Modelling and Prediction of Athletic Readiness based on Sleep and Recovery Patterns

Krina Khakhariya, Urjit Mehta, Brijesh Munjiyasara, Khushi Agrawal, and Jafri Syed Mujtaba
Ahmedabad University
Email: {krina.k, urjit.m, brijesh.m, khushi.a2, and jafri.h}@ahduni.edu.in

***Abstract*—Athletic readiness is a primary driver of performance, especially in high-intensity sports such as collegiate basketball, where athletes balance regular games, travel, rigorous training, and class schedules. These demands usually result in sleep deprivation and poor recovery, both of which are detrimental to physical and mental performance. In this study, we try to model and forecast an athlete's readiness score (RSImod) from sleep and recovery behaviors using machine learning methods. Different imputation methods such as mean, median, mode, KNN, interpolation, MICE, and EM are applied in the dataset for preprocessing to ensure data quality. Different predictive models such as Linear Regression, Ridge and Lasso Regression, Random Forest, and Gradient Boosting (XGBoost, LightGBM) are trained and tested using MAE, RMSE, and R^2 score. Explainable AI (xAI) methods are also employed to model explainability and to identify the most important factors driving readiness. Through an analysis of the interaction between sleep, recovery, and performance, the study offers actionable insights for sports scientists and coaches to optimize training schedules, reduce injury risk, and maximize player performance. The study emphasizes the contribution of sleep management strategies in competitive sports and suggests a data-driven approach for monitoring athlete welfare.**

***Index Terms*—Data mining, Regression, Machine Learning, Athletic Readiness, Sleep and Recovery Pattern, Sport Analytics**

I. Introduction

In the world of competitive sports, an athlete's performance and recovery are the two most crucial factors if the athlete is to succeed. Naturally, sleep is the main factor in human readiness optimization and also the obvious one in athlete injuries reduction. The college basketball season, along with weekly games and training, usually disrupts player sleep patterns, causing them to feel tired and leading to longer recovery times. The study will respect the importance of sleep in the lifetime of college basketball players and will describe sleep and recovery tendencies. Students whose lives depend on the examination of their well-being will feel that all that is being talked about matters to them. In the present case, therefore, the author is concerned with the time a student has to relax and try to apply what's been learned in schools.

The most important end result of the study will be the thorough analysis of the consequences of deficient sleep. Sleep quality will be tested most accurately by conducting overnight sleep studies at the beginning of the study, periodical web-based questions over a seven-day period, and following that, using sleep diaries and actigraphy for weeks at a time. Sample data from the sensors are transferred to a microcomputer board, for feature extraction, digital signal processing, and machine learning, to achieve readiness quantification. Another critical augury is focused on stress levels as they are determining the athlete's performance.

Estimation of models' performance through key metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 Score would enable us to understand the strengths and weaknesses of these prediction models In consequence guiding the most accurate approach to athlete readiness forecasting. This study, by using machine learning algorithms along with sleep and recovery data, has become a part of the increasing body of knowledge on athlete performance optimization. This study's results might be a beneficial factor, helping to design personalized training schedules, improving recovery techniques and overseeing the reduction of injury chances among student-athletes.

A. Related Works

A new research was carried out by Mah et al. (2011) that became a landmark in the field of the effects of sleep extension on the athletic performance of collegiate basketball players [1]. The study discovered that along with longer sleep, the efficiency of the reaction time was raised remarkably, the feeling of sleep oversleeping was dropped, the increase in sprint speed was observed, all the shots were aimed at the target, and the mood was better managed. This research stressed the importance of good sleep as a vital part of physical exercise by providing early evidence that sleeplessness is the main factor responsible for mental and physical disorders of athletes. The researchers' results were indicative of sleep management strategies in the sports training programs, which were an encouragement for athletes to get better performance outputs.

The study of Burke et al. (2020) was an extension of the study of Mah et al. (2011). Its main focus was to determine the extent of the relationship between sleep quality and the risk of injury among college football players. The study, supported by logistic regression models, revealed the indirect relationship between low sleep quality and injury [2]. Instead of the solitary role of sleep deprivation, this study provided evidence that besides fatigue, individuals take more time to recover after such periods, this being the root cause of performance decline which, in turn, increases injury susceptibility. The authors of this research indicated from just sleep amount to sleep quality, which means that the regularity of sleep as well as

its incorporation of recovery are the factors that have to be kept in control if a sportsman does not want to be injured.

Recent research was conducted by Sargent et al. (2021) who looked into the gap between athletes' actual sleep and the amount of sleep they believed they needed [3]. The probable conclusion of the survey was that most athletes miscalculations resulted in not enough rest. Hence, performance was compromised. Using wrist actigraphy, sleep diaries, and mixed-effects models, the researchers were capable of finding a very substantial relationship between insufficient sleep and the efficiency of recovery. The study also showed that the athletes must be given this information on how they would behave if they decide to sleep better and thus allow their bodies to restore and recover properly. Education on the issue of the importance of a good night's sleep for an athlete is important, not only about what they do with this extra time but also being able to evaluate and put it on top of their schedules is a priority.

Senbel et al. (2022) visited Division-1 women's basketball players and did an analysis on the implications of the joint action of sleep and training load on injuries and game performance in the case of the pandemic [4]. As opposed to previous investigations conducting casual observations, this research adopted machine learning techniques, including ensemble classifiers and imputation methods, for coming up with predictive models. Besides, the research revealed that a close connection does exist between regular sleep and reduced weariness in athletes, thus, proper sleep management should to some extent be involved in the predictive capabilities of athletes and readiness and performance.

II. Dataset

The data set consists of several sleep quality, physiological, and readiness features of athletes. It consists of features like heart rate, sleep duration, efficiency, disturbance, and different sleep cycle values as Table. I. The correlation analysis shows underlying relationships between the factors and indicates the most influential factors for sleep quality and recovery. Through analyzing these relationships, we can get a better understanding of how sleep affects the general performance and health of athletes.

III. Proposed Methodology

This study aims to predict the Readiness Score (RSImod) of university basketball players based on recovery and sleep patterns using a robust machine learning pipeline. The approach used in this study consists of five major phases: Data Preprocessing, Imputation Methods, Model Training, Evaluation Metrics, and Performance Analysis as in fig 1.

A. Data Preprocessing

The observations include athlete-specific sleep, recovery, and readiness scores. Missing values, outliers, and inconsistencies are handled prior to using machine learning models.

- Handling Missing Data: Multiple imputation procedures handle the missing values in the data.

TABLE I
Feature Description Table

Feature	Description
Athlete	Athlete identifier
Date	Date of recorded data
Day.of.Week	Day of the week
RHR	Resting Heart Rate (bpm)
HRV	Heart Rate Variability (ms)
Recovery	Readiness recovery score
Sleep.Score	Overall sleep quality
Hours.in.Bed	Total time in bed (hrs)
Hours.of.Sleep	Actual sleep duration (hrs)
Sleep.Need	Recommended sleep (hrs)
Sleep.Eminency	Sleep efficiency (%)
Wake.Periods	Wake-up frequency
Sleep.Disturbances	Sleep disruptions count
Latency.min.	Time to fall asleep (min)
Cycles	Number of sleep cycles
REM.Sleep	REM sleep duration (hrs)
Deep.Sleep	Deep sleep duration (hrs)
Light.Sleep	Light sleep duration (hrs)
Awake	Time awake during sleep (hrs)
Sleep.Debt	Sleep deficit (hrs)
Sleep.Consistency	Sleep timing consistency
Respiratory.Rate	Breathing rate (bpm)
Total.Cycle.Sleep.Time	Total sleep cycle time (hrs)
REM.Percentage	REM sleep percentage (%)
Deep.Sleep.Percentage	Deep sleep percentage (%)
Restorative.Sleep	Restorative sleep (hrs)
Restorative.Sleep.%	Restorative sleep percentage (%)
RSI	Recovery Sleep Index

- Feature Selection: Relevant sleep and recovery-related features are selected to enhance predictive ability.
- Data Normalization: Required standardization methods (e.g., Min-Max Scaling, Standard Scaling) are used.

B. Earlier Imputation Techniques

To ensure completeness of the data, various imputation techniques are employed:

- Mean, Median & Mode Imputation: Missing value imputation using central tendency measures.
- Interpolation + Forward-Fill (F-fill) & Backward-Fill (B-fill): Trend analysis gap filling.
- MICE (Multiple Imputation by Chained Equations): Statistical method to manage missingness.
- KNN (K-Nearest Neighbors) Imputation: Missing value imputation based on similarity measures.
- EM (Expectation-Maximization): Probabilistic estimation for managing missing data.

C. Current Imputation Techniques

Missing values were common in some features of the dataset due to device non-compliance (e.g., uncharged WHOOP strap, app not in use), inconsistent behavior from athletes, and weekly sampling schedules (especially for RSI). Two strategies were used to mitigate this issue:

1) **RSI Imputation:** RSI was measured weekly and had a significant number of missing values. The dataset, containing 3,100 rows, was expected to have 440 RSI entries; however, only 185 were recorded. This implies that more than 90% of RSI values were missing. To retain its temporal characteristics, backward filling was

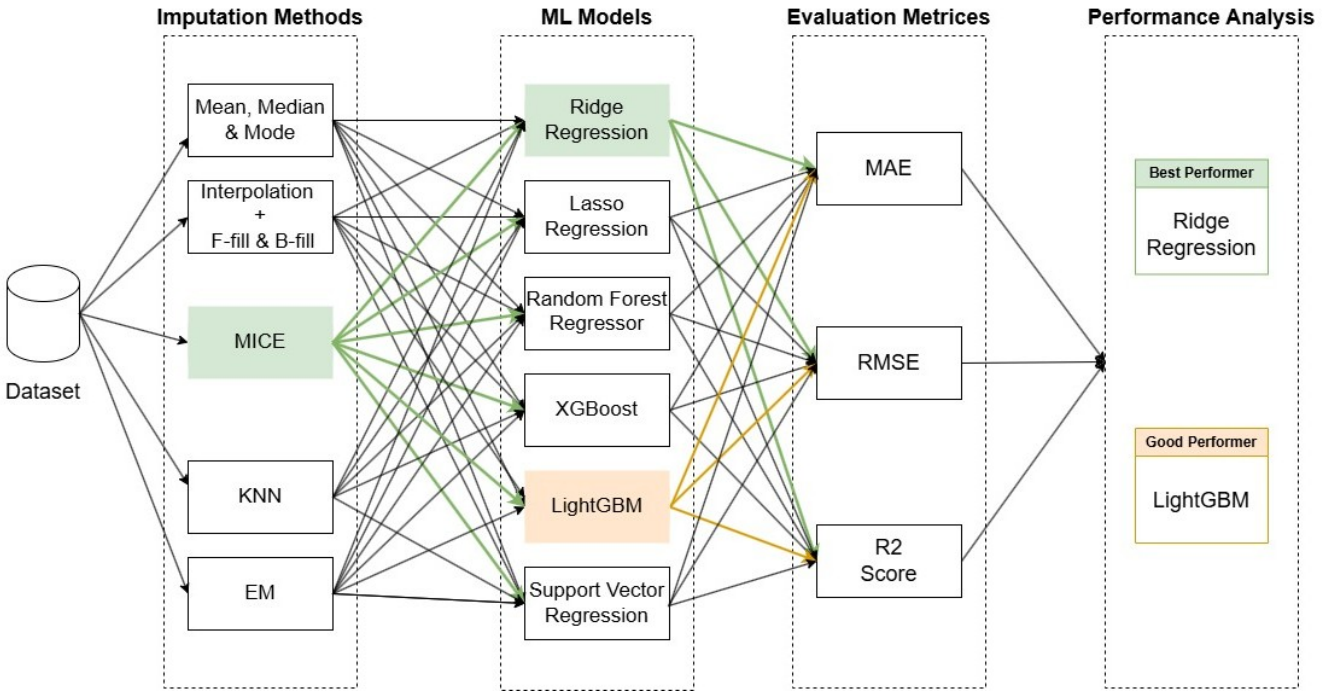

Fig. 1: Overview of the Earlier Modeling Approach

used with a maximum window of 13 days. This approach allowed RSI values to be extended within a realistic timeframe, maintaining a bi-weekly resolution. After imputation, rows with the remaining missing RSI values were removed, resulting in a dataset with 1,750 rows.

2) **MICE Imputation for Other Features:** Multivariate Imputation by Chained Equations (MICE) was used for all other numeric features, such as recovery metrics and sleep parameters. Many of these features had around 1,100 missing values (approximately 30% missingness). MICE performs iterative, feature-wise modeling using all available data and is particularly effective for datasets with high levels of missingness ($> 50\%$). Non-numeric features such as `day_of_week`, `athlete`, and `date` were excluded from imputation.

The quality of imputation was validated by comparing the mean and standard deviation of the original (non-missing) and imputed values. For each feature, the percentage error in the mean and standard deviation was computed. The errors for most features were found to be well below 5%, indicating that the imputation process preserved the underlying statistical properties and introduced minimal bias.

TABLE II
Comparison of Original and Imputed Data: Mean and Standard Deviation Errors

Feature	Orig. Mean	Imputed Mean	% Mean Err	Orig. Std	Imputed Std	% Std Err
rhr	59.69	59.59	0.16	9.00	8.27	8.17
hrv	84.07	85.80	2.07	36.11	33.63	6.86
recovery	59.55	61.45	3.19	22.66	19.06	15.91
sleep_score	76.39	76.27	0.17	18.55	15.65	15.67
hours_in_bed	7.78	7.76	0.33	1.93	1.62	15.91
hours_of_sleep	6.89	6.87	0.26	1.65	1.38	16.23
sleep_need	8.91	8.92	0.09	1.12	0.97	13.21
sleep_efficiency	88.87	88.96	0.10	6.30	5.32	15.48
wake_periods	14.87	14.19	4.56	5.61	5.04	10.42
sleep_disturbances	11.63	11.70	0.60	5.61	5.03	10.29
latency_min	2.89	2.76	4.21	5.28	4.20	20.43
cycles	5.30	5.24	1.26	1.26	1.05	16.17
rem_sleep_hours	2.04	2.02	1.20	0.79	0.67	14.43
deep_sleep_hours	1.37	1.36	0.74	0.46	0.39	13.90
light_sleep_hours	3.47	3.49	0.48	1.21	1.04	14.35
awake_hours	0.90	0.89	0.70	0.62	0.54	12.82
sleep_debt_hours	0.98	0.98	0.64	0.73	0.62	15.83
sleep_consistency	62.59	60.69	3.04	14.97	13.45	10.13
respiratory_rate	16.46	16.51	0.36	1.58	1.63	3.14
total_cycle_sleep_time_hours	7.18	7.15	0.38	1.79	1.51	15.63
rem_percentage	26.37	26.12	0.96	8.50	7.25	15.53
deep_sleep_percentage	17.79	17.75	0.18	4.79	4.03	15.97
restorative_sleep_hours	3.41	3.42	0.26	0.96	0.92	1.77
restorative_sleep	44.14	44.34	0.45	11.07	9.19	17.00
RSI	0.38	0.37	1.68	0.08	0.08	1.93

D. Synthetic Data Generation & Training

To create the synthetic data, we employed the CTGAN (Conditional Tabular Generative Adversarial Network) model, which is a GAN-based approach designed precisely to handle the complexities of tabular data. The original dataset contained both continuous and categorical variables as well as an overwhelming number of missing values, problems that are inconvenient for typical generative models. CTGAN was particularly helpful in this sense, thanks to its ability to model imbalanced distributions while guaranteeing the variable interactions through conditional generation. Although the model was successfully utilized in generating synthetic samples through learning from the underlying structure of the sanitized data, the utility of the generated data is still under consideration since current model results have fallen short of our expectations. This approach, though, served to alleviate the difficulties presented by missing values and allowed for additional experimentation with machine learning methods, all while preserving data privacy.

TABLE III
Results of Imputed Dataset

Dataset using imputed techniques	Model	MAE	RMSE	R^2 Score
Multiple Imputation by Chained Equation (MICE)	Ridge	0.002780	0.000159	0.898314
	Lasso	0.019029	0.000919	0.411488
	Random Forest	0.005965	0.000325	0.644861
	LightGBM	0.006435	0.000319	0.650644
	XGBoost	0.007474	0.000379	0.585643
Expectation-Maximization (EM)	Ridge	0.003854	0.000289	0.684196
	Lasso	0.011201	0.000457	0.499777
	Random Forest	0.005965	0.000325	0.644861
	LightGBM	0.006435	0.000319	0.650644
	XGBoost	0.007474	0.000379	0.585643

E. Training Machine Learning Model

A few machine learning regression models are trained on the imputed data:

- Linear Regression: An initial step to find linear associations.
- Ridge & Lasso Regression: Regularized regression models to deal with multicollinearity.
- Random Forest Regressor: An ensemble model to identify non-linear trends.
- Gradient Boosting (XGBoost, LightGBM): Sophisticated boosting models with improved accuracy.

Each model is trained using a variety of imputation techniques, and hyperparameter tuning is conducted to achieve optimal performance.

F. Model Evaluation

The models are benchmarked using the following key performance indicators:

- Mean Absolute Error (MAE): Approximates absolute differences between actual and predicted values.
- Root Mean Square Error (RMSE): Measures model performance by penalizing large errors.
- R^2 Score: Indicates the extent to which the model explains variation in the data.

G. Performance Analysis & Best Model Selection

Upon comparison of all models, the imputation method + ML model with the lowest MAE and RMSE and highest R^2 score is chosen. The results are plotted to present insights into how sleep and recovery trends influence athlete readiness.

IV. Experiments

A. Experimental Setup

The experiments were conducted on Google Colab, utilizing its cloud-based computational resources without GPU acceleration. The proposed framework was implemented in Python, focusing on various imputation methods and machine learning models for predicting Readiness Score (RSImod).

B. Model Evaluation Report

We employ important evaluation metrics in forecasting RSI Readiness to appraise the performance and accuracy of our predictive models.

Model comparisons using R^2 accuracy show sharp differences in performance between various imputation methods. LightGBM is the top model with the highest R^2 value (0.803) on the imputed_MICE dataset and performs consistently on other datasets. XGBoost comes in second, with good generalization evidenced by an R^2 of 0.761 (MICE) and 0.586 (EM) and a strong recommendation for prediction tasks. Random Forest displays mediocre performance, with high-scoring performances but less consistency than XGBoost and LightGBM. Linear models such as Ridge and Linear Regression have high performance in certain instances, i.e., imputed_MICE ($R^2 \approx 0.899$), but perform poorly with other imputation methods, revealing sensitivity to missing data management. Conversely, Lasso Regression is the worst, with the lowest R^2 values for all datasets, indicating that its feature selection mechanism is not ideal for this situation. Of imputation methods, MICE is the best, resulting in the highest R^2 values for all models, while EM is moderately successful and interpolation is the worst, with little predictive accuracy. Finally, LightGBM and XGBoost are the two top models, Random Forest and Ridge are middle-of-the-pack options, and Lasso with interpolation-based imputation is the worst method as in Table III.

C. Limitation and Challenges Faced

Many challenges were faced with data preparation and model development that had to do with lacking values, imbalanced data, and model accuracy across the various datasets.

- Missing RSI Values: Many readiness scores (RSImod) were missing, and this made it hard to train initially.
- Imputation Affected Results: Accuracy of the models varied based on what was used to impute the data.
- Unbalanced Data: Certain features had extremely small numbers of values, which hindered learning patterns.

- Synthetic Data Caution: Data produced by CTGAN needed to be carefully inspected in order to avoid unrealistic values.
- Inconsistent Model Performance: There were some models that did not perform well over various datasets.

V. Conclusion

Our research examined sleep and recovery pattern correlations and how they affect readiness for sport (RSImod). Utilizing several predictive algorithms, such as Linear Regression, Ridge/Lasso Regression, Random Forest, and XGBoost/LightGBM, we compared their success using the R^2 score as our main benchmark.

Results reveal that XGBoost and LightGBM achieved the best predictability, and they were capable of identifying complicated relationships between recovery, sleep, and readiness. Random Forest performed averagely, with a good balance of accuracy and interpretability, whereas Linear Regression and Ridge/Lasso Regression failed to perform, failing to capture non-linear relationships in the data. Explainable AI (xAI) methods emphasized the prominent effect of sleep duration and recovery scores on RSImod predictions. These results can be used by athletes and coaches to optimize training schedules for peak performance and less fatigue. Next studies may investigate the use of personal model tuning or clustering-based solutions for individual readiness prediction.

VI. Future Work

The present work is limited by the small size of the available data, which reflects on the quality and generality of the model built. Addressing this limit, the upcoming work will proceed with enriching the dataset with data generation. The growth in dataset size shall result in offering a richer estimation of the hidden data distribution and hence improved performance.

Once the dataset becomes big enough, deep learning models will be investigated in order to take advantage of the higher volume of data. Different architectures will be examined to find the most efficient methodology for the problem domain at hand. Moreover, thorough model testing and tuning will be done to evaluate how the larger dataset affects model robustness and performance. Comparative studies will be conducted among models developed based on the original and supplemented datasets in order to measure improvement and to test the efficacy of the suggested process of data generation.

This future direction is intended to produce a more viable dataset and utilize sophisticated deep learning methods, in order to enhance the reliability and performance of the proposed approach.

References

[1] C. D. Mah, K. E. Mah, E. J. Kezirian, and W. C. Dement, "The effects of sleep extension on the athletic performance of collegiate basketball players," *Sleep*, vol. 34, no. 7, pp. 943–950, 2011.

[2] T. M. Burke, P. J. Lisman, K. Maguire, L. Skeiky, J. J. Choynowski, V. F. Capaldi, J. N. Wilder, A. J. Brager, D. A. Dobrosielski *et al.*, "Examination of sleep and injury among college football athletes," *The Journal of Strength & Conditioning Research*, vol. 34, no. 3, pp. 609–616, 2020.

[3] C. Sargent, M. Lastella, S. L. Halson, and G. D. Roach, "How much sleep does an elite athlete need?" *International journal of sports physiology and performance*, vol. 16, no. 12, pp. 1746–1757, 2021.

[4] S. Senbel, S. Sharma, M. S. Raval, C. Taber, J. Nolan, N. S. Artan, D. Ezzeddine, and T. Kaya, "Impact of sleep and training on game performance and injury in division-1 women's basketball amidst the pandemic," *IEEE Access*, vol. 10, pp. 15 516–15 527, 2022.