684747	0.402126404235848	0.360787597088553
	MAE (Mean Absolute Error)	0.0170034144371387	1.15667167790306	0.158198919961618	0.349755207292421	0.364503793811486
	R- Squared (R^2) Score	0.969686950675921	0.998153694341059	0.978076022243923	0.999447064622072	0.999429723959793
Ensemble XGBoost + Random Forest	*MSE (Mean Squared Error)*	0.000664658199363002	1.40983580815983	0.0405198087089743	0.10264086799062	0.295341000720988
	MAE (Mean Absolute Error)	0.0202133637559491	0.90719231974956	0.14391843467535	0.181487092835895	0.329828903370596
	R- Squared (R^2) Score	0.958028082291557	0.998650244784118	0.978134331952771	0.999858865852788	0.999533171600795
Stacking | GBM + Random Forest	*MSE (Mean Squared Error)*	0.000297045310253663	1.02396693352305	0.0236786791191387	0.213956789186569	0.556577420927864
	MAE (Mean Absolute Error)	0.0122657195899431	0.821242573380043	0.115921149142953	0.212629123137716	0.339762409755163
	R- Squared (R^2) Score	0.981242146219524	0.999019669736423	0.987222295615101	0.999705803257774	0.999120250334999
Ensemble XGBoost + Random Forest + Gradient Boosting	*MSE (Mean Squared Error)*	0.000442105082681407	1.39281761318321	0.0272207797110945	0.0504194482458009	0.225540228048965
	MAE (Mean Absolute Error)	0.0173922873822539	0.92028828949341	0.116644206560173	0.139026775711396	0.326131100292132
	R- Squared (R^2) Score	0.972081893871809	0.998666537743412	0.985310875048191	0.999930671807727	0.999643501635874

FIGURE 3. **The Evaluation Metrics for each of the modality with respect to all five models used. Notable findings for Stacking | GBM + Random Forest denote better in 2 scores, almost similar in 1, and bad in 2 scores in comparison to the Ensemble XGBoost + Random Forest model. Similarly, for Ensemble XGBoost + Random Forest + Gradient Boosting, key findings denote better in 2, bad in 2 scores, and equivalent to 1, with respect to the Stacking | GBM + Random Forest model; it performs better in every aspect with respect to the Ensemble XGBoost + Random Forest model.**



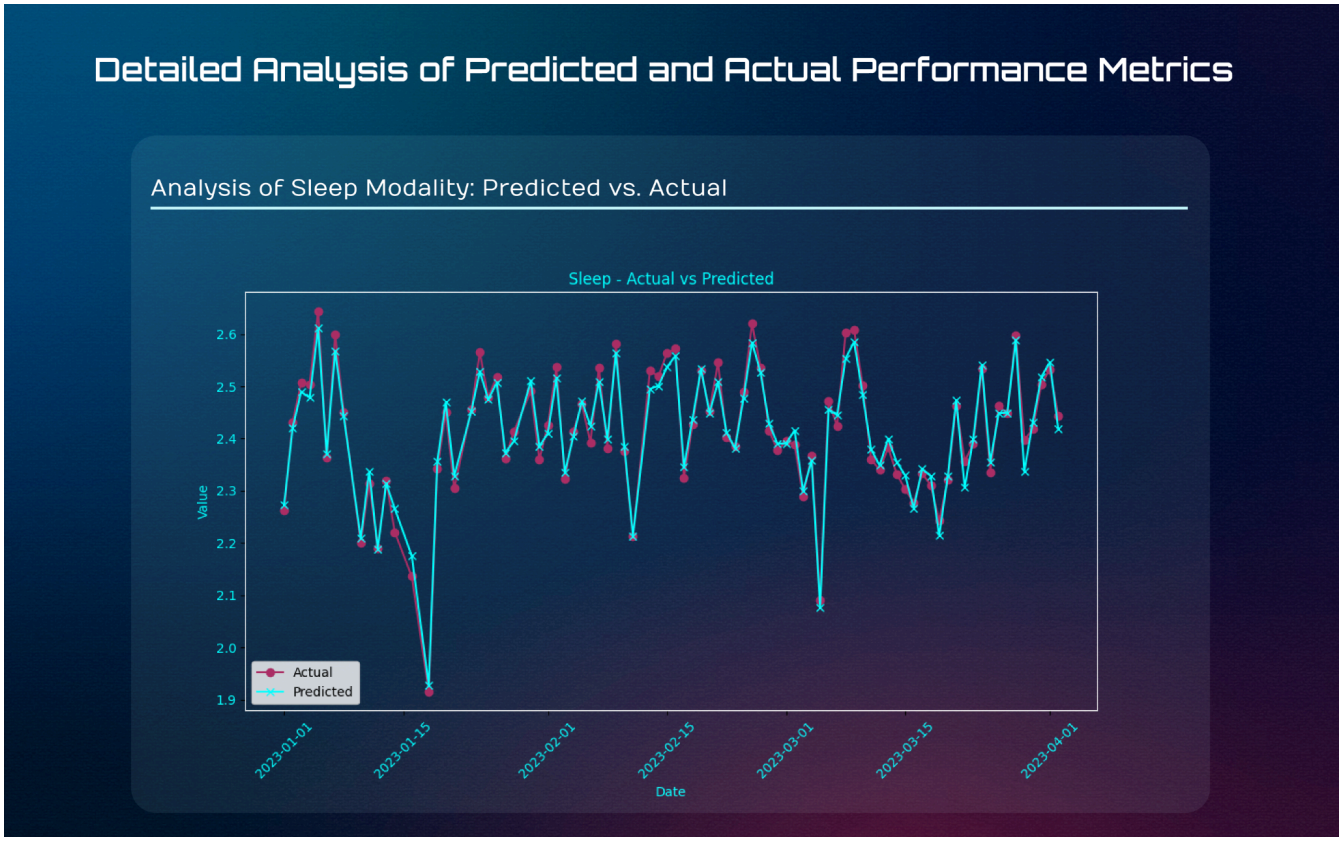


FIGURE 3. Athlete Dashboard that utilizes the concept of Data Visualization of statistics related to the five modalities along with the Predictive Analytics section that lets users predict data by segregation of training and testing data on the basis of years; the resulting accuracy graphs are shown in detail in the next screen.

V. REFERENCES

[1] *Taber, C. B., Sharma, S., Raval, M. S., Senbel, S., Keefe, A., Shah, J., ... & Kaya, T. (2024). A holistic approach to performance prediction in collegiate athletics: player, team, and conference perspectives. Scientific Reports, 14(1), 1162.*

[2] *Senbel, S., Sharma, S., Raval, M. S., Taber, C., Nolan, J., Artan, N. S., ... & Kaya, T. (2022). Impact of sleep and training on game performance and injury in division-1 women's Basketball Amidst the Pandemic. Ieee Access, 10, 15516-15527.*

[3] *Juliano, E., Thakkar, C., Taber, C., Raval, M., Kaya, T., & Senbel, S. (2023, October). A dynamic online dashboard for tracking the performance of division 1 basketball athletic performance. In 2023 IEEE 28th Pacific Rim International Symposium on Dependable Computing (PRDC) (pp. 314-318). IEEE.*

[4] *Sharma, S. U., Divakaran, S., Kaya, T., & Raval, M. (2022, July). A Hybrid Approach for Interpretable Game Performance Prediction in Basketball. In 2022 International Joint Conference on Neural Networks (IJCNN) (pp. 01-08). IEEE.*

[5] *Ibáñez, S. J., Sampaio, J., Feu, S., Lorenzo, A., Gómez, M. A., & Ortega, E. (2008). Basketball game-related statistics that discriminate between teams' season-long success. European journal of sport science, 8(6), 369-372.*

[6] *Sarlis, V., & Tjortjis, C. (2020). Sports analytics—Evaluation of basketball players and team performance. Information Systems, 93, 101562.*

[7] *Li, B., & Xu, X. (2021). Application of artificial intelligence in basketball sport. Journal of Education, Health and Sport, 11(7), 54-67.*

[8] *Sarlis, V., & Tjortjis, C. (2020). Sports analytics—Evaluation of basketball players and team performance. Information Systems, 93, 101562.*

[9] *Nagarajan, R., & Li, L. (2017, November). Optimizing NBA player selection strategies based on salary and statistics analysis. In 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech) (pp. 1076-1083). IEEE.*

[10] *Cao, C. (2012). Sports data mining technology used in basketball outcome prediction.*

[11] *Imbach, F., Perrey, S., Chailan, R., Meline, T., & Candau, R. (2022). Training load responses modelling and model generalisation in elite sports. Scientific Reports, 12(1), 1586.*

[12] *Sarto, F., Impellizzeri, F. M., Spörri, J., Porcelli, S., Olmo, J., Requena, B., ... & Franchi, M. V. (2020). Impact of potential physiological changes due to COVID-19 home confinement on athlete health protection in elite sports: a call for awareness in sports programming. Sports medicine, 50, 1417-1419.*

[13] *Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.*

[14] *Mera-Gaona, M., Neumann, U., Vargas-Canas, R., & López, D. M. (2021). Evaluating the impact of multivariate imputation by MICE in feature selection. Plos one, 16(7), e0254720.*

[15] *Madley-Dowd, P., Hughes, R., Tilling, K., & Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. Journal of clinical epidemiology, 110, 63-73.*