SN Computer Science (2026) 7:288
https://doi.org/10.1007/s42979-026-04882-0

ORIGINAL RESEARCH

Data-Driven Imputation for Cohort Studies Using Collegiate Basketball Data

Srishti Sharma[1] · Hetav Raval[2] · Vishal Barot[3] · Srikrishnan Divakaran[1] · Tolga Kaya[4] · Christopher Taber[5] · Mehul S. Raval[1]

Received: 27 December 2025 / Accepted: 6 March 2026

Abstract

Missing data remains a critical challenge in cohort studies. This study introduces a novel missing-value imputation technique that integrates feature sensitivity analysis, factor analysis, clustering, and predictive modelling to enhance accuracy, reliability, and interpretability. The dataset comprises 42 features collected from 16 collegiate female basketball athletes over 26 weeks, including sleep and cardiac rhythms, training loads, cognitive states, travel, and countermovement jump performance. The objective is to model the impact of these contextual stressors on athletic readiness, quantified via the Reactive Strength Index modified (RSImod). Our proposed methodology achieved up to an 80.85% reduction in MSE and a 79.99% increase in R^2 scores in RSImod predictions on the imputed dataset, demonstrating substantial improvements over existing state-of-the-art approaches. The average reduction in computation time across the evaluated state-of-the-art methods was approximately 21.6%. External validation on an independent wearable-based sleep–HRV dataset (49 participants) confirmed generalisability, with the proposed model achieving an RMSE of 0.643 and outperforming baseline methods by 10–16%. Further, ablation analysis showed clear contributions from each module: clustering, hybrid feature weighting, factor analysis, and the pure XGBoost variant. Missing-data simulations confirmed robustness, with RMSE increasing from 0.74–0.95 (MCAR) to 0.80–1.06 (MAR) and 0.91–1.32 (MNAR), reflecting graceful degradation under structured missingness. Interpretability was enhanced using SHAP (SHapley Additive exPlanations) and ALE (Accumulated Local Effects) analyses, providing actionable insights for coaches and practitioners.

Keywords Athletic readiness · Clustering · Data imputation · Explainable AI · Factor analysis · Feature sensitivity · Prediction

✉ Mehul S. Raval
mehul.raval@ahduni.edu.in

Srishti Sharma
srishti.s1@ahduni.edu.in

Hetav Raval
raval.het@northeastern.edu

Vishal Barot
Vishal.barot@gtu.edu.in

Srikrishnan Divakaran
srikrishnan.divakaran@ahduni.edu.in

Tolga Kaya
kayat@sacredheart.edu

Christopher Taber
taberc@sacredheart.edu

[1] School of Engineering and Applied Science, Ahmedabad University, Navrangpura, Ahmedabad 380009, Gujarat, India

[2] Khoury College of Computer Sciences, Northeastern University, Fore Street, Portland 04101, Maine, United States of America

[3] Computer Engineering, Gujarat Technological University-Institute of Technology and Research (ITR), Mevad, Mehsana 384460, Gujarat, India

[4] School of Computer Science and Engineering, Sacred Heart University, Park Avenue, Fairfield 06825, Connecticut, United States of America

[5] Exercise Science, Sacred Heart University, Park Avenue, Fairfield 06825, Connecticut, United States of America

Published online: 18 March 2026

Introduction

Cohort studies follow a group of individuals, typically sharing a common characteristic or exposure, and observe them longitudinally to assess changes and outcomes [1]. In recent years, they have become increasingly prevalent in scientific research, offering a robust framework to investigate outcomes, interventions, and various phenomena over time. They span diverse domains, including public health, psychology, sociology, economics, education, environmental science, and, increasingly, sports science [2]. In sports science, they help understand how the applied stresses of training and competition alter an athlete's health and readiness, influence training protocols, impact injury risk, and enhance overall sport performance [3]. However, sports datasets often contain missing values due to athlete noncompliance, data collection errors, and equipment malfunctions. This leads to incomplete datasets and compromises the reliability and interpretability of findings [4].

There are three types of missing data mechanisms: Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR) [5]. MCAR implies that missing data are independent of the observed data, while MAR indicates that missing data correlate with the observed data but not with the missing values [6]. In contrast, MNAR suggests that missing data depends on both observed and missing data, making unbiased estimation challenging. Data collected in cohort studies establish associations among variables based on participant characteristics, generating both MAR and MNAR data types [6]. Imputing MAR allows for a more reliable estimation of missing values without introducing additional uncertainty. Therefore, for this study, all data in the database are categorised as MAR data [7].

In current practice, cohort studies increasingly leverage machine learning (ML) to uncover complex relationships and make predictive inferences from large-scale athlete monitoring data. These datasets often include longitudinal physiological, psychological, and performance metrics. However, such real-world data is prone to missing entries due to device failures, inconsistent reporting, or logistical constraints during data collection. Since ML models are highly sensitive to data quality, effective strategies for handling missing data are essential to ensure robust, unbiased, and interpretable outcomes [4]. These include removing instances or features with missing values (resulting in data loss), allowing the ML algorithm to handle missing values during its execution, or replacing them with estimated ones through missing value imputation (MVI), closely approximating the real values [4].

Recent research on data imputation in sports science has addressed critical challenges in managing missing data to ensure accurate workload and performance analysis. One study aimed to identify optimal imputation methods for high school athletes' workload and found that machine learning models combining session and individual context significantly improved accuracy [8]. Another study utilised Multivariate Imputation by Chained Equations (MICE) to impute data for basketball players, demonstrating high predictive accuracy for game performance and injury and emphasising the importance of addressing data gaps to enhance decision-making [4]. Yet another sports science study focused on professional soccer players' training load, revealing that the Daily Team Mean method effectively mitigated inaccuracies caused by missing data [9].

The literature on imputation methods for handling missing data highlights several prominent techniques widely used across statistical and machine learning domains. Simple imputation replaces missing values with the mean or mode of the available data, but may yield poor results in complex datasets [7]. Regression imputation predicts missing values using regression equations fitted on complete data [7]. Expectation-maximization (EM) iteratively estimates missing values based on existing data distributions [10]. Multiple imputations (MICE) generate multiple datasets to reflect uncertainty in missing values, aggregating results for final imputations [4, 11]. K-nearest neighbour (KNN) imputation estimates missing values based on similarities with neighbouring samples [12]. Clustering imputation groups data into clusters to impute missing values within similar groups [10]. Decision tree (CART) and random forest (RF) methods build predictive models from complete data and use these to impute missing values [13]. These traditional imputation techniques have limitations, such as assuming linear relationships (e.g., Regression), sensitivity to initialization (e.g., EM), and reliance on similarity metrics (e.g., KNN). In the context of real-world sports science data–such as ours, which exhibits non-linear patterns, individual variability, and temporal dependencies–these assumptions often fail to hold. For instance, MICE relies on accurate distributional assumptions, which are hard to validate in highly dynamic, athlete-specific datasets. MICE requires accurate distribution assumptions. Clustering imputation may introduce bias when clusters are poorly defined. CART is prone to overfitting, while RF can bias results if feature selection is inadequate [1].

Recent advancements in deep learning have significantly improved data imputation in medical and health domains. Traditional methods such as mean, median, and KNN often fail to capture complex relationships among variables [1]. Overcomplete Denoising Autoencoders (ODAE) and similar deep learning models, which treat missing data as noise and use denoising regularisation, have shown superior performance by effectively learning hidden data representations

[14]. These models achieve lower mean squared error and higher prediction accuracy, proving more reliable for imputation tasks than conventional methods [15]. However, deep learning methods are less useful in cohort studies due to their need for large training data, inherent complexity, and lack of interpretability, critical for understanding subtle nuances in longitudinal data and deriving actionable insights for personalised interventions [1, 15]. Given the relatively small size of many cohort-study datasets and the need for interpretable modelling in applied sports science, this study did not evaluate deep learning approaches experimentally.

This research focuses on a collegiate women's basketball team, aiming to provide actionable advice based on the model's outcomes. Therefore, quantifying and explaining features and predicted outcomes is crucial. We propose a data-driven MVI method that adaptively learns from the dataset. The technique first identifies the most influential features using a hybrid feature importance analysis, followed by factor analysis, to generate latent factors that compactly represent the original dataset. These steps reduce data dimensionality, achieve a quadratic speed-up, and improve the interpretability of outcomes. It then leverages clustering and a hybrid decision tree (DT)- based ensemble boosting model to improve the accuracy and reliability of imputed values.

The method was first validated on our collegiate dataset by assessing its impact on athlete readiness modelling, where improved imputations yielded lower MSE and higher R^2 values for RSImod predictions. A systematic module-wise ablation was performed to quantify the contribution of each component, confirming the incremental value of factor reduction, hybrid feature weighting, and athlete-specific clustering. Assuming data missingness under the MAR mechanism, we assessed the impact of increasing missingness rates (5%, 10%, 15%, and 20%) on each component of the proposed method. The approach was then tested on an external real-world Wearable-HRV sleep diary dataset [16] to examine cross-cohort generalizability. Next, we benchmarked empirical runtime across increasing cohort sizes to assess computational efficiency, ensuring that the proposed approach remains scalable for practical deployment in applied sports-science environments. Finally, SHAP (SHapley Additive exPlanations) and ALE (Accumulated Local Effects) were plotted to enhance interpretability by quantifying feature contributions and visualizing their marginal effects on the model outputs under varying missingness conditions.

Research Contributions:

1. We designed a feature weighting framework that combines Pearson correlation, Random Forest, and XGBoost importance scores, using Softmax weighting to identify the most influential predictors for imputation.
2. We applied factor analysis to reduce dimensionality by extracting interpretable latent physiological factors, followed by k-means clustering to capture inter-individual variability among athletes.
3. Missing values are reconstructed using cluster-specific XGBoost regressors, with predictions aggregated through similarity-based Softmax weighting.
4. We integrated SHAP and ALE for interpretability on the collegiate basketball cohort dataset.

The rest of the paper is organised as follows: Section II describes the dataset. Section III outlines the methodology for implementing the proposed MVI technique. Section IV summarises the results and discusses the interpretations drawn, and Section V is the conclusion.

Dataset

This real-world dataset includes sleep-recovery patterns, subjective training load, cognitive state information, countermovement vertical jumps, and travel data of collegiate women's basketball athletes. This data was collected to assess the fatigue caused by various internal and external physical, physiological, and cognitive stressors and their impact on athletic readiness.

Subjects

From October 2021 to March 2022, sixteen female basketball players (mean age: 21 years; average height: 174.21 cm; mean body mass: 73.98 kg) from Sacred Heart University, CT, USA, underwent comprehensive testing and monitoring. Before participation, all participants received detailed explanations of the study procedures and provided informed consent (Institutional review board approval number 170720a).

Sleep and Recovery Data

All athletes were given Whoop straps (WHOOP, Boston, MA, USA) to wear continuously throughout the data collection period, except during games and practice. The straps monitored daily activity and sleep, recording data using Whoop's specialised software. The study analysed 22 metrics per athlete daily, including resting heart rate, heart rate variability, sleep parameters, and recovery metrics. Third-party testing has validated Whoop's reliability and accuracy compared to polysomnography for sleep and heart rate assessments [17].

Training Data

The training load was quantified weekly by aggregating the workload from sports practice, metabolic conditioning, strength training, and gameplay. After each session, athletes reported their perceived exertion using a 1–10 Likert scale, which was then multiplied by the session duration to compute the session rating of perceived exertion (sRPE) [17]. Total Weekly Load (TWLoad) and its standard deviation were derived from these sRPE values across the week. Additionally, the weekly resistance training load was computed by summing the total weight lifted during resistance sessions $(sets \times repetitions \times load)$. Training monotony was calculated as the average daily load normalised by the weekly standard deviation of the training load. In contrast, training strain was determined by multiplying TWLoad by the monotony score [17].

Short Recovery Short Stress Questionnaire

Twice weekly, athletes utilised an online dashboard to complete a brief questionnaire assessing their emotional and mental states. The survey included eight questions; four focused on recovery and the remainder on stress, each rated on a 0-6 Likert scale [18]. Table 1 lists the internal and external load quantifying features.

Vertical Jump Data

Subjects performed weekly countermovement jumps (CMJs) on the first practice day each week, typically on Monday or Tuesday. After a standardised general warm-up, they executed two submaximal jumps at 50% and 75% of perceived maximum effort, with 30 s of passive rest between repetitions. Dual force plate (Vald Force Decks, Brisbane, AUS) sampling at 1000 Hz recorded all jumps; data were collected and analysed using proprietary Force Decks software. The RSImod, jump height via flight time, and peak power were the key performance indicators (KPIs) monitored during routine athlete testing and monitoring [17]. RSImod, derived from CMJ data, measures an athlete's ability to generate maximal vertical impulse quickly, integrating jump height and contact time.

Travel Data

During the season, athletes travel to different states for games. We gathered travel dates from their game-schedule journal, marking them as 0 or 1, and recorded the travel hours.

Wearable-HRV Sleep Diary Dataset to Test Generalisation

We used "A continuous real-world dataset comprising wearable-based heart rate variability alongside sleep diaries" [16] for validation of the imputation efficacy of our proposed MVI scheme. It contains data from 49 adult participants (mean age $\approx$ 28.4 years; 51% female), each monitored continuously for four weeks, resulting in over 1,300 participant-days of physiological and behavioural records. The dataset integrates wearable sensor data with self-reported sleep diaries, providing a rich set of features. Wearable-derived measurements include heart rate (HR), heart rate variability–specifically, the root mean square of successive differences (RMSSD) and the standard deviation of normal-to-normal intervals (SDNN) – extracted from photoplethysmography (PPG) signals; resting heart rate; movement intensity; and time-stamped activity information. The sleep diary component provides nightly sleep duration, sleep efficiency, sleep onset and wake times, and subjective evaluations of sleep quality, including proportions of deep, light, and REM sleep.

Methodology

We designate the feature requiring imputation as the target feature and determine the importance scores of other features using Pearson's correlation, Random Forest-based

Table 1 Internal and external load quantifying features

Source	Features
Sleep and Recovery (22)	hours of sleep, hours in bed, awake hours, wake periods, sleep disturbances, resting heart rate (RHR), heart rate variability (HRV), respiratory rate, recovery, deep sleep hours, light sleep hours, restorative sleep hours, rapid eye movement (REM) sleep hours, total cycle sleep hours, sleep consistency, sleep efficiency, sleep score, sleep debt hours, total cycle nap hours, latency, sleep cycles, sleep need
Training (6)	strain, monotony, resistance training (RT) volume load, total workload (TWLoad), daily average, weekly standard deviation (SD)
Short recovery short stress (SRSS) questionnaire (8)	overall recovery (OR), overall stress (OS), negative emotional state (NES), lack of activation (LA), muscular strength (MS), physical performance capabilities (PPC), mental performance capabilities (MPC), emotional balance (EB)
CMJs (4)	peak power, jump height (JH), body weight, reactive strength index modified (RSImod)
Travel Schedule (2)	Travel (yes/no), Hours of travel

feature importance, and XGBoost-based feature importance for the target feature (*Section 3.1*). Following this, factor analysis addresses multicollinearity and reduces dimensionality, improving data interpretation and processing efficiency (*Section 3.2*). Athlete clustering via k-means further refines the approach by grouping similar athletes (*Section 3.3*). Finally, our model employs XGBoost regressors trained on these clusters, using similarity scores and softmax probabilities to accurately predict missing values (*Section 3.4*). Figure 1 shows the proposed methodology, detailing the techniques step-wise.

Feature Importance Analysis

The feature requiring imputation is designated as the target feature. We determine the importance scores of other features in the dataset relative to the target feature using three distinct techniques: Pearson's correlation [19], Random Forest-based feature importance, and XGBoost-based feature importance. Pearson's correlation captures linear relationships, while ensemble methods such as Random Forest and XGBoost handle complex interactions and nonlinearity [4]. We proposed a hybrid approach for computing an aggregate feature importance (FI) score to provide a comprehensive and nuanced feature selection process, particularly for complex datasets where individual techniques may have limitations [1].

Figure 2 presents the feature importance analysis. For each FI technique, we identify the top M ($M << N$) features based on their significance value (p-values), ensuring only relevant features are selected for model training. Subsequently, we train three models using the selected feature subsets to predict RSImod scores. To evaluate the effectiveness of each method in generating an optimal feature subset for RSImod score prediction, we assess all three models using Mean Squared Error (MSE), defined as the average squared difference between the observed and predicted RSImod scores [20]. The MSE values obtained from these models are then processed using a Softmax function to generate probabilistic scores for each technique, reflecting their respective weights in the overall analysis [21]. By normalising the FI scores for all three techniques to the range 0 to 1, we computed a weighted-average feature importance score for each feature, with the probabilistic score from each technique serving as the weight. We identify the top K features using this aggregated score, with K significantly smaller than N.

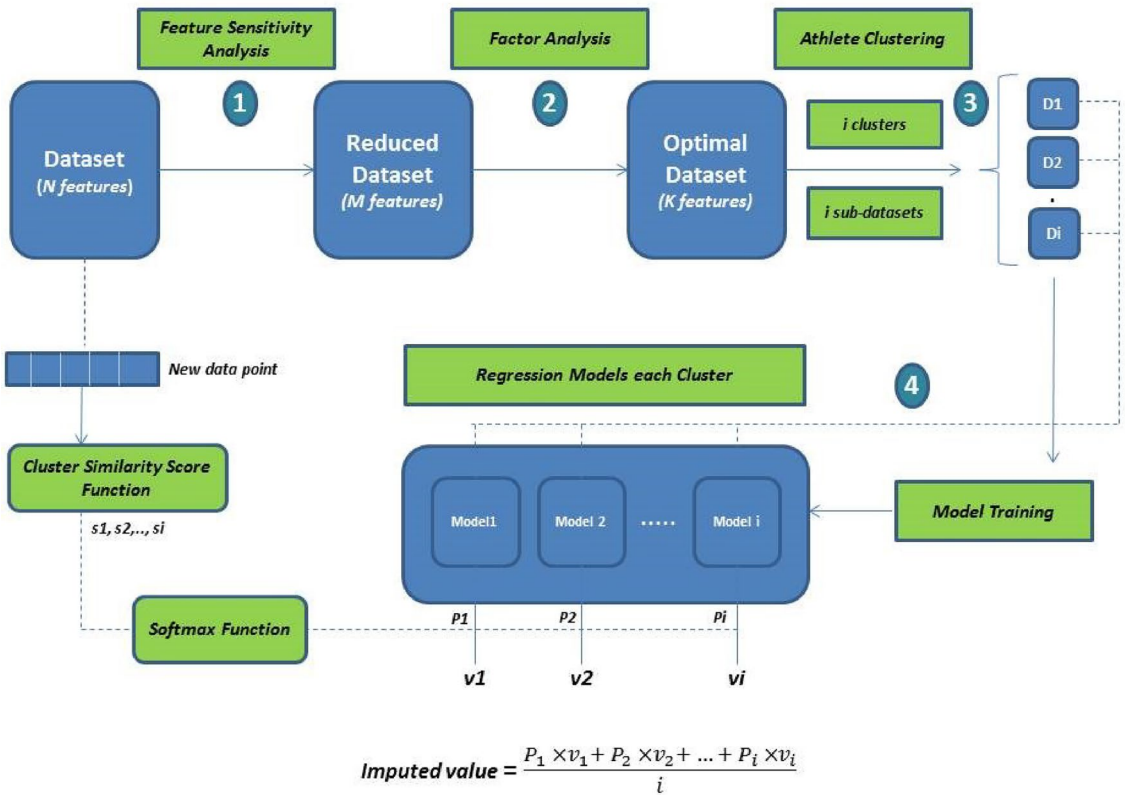

Fig. 1 Proposed methodology depicts how a new data point is imputed

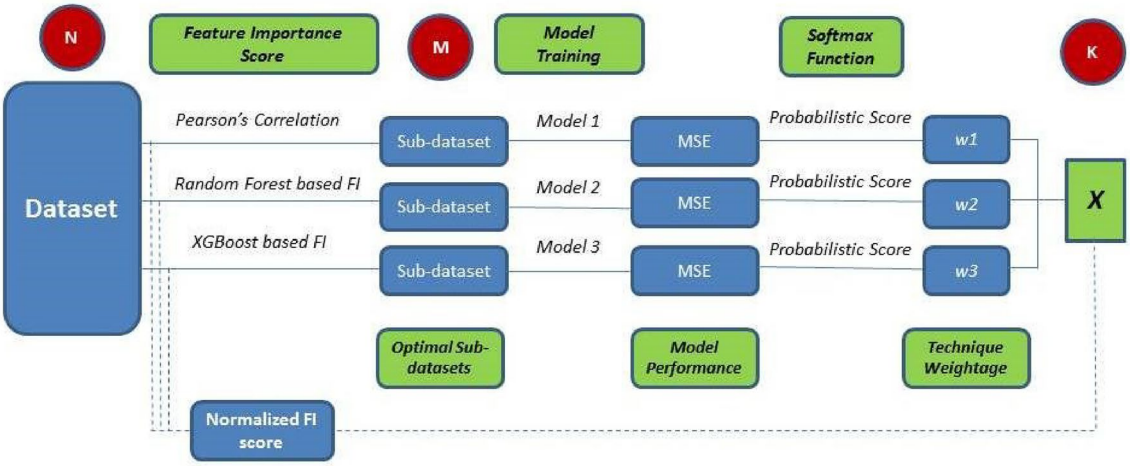

Fig. 2 Hybrid feature importance analysis technique

Factor Analysis

We observed several instances of multicollinearity and linear dependencies in the dataset. To preserve the significance of these features without discarding any of them, we combined them into latent features via factor analysis [20]. This approach helped reduce dimensionality and simplify data interpretation. We conducted a factor analysis on the optimal subset of features to identify the underlying latent factors that compactly represent the dataset. Each factor (F_i) is expressed as a linear combination of features (f_x) sharing common variance, with the resulting factor loadings (l_x) serving as weights [20].

$$F_i = \sum_{x=1}^{n} l_x \times f_x \tag{1}$$

Athlete Clustering

Athlete clustering is a statistical method to identify patterns and similarities by grouping athletes based on shared traits or behaviours within a dataset [22]. It allows us to leverage information from similar athletes to estimate the missing values more accurately [22].

In cohort studies involving human subjects, physiological and behavioural responses often vary substantially across individuals due to differences in training history, recovery capacity, sleep patterns, and adaptation to workload. A single global regression model may therefore fail to capture these heterogeneous patterns. Clustering athletes prior to imputation enables the model to learn subgroup-specific relationships between predictors and outcomes, allowing missing values to be reconstructed using data from physiologically similar individuals rather than the entire population. We performed athlete clustering using k-means on the identified latent factors. The Silhouette score was used determine the optimal number of clusters [23].

Proposed Hybrid Model for Imputation

The dataset was systematically partitioned into four distinct sub-datasets, each representing an athlete cluster by allocating each athlete's data records to their respective cluster. XGBoost regressor, known for its speed, regularisation capabilities, ability to handle missing data, and feature interpretability, was utilised in this process [24]. Subsequently, four XGBoost regressors were trained, each on one of these sub-datasets, to predict the missing feature values.

Algorithm: Missing Value Imputation

Step 1: Cluster assignment and similarity score calculation For a new record x, the similarity score s_i is calculated as the negative Euclidean distance between the record and the centroid of the cluster μ_i:

$$s_i = -\|x - \mu_i\| \tag{2}$$

where $\|x - \mu_i\|$ is the Euclidean distance between the record x and the centroid μ_i of cluster i.

Step 2: Softmax probabilities The similarity scores are converted into probabilistic weights using the softmax function:

$$P_i = \frac{e^{s_i}}{\sum_{j=1}^{k} e^{s_j}} \tag{3}$$

where e is the base of the natural logarithm, s_i is the similarity score for cluster i, and k is the number of clusters.

Step 3: Missing value imputation The final feature value $\hat{y}$ is computed as a weighted average of the predictions from the XGBoost regressors, with the softmax probabilities as weights:

$$\hat{y} = \sum_{i=1}^{k} P_i \times \hat{y}_i \tag{4}$$

where $\hat{y}_i$ is the feature value predicted by the XGBoost regressor trained on cluster i, and P_i is the softmax probability for cluster i.

To impute missing feature values, a new record was first classified using an unsupervised machine learning model trained with the k-means clustering technique to determine the athlete's cluster affiliation. The similarity score for this record with each cluster was calculated as the distance from its centroid. These similarity scores were passed through a softmax function, yielding probabilistic weights for each cluster. The final feature value was imputed as a weighted average of the four XGBoost regressors' predictions, with softmax probabilities serving as weights corresponding to each regressor's cluster.

Experimental Evaluations, Results and Discussions

Feature Importance

The dataset has 42 features in total, with an overall missing-data rate of 32%. As this dataset's objective is modelling and predicting athletic readiness, RSImod serves as the KPI, which is the target feature. We first computed Pearson's correlation coefficients for all other RSImod features to identify which features contributed most to the RSImod score's variability. Apart from the RSImod score, the features identified as the most significant ones -hours of sleep (0.72), TWLoad (0.68), HRV (0.60), and sleep need (0.54) were selected for imputation. The methodology and results of the preprocessing steps have been explained in terms of RSImod score imputation.

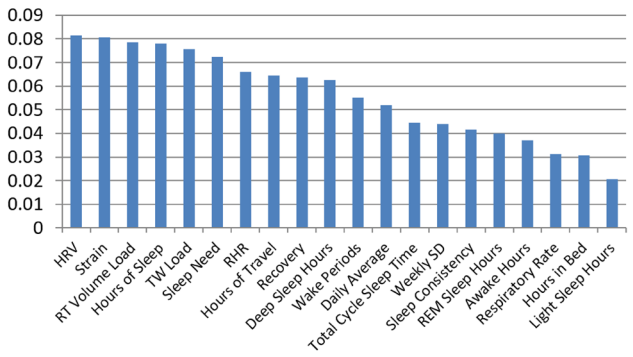

Fig. 3 Top 20 features constituting the optimal dataset

To impute the RSImod score for the week's athletes who missed their resistance and strength training sessions, we set the RSImod score as the target feature. The dataset included daily recordings for each of the 41 ($N = 41$) independent features, while the weekly CMJs provided a single RSImod score per week. To facilitate imputation, we first averaged daily readings for each feature into a single weekly reading. Over 26 weeks, 416 RSImod scores were recorded ($n = 416$, $mean = 0.37 \pm 0.08$) with 98 missing entries (used as a test set).

Using the three distinct techniques, we first calculated the FI score of the 41 independent features. We employed forward selection and backward elimination to identify the features that most significantly impacted RSImod prediction, with the threshold for the number of features set by the p-value. On average, the top 15 features with $p < 0.05$ improved model performance and were included in the reduced data subsets. We calculated the models' MSEs, and the FI scores for all techniques were normalised to the range 0-1. The MSE values from these models were then processed by a Softmax function, yielding a probabilistic score for each technique reflecting its weight in the overall analysis. We computed each feature's average importance score using these weights. Using this aggregated score and the p-values from model fits, we identified the top 20 features (see Fig. 3) as the optimal subset for further analysis.

Factor Analysis

Our dataset now consisted of 21 features (20 input features and RSImod). To assess the suitability of the dataset for factor analysis, we conducted the Kaiser-Meyer-Olkin (KMO) test, which measures sampling adequacy, and Bartlett's test of sphericity, which examines the redundancy among features that could be summarised into factors. The KMO test yielded a value of 0.69 (> 0.50), and Bartlett's test resulted in a significance value of 0.000 (< 0.005). These results indicated that the dataset was suitable for factor analysis. Consequently, we performed factor analysis, which computed four latent factors. The features were then grouped into four discrete clusters (factors) based on their correlations with the identified factors, yielding a 5-feature dataset (4 factors and RSImod).

- The features "hours of sleep," "hours in bed," "total cycle sleep time," "awake hours," "wake periods," "deep sleep hours," "REM sleep hours," "light sleep hours," "sleep consistency," and "sleep need," collected using

the WHOOP strap, showed the highest correlation with *factor 0* based on their factor loadings. We computed the value for factor 0 as a linear combination of these features, using their factor loadings as weights, and named this factor "Sleep."

- The features "HRV," "RHR," "recovery," and "respiratory rate," also collected from the WHOOP strap, showed the highest correlation with *factor 1*. We calculated the value for factor 1 for each record by using the values of these features and their factor loadings as weights, naming this factor "Cardiac Rhythm."
- Similarly, features "TWLoad," "strain," "RT volume load," "weekly SD," and "daily average," collected during strength and resistance training sessions, constituted *factor 2*, which we named "Training Strain." The "hours of travel" feature formed *factor 3*, named "Travel Schedule."

In Table 2, we compare the MSE and adjusted R^2 scores of our approach using latent factors with those of multilinear regression, incorporating all input features from the dataset, to assess the efficacy of factor analysis in reducing redundancy without significant information loss.

Notice that the optimal dataset after factor analysis improved the models prediction accuracy (slightly smaller MSE) and reduced variability (slightly increased adjusted R^2 score). It indicates that factor analysis effectively captured the essential information in the data by eliminating redundant features, thereby reducing data dimensionality.

Table 2 MSE AND R^2 scores with features and factors

Dataset	MSE	R^2
Original (41 features)	0.048	0.560
Optimal (4 factors)	0.028	0.694

Cluster Analysis and Module-Wise Ablative Study

Among the features categorised under "Sleep," "hours of sleep" was identified as the most significant, having the highest factor loading. For "Training Strain," the most important feature was "strain"; for "Cardiac Rhythm," it was "HRV"; and for "Travel Schedule," it was "hours of travel". Following the silhouette score recommendation, we utilised the k-means clustering algorithm to partition the dataset into four distinct athlete clusters. Table 3 summarises the average cluster statistics.

Overall, we observed an imbalance in the dataset: 38% of the records had high or very high RSImod scores, while 62% had low or moderate RSImod scores. To address this imbalance, we applied the SMOGN [4] technique for data balancing, increasing the overall sample size to 712. This adjustment ensured that 50% of the records belonged to high or very high levels of athlete readiness, and the remaining 50% to low or moderate levels. We partitioned the dataset into four subsets by grouping the athlete data points into each cluster.

To validate model efficacy, we introduce missing data into the dataset by deliberately removing 213 RSImod score values, resulting in 499 records allocated for training and those 213 for testing using a traditional 70:30 training-to-test ratio. Over each of these subsets, we trained an XGBoost regressor, with RSImod as the target feature. We first identified the athlete's cluster affiliation using an unsupervised machine learning model for each record in the test dataset. Next, we computed the distance from the record to the centroid of each cluster to obtain a similarity score. This similarity score was passed through the Softmax function to generate probabilistic weights for each cluster. The final RSImod score was imputed as a weighted average of the XGBoost regressor predictions, with the Softmax probabilities serving as weights corresponding to their respective clusters. The hybrid model achieved an MSE of 0.0102, outperforming the individual cluster-based models with MSEs of 0.0265, 0.0287, 0.0271, and 0.0197 for Clusters 1, 2, 3, and 4, respectively.

Now, to quantify the contribution of each component of the proposed imputation pipeline, we conducted a systematic module-wise ablation study comparing feature-importance weighting, factor analysis, clustering, and the full hybrid model. We report results on two complementary evaluation sets to capture both realistic performance and robustness. First, we use 213 test records drawn from the SMOGN-balanced cohort (712 total records) to assess model accuracy under a single, practically relevant missingness scenario that preserves the full distribution of athlete readiness levels. This larger test set provides stable estimates of overall MSE and R^2 for the proposed hybrid imputation pipeline. In addition, we analyse a focused subset of 78 RSIMod records from raw data on which we systematically impose increasing levels of MAR missingness (0–15%). This smaller but controlled subset allows us to isolate the effect of missingness severity and to perform module-wise ablation, without

Table 3 Average cluster statistics

Cluster	Athletes	Average sleep hours	Training	RSImod
Moderate performers (Cluster 1)	4, 6, 13, 16, 17	6.78	2060	0.32
Intensive trainers (Cluster 2)	2, 8, 10, 12, 16	6.59	2268.5	0.24
High performers (Cluster 3)	3, 7, 9, 11	7.28	1935	0.39
Suboptimal performers (Cluster 4)	1, 5, 15	7.28	1890	0.27

Training is measured in arbitrary units and rsimod as jump height/time to takeoff

Table 4 Module-wise ablation results (MSE and R^2 summary) on RSIMod testset with 213 records

Configuration	MSE	R^2
Pure XGBoost	0.051	0.538
Hybrid FI removed	0.046	0.574
Factor analysis removed	0.053	0.512
Clustering removed	0.042	0.618
Full model (FI + FA + Cluster + Hybrid XGB)	0.036	0.667

the confounding influence of repeated resampling of the entire cohort.

Table 4 presents module-wise ablation results on 213 records using MSE and R^2, with the pure XGBoost model serving as the baseline. Removing factor analysis produced the largest decline in performance (MSE = 0.053, R^2 = 0.512), indicating that latent factor extraction plays the most critical role in reducing redundancy and preserving predictive structure. Excluding hybrid feature-importance weighting also degraded performance relative to the baseline (MSE = 0.046, R^2 = 0.574), showing that sensitivity-guided feature weighting improves the alignment between predictors and the target variable. Removing clustering resulted in moderate performance loss (MSE = 0.042, R^2 = 0.618), confirming that athlete-specific grouping enhances reconstruction accuracy. The full hybrid model, which integrates feature weighting, factor analysis, clustering, and XGBoost prediction, achieved the best overall performance (MSE = 0.036, R^2 = 0.667), demonstrating that each module contributes incrementally and that the modules collectively yield a more accurate and stable imputation framework.

Figure 4 shows module-wise ablation results for increasing MAR missingness levels (0%, 5%, 10%, and 15%) across 78 records from the raw dataset, extending the analysis from Table 4. Removing hybrid feature-importance weighting resulted in the largest increase in MSE (0.00693–0.00912) and the greatest drop in R^2 (0.622–0.493), underscoring its critical role in aligning predictors with targets. Excluding factor analysis moderately raised MSE (0.00605–0.00669) and reduced R^2 (0.664–0.630), highlighting the importance of latent factor extraction. Interestingly, removing clustering reduced the MSE (0.00429–0.00604) while maintaining high R^2 (0.764–0.670), suggesting that athlete-specific clustering can be counterproductive in the presence of missing data. The Full Model remained robust (MSE: 0.00644–0.00832; R^2: 0.590–0.545), confirming that hybrid weighting and factor analysis are essential, while the integrated pipeline balances robustness and predictive accuracy. Together, these factors mean that absolute error and R^2 values are not expected to match across the two experiments, even though the relative conclusion–that the full hybrid model outperforms its ablated variants–remains consistent.

Table 5 RMSE comparision with state-of-the-art methods

Technique/feature	Hours of sleep	Sleep need	TWLoad	OR	HRV
KNN [12]	0.78	1.28	0.98	1.1	0.88
MICE [11]	0.74	1.25	0.95	1.08	0.86
EM [10]	0.82	1.35	1.05	1.18	0.95
CART [13]	0.79	1.3	1.00	1.12	0.9
XGBoost [13]	0.81	1.32	1.02	1.15	0.92
Proposed	0.72	1.21	0.93	1.05	0.84

Comparison with State-of-the-Art

To further validate the efficacy of the imputations performed by the proposed approach, it was compared against five state-of-the-art missing value imputation (MVI) techniques: KNN [12], MICE [11], EM [10], CART [13], XGBoost [13]. The evaluation criteria used were Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). Table 5 and 6 present the RMSE and MAE values obtained for each technique across five different features, respectively.

The RMSE and MAE values for sleep features, training metrics, cognitive variables, and cardiac rhythm parameters

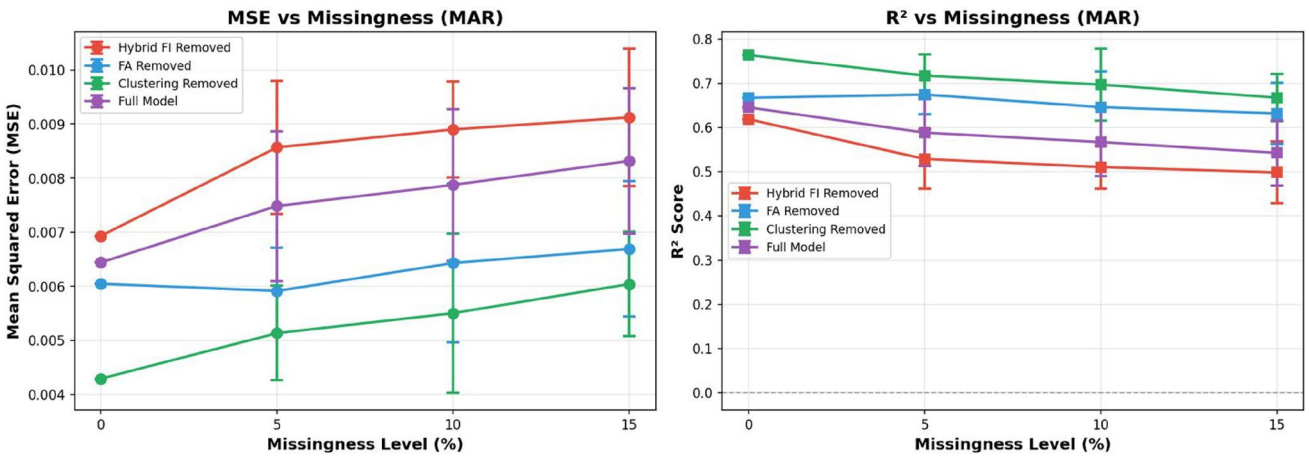

Fig. 4 Module-wise impact due to data missingness on RSIMod raw dataset with 78 records

Table 6 MAE comparison with state-of-the-art methods

Technique/feature	Hours of sleep	Sleep need	TWLoad	OR	HRV
KNN [12]	0.62	1.02	0.79	0.85	0.70
MICE [11]	0.60	1.00	0.77	0.84	0.68
EM [10]	0.65	1.08	0.84	0.92	0.75
CART [13]	0.63	1.04	0.80	0.87	0.71
XGBoost [13]	0.64	1.06	0.82	0.90	0.73
Proposed	0.58	0.97	0.76	0.82	0.67

Table 7 MSE AND R^2 comparison of RSImod score over imputed datasets

Imputation technique	MSE	R^2 score
KNN [12]	0.0534	0.487
MICE [11]	0.0312	0.680
EM [10]	0.0490	0.512
CART [13]	0.0361	0.734
XGBoost [13]	0.0283	0.760
Proposed approach	0.0102	0.897

consistently showed lower error rates with our approach compared to KNN [12], MICE [11], EM [10], CART [13], XGBoost [13]. Specifically, for hours of sleep, our technique achieved an RMSE of 0.72 and an MAE of 0.58, while the best-performing comparative technique, MICE, achieved an RMSE of 0.74 and an MAE of 0.60. Similar trends were observed across all features, demonstrating the robustness and efficacy of our proposed method in handling missing data.

Next, RSImod predictions were generated using XGBoost regressor models trained on datasets imputed by these MIV techniques (RSImod score imputation for 98 missing values). Table 7 compares model performances regarding MSE and R^2 score of our proposed data-driven MVI scheme and state-of-the-art imputation techniques.

The model fit over the dataset imputed using our proposed approach achieved the lowest MSE of 0.0102, substantially outperforming KNN (0.0534), MICE (0.0312), EM (0.0490), CART (0.0361), and XGBoost (0.0283). Moreover, it exhibited the highest R^2 score of 0.897, indicating superior model fit and predictive accuracy. The superior performance of our proposed technique can be attributed to several key factors. Firstly, our approach integrates a hybrid methodology that leverages feature importance analysis, factor analysis, and athlete clustering to identify and prioritise influential features for imputation. By accounting for both linear relationships and complex interactions using ensemble methods such as XGBoost, our technique effectively captures the nuanced patterns in the dataset, thereby improving the accuracy of imputed values. Furthermore, incorporating athlete clustering ensures that missing values are estimated based on athletes' similarities, enhancing the relevance and precision of the imputation process and tailoring it to individual performance profiles. This comprehensive approach addresses challenges with missing data and optimises predictive modelling outcomes for assessing athletic readiness and performance indicators.

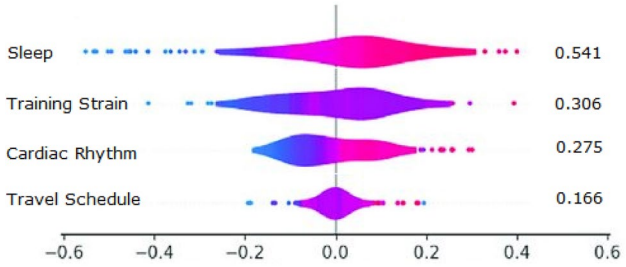

Fig. 5 SHAP value explanation of predicted RSImod score - features are listed by importance, with positive SHAP values indicating a positive impact on the prediction and negative values indicating a negative impact. The colour gradient reflects the magnitude of the impact, with darker colours representing stronger influences

eXplainable AI - SHAP and ALE Plots

Additionally, we generated SHAP (SHapley Additive exPlanations) values for the predicted RSImod values. These values provide a clear and interpretable rationale behind each imputed value [25]. As depicted in Fig. 5, SHAP values for a predicted RSImod score highlight sleep's paramount importance. Effectively managing training strain and optimising cardiac rhythm can enhance athlete readiness and performance. Conversely, the travel schedule exhibits the least significance.

By generating SHAP values for the predicted RSImod values, our approach not only imputes missing data but also quantifies each feature's contribution to the imputation outcome. This enables us to understand which features significantly influence the imputed RSImod score and how they interact during the imputation process.

Moreover, we leveraged Accumulated Local Effects (ALE), a model-agnostic technique that quantifies the isolated effect of individual features on model predictions. It addresses the issues of correlation in traditional Partial Dependence Plots (PDPs). ALE was selected given the correlation of features present in our athlete monitoring data (e.g., Sleep, Strain, and HRV). ALE estimates the true effect of each feature by calculating local differences within small intervals, rather than averaging across unrealistic combinations of features.

The ALE plots in Fig. 6 indicate that training strain has a predominant effect on RSI, with higher training loads leading to reduced reactive strength, particularly beyond moderate strain levels, suggesting accumulated fatigue. In contrast, HRV shows a clear positive association with RSI, with higher HRV values associated with improved

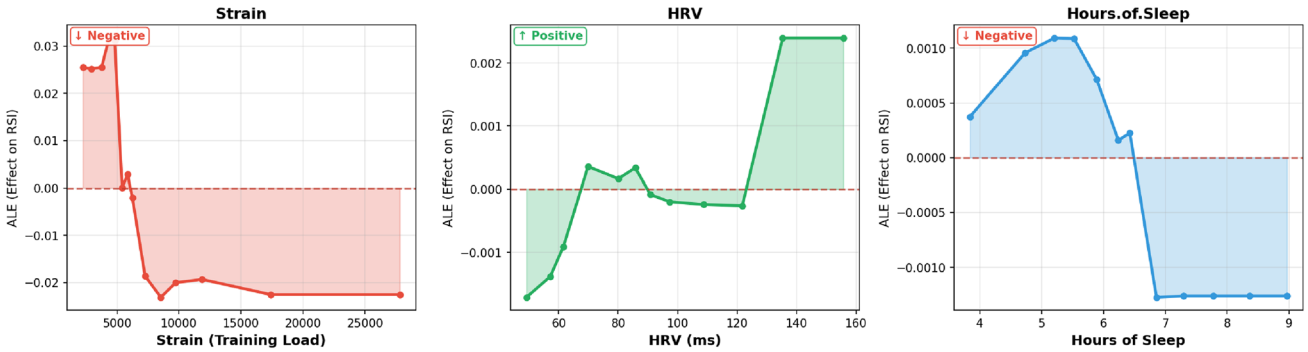

Fig. 6 ALE plots showing feature effects on RSI prediction. Strain negatively impacts RSI at higher values; HRV positively correlates with RSI; Sleep shows optimal effects at 5-6 h

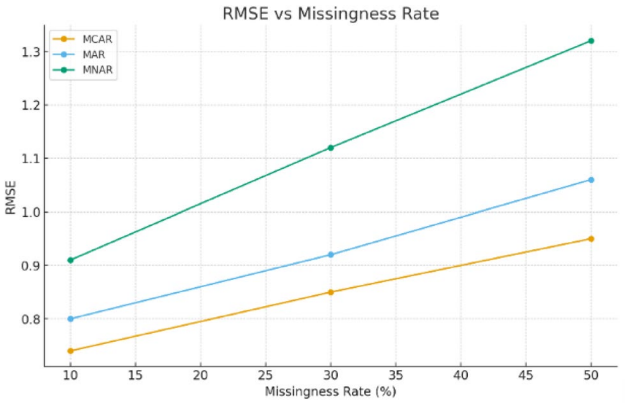

Fig. 7 RMSE achieved by the proposed imputation scheme across different sub-datasets

neuromuscular readiness and better explosive performance. Hours of sleep show a non-linear relationship with RSI: moderate sleep duration has a small positive effect, while insufficient or excessive sleep is associated with a decline in reactive strength, highlighting the importance of optimal sleep balance for performance. The ALE results match the SHAP interpretations: Sleep has the greatest influence in both analyses (SHAP importance: 0.541), followed by Training Strain (SHAP importance: 0.306), which corresponds to the strong negative ALE effect observed at high Strain values. HRV, captured within the Cardiac Rhythm factor (SHAP: 0.275), demonstrates the positive monotonic relationship seen in the ALE plot. Both methods showcase that sleep-related variables and training load are the primary contributors to RSImod prediction, validating the model's physiological interpretability.

Simulation of Missing Data Mechanisms (MCAR, MAR, MNAR)

Our dataset was predominantly MAR. However, to generalise our proposed imputation scheme to real-world uncertainties in cohort studies, we applied it to three synthetically generated sub-datasets. In these three sub-datasets, missingness was simulated under three mechanisms: MCAR, where values were removed uniformly at random; MAR, where missingness depended on observed variables such as high TWLoad leading to missing HRV; and MNAR, where the probability of missingness depended on the unobserved value itself, for example, lower RSImod being more likely to be missing. Missingness rates of 10%, 30%, and 50% were applied across the dataset. Figure 7 compares RMSE achieved by the proposed imputation scheme across different sub-datasets.

At a low missing rate of 10%, the model achieves the lowest error under MCAR (RMSE = 0.74), followed by MAR (RMSE = 0.80), while MNAR exhibits the highest error (RMSE = 0.91). As the proportion of missing data increases to 30%, RMSE values rise consistently across all mechanisms, reaching 0.85 for MCAR, 0.92 for MAR, and 1.12 for MNAR. This trend becomes more pronounced at a 50% missing rate, where RMSE further increases to 0.95 (MCAR), 1.06 (MAR), and 1.32 (MNAR).

External Dataset Validation for Generalisability

This evaluation used the dataset "A continuous real-world dataset comprising wearable-based heart rate variability alongside sleep diaries" ([16]), which provides high-quality in-the-wild physiological recordings. The proposed imputation technique was applied on the dataset to impute features HRV, RHR, sleep duration, sleep efficiency, and sleep-stage composition. Table 8 presents the average imputation performance on the dataset, showing clear improvements from traditional statistical imputation schemes to machine-learning approaches, with the proposed hybrid model achieving the lowest RMSE (0.643) and MAE (0.521). Compared with XGBoost, the proposed method reduces RMSE by 10.8% and MAE by 10.7%, and it outperforms MICE by 16.3% RMSE and 14.9% MAE. These gains reflect stronger recovery of physiological signals–particularly HRV and

Table 8 Average RMSE and MAE comparison between proposed MVI scheme and state-of-the-art techniques over the external dataset

Method	RMSE	MAE
KNN [12]	0.842	0.671
MF	0.811	0.654
MICE [11]	0.768	0.612
XGBoost [13]	0.721	0.583
Proposed approach	0.643	0.521

sleep-stage features–demonstrating that the hybrid architecture generalizes well to real-world datasets and maintains superior prediction accuracy across all measured outcomes.

Theoretical Time Complexity Analysis

The time complexity of the proposed methodology can be broken down as follows:

Feature Importance Analysis

- Pearson's Correlation: Calculating the correlation between two features has a time complexity of $O(N)$, where N is the number of data points. Since we calculate this for all pairs, the total complexity for Pearson's correlation for M features is $O(M^2N)$.
- Random Forest Feature Importance: Training a Random Forest with T trees, each with depth D, has a $O(TDN.\log N)$ complexity. Extracting feature importance is usually $O(T.M)$.
- XGBoost Feature Importance: XGBoost's training complexity is $O(K.N.logN)$, where K is the number of boosting rounds. Extracting feature importance adds a complexity of $O(M)$.
- Aggregated Feature Importance Score Calculation: Combining feature importance scores from the three methods involves normalization and averaging, with a $O(M)$ complexity.

Factor Analysis

The time complexity of performing factor analysis primarily depends on the matrix operations involved. Due to the eigenvalue decomposition step, the complexity of M features and N samples is typically $O(M^3 + M^2N)$.

Athlete Clustering

K-means clustering has a complexity of $O(I.K.N.M)$, where I is the number of iterations, K is the number of clusters, N is the number of data points, and M is the number of features.

Table 9 Time complexity comparison with state-of-the-art techniques

MIV technique	Time complexity
KNN [12]	$O(N^2M)$
MICE [11]	$O(I.M.N.logN)$
EM [10]	$O(I.N.M^2)$
CART [13]	$O(N.logN.M)$
XGBoost [13]	$O(K.N.logN)$
Proposed approach	$O(M^3 + M^2N + NlogN)$

Training XGBoost Regressors

Training an XGBoost model has a complexity of $O(K.N.logN)$. Since we train one regressor per cluster, this becomes $O(C.K.N.logN)$, where C is the number of clusters.

Missing Value Imputation

- Cluster Assignment: Calculating similarity scores with each cluster has an $O(C.M)$ complexity.
- Softmax Probabilities: Computing softmax has a complexity of $O(C)$.
- Weighted Prediction: Calculating the weighted average of predictions has a complexity of $O(C)$.
- Thus, the overall time complexity is $O(M^3 + M^2N + N.logN)$.

Table 9 compares the time complexity of the proposed approach with state-of-the-art MVI techniques. While the proposed methodology exhibits a time complexity comparable to that of KNN and EM techniques, it is higher than that of CART, MICE, and XGBoost. However, these considerations are outweighed by its substantial improvements over state-of-the-art imputation techniques in terms of RMSE (up to 12.20%) and MAE (up to 10.77%). Moreover, predictions of athletic readiness (RSImod) scores over the dataset, imputed using our proposed technique, demonstrate improvements, achieving up to 80.85% in MSE and up to 79.99% in R^2 compared to the current state-of-the-art. In cohort studies that require robust imputation methods, imputation quality directly affects the reliability of subsequent analyses and conclusions. The proposed methodology's ability to achieve significantly lower error rates enhances the accuracy and trustworthiness of imputed data, thereby improving the overall validity and insights derived from cohort studies.

We assessed the practical efficiency of our scheme through empirical benchmarking as shown in (Table 10). To evaluate real-world performance, runtime was recorded for six imputation methods–KNN, EM, MICE, CART, XGBoost, and the proposed Hybrid model–under identical computational conditions. The cohort size was progressively

Table 10 Runtime (seconds) vs sample size

Records	KNN [12]	MICE [11]	EM [10]	CART [13]	XGBoost [13]	Proposed
100	00.92	01.80	02.30	00.41	00.38	00.67
300	03.20	05.90	07.40	01.20	01.10	02.00
500	05.40	09.70	11.80	02.10	01.90	03.40
700	07.80	13.50	16.40	03.00	02.60	04.80
1000	11.40	19.80	23.70	04.30	03.80	06.90

increased to n=100, 300, 500, 700, and 1000 records, replicating the growth pattern of our longitudinal athlete-monitoring datasets.

Limitations

Despite the promising results, the study has certain limitations. The primary dataset consists of monitoring data from 16 athletes observed over 26 weeks, which is typical for elite sports cohort studies but remains relatively small compared with large machine-learning datasets. Small cohort sizes may limit the ability of highly complex models to generalise across populations and may increase sensitivity to noise in the observations. To mitigate this issue, we validated the proposed framework on an independent wearable HRV–sleep dataset comprising 49 participants and approximately 1,300 participant-days of physiological and behavioural records, providing a substantially larger dataset for assessing the generalisability of the proposed imputation framework.

Conclusion

This study proposed a data-driven missing value imputation (MVI) framework for longitudinal cohort studies, demonstrated through a collegiate women's basketball dataset comprising 42 physiological, training, cognitive, travel, and performance features collected over 26 weeks. The primary objective was to robustly reconstruct missing data and improve modelling of athletic readiness, quantified using RSImod, under realistic conditions of high missingness and athlete-specific variability. Comparative evaluations showed that the proposed approach consistently outperformed KNN, EM, MICE, CART, and XGBoost, achieving up to 12.20% lower RMSE and 10.77% lower MAE at the feature level, and up to 80.85% lower MSE and 79.99% higher R^2 for RSImod prediction.

The proposed imputation pipeline follows a structured, modular design comprising four key stages: (i) hybrid feature-importance analysis to identify predictors most sensitive to the target variable, (ii) factor analysis to address multicollinearity and extract latent physiological constructs, (iii) athlete-specific clustering to capture inter-individual variability, and (iv) cluster-weighted ensemble regression using XGBoost for final value reconstruction. An ablative analysis was conducted to evaluate each component of the proposed pipeline. A hybrid feature-importance analysis identified the most influential predictors of RSImod, followed by a factor analysis to reduce redundancy and derive four latent factors–Sleep, Cardiac Rhythm, Training Strain, and Travel Schedule. Athlete clustering enabled personalised imputation, and cluster-weighted XGBoost regressors were used to estimate missing values. Ablation studies confirmed that factor analysis contributed most to performance, followed by hybrid feature weighting and clustering, with the full model achieving the best results (MSE = 0.036, R^2 = 0.667), outperforming a pure XGBoost baseline.

Robustness analyses under MCAR, MAR, and MNAR missingness demonstrated graceful degradation as the proportion of missingness increased. External validation on an independent wearable HRV–sleep dataset further confirmed strong generalisability, with the proposed approach achieving average RMSEs of 0.643 and MAEs of 0.521, outperforming KNN, MICE, and XGBoost by 10–16% across physiological features. SHAP and ALE analyses enhanced interpretability, identifying sleep, training strain, and cardiac rhythm as dominant contributors.

Theoretical analysis shows that the proposed framework has a higher asymptotic complexity than simpler MVI methods due to factor analysis and clustering; however, empirical benchmarks demonstrate scalable runtime with increasing cohort sizes and consistently lower execution times than KNN, EM, and MICE for practical dataset scales. The proposed method substantially improves missing value reconstruction and athletic readiness modelling, offering a reliable solution for real-world longitudinal cohort studies.

Acknowledgements The authors would like to thank Sacred Heart University, USA, all athletes participating in the study, all graduate assistants, and staff for their help in creating data pipelines. The author also extends thanks to undergraduate students at Ahmedabad University for working on this definition during their course project.

Author Contributions Srishti Sharma: Conceptualization; Literature review; methodology; writing – original draft. Hetav Raval: Literature review, Writing, preparation of figures and tables; manuscript formatting; Vishal Barot: review; editing; administrative support. Srikrishnan Divakaran: Statistical analysis (RMSE/MAE/MSE, robustness checks, external validation); interpretation of results; writing – review & editing. Tolga Kaya: Data curation (wearable and performance data collec-

tion); investigation; validation; resources; writing – review & editing. Christopher Taber: Domain expertise (sports science/strength & conditioning); protocol design; interpretation for practice; writing – review & editing. Mehul S Raval: Predictive modelling and experiments (XGBoost, ablations, simulations); formal analysis; visualization (SHAP, ALE plots); writing – review & editing.

Funding Not Applicable

Data Availability Due to the personal nature of the data, it cannot be publicly released. However, an anonymized sample dataset is available on the GitHub repository (a link can be provided upon acceptance).

Materials Availability Not Applicable.

Code Availability Code will be available on the GitHub repository upon paper acceptance.

Declarations

Conflict of Interest On behalf of all authors, the corresponding author states that there is no Conflict of interest.

Ethics Approval and Consent to Participate Before participation, all participants received detailed explanations of the study procedures and provided informed consent with Institutional review board approval number 170720a at Sacred Heart University, USA.

Consent for Publication Not applicable.

References

1. Li J, Guo S, Ma R, He J, Zhang X, Rui D, Guo H Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets. BMC Med Res Methodol 2024; 24(1), 41
2. Harron K, Doidge JC, Goldstein H. Assessing data linkage quality in cohort studies. Ann Hum Biol. 2020;47(2):218–26.
3. Tenan MS. Missing data in sport science: a didactic example using wearables in american football. Sports Med. 2023;53(6):1109–16.
4. Senbel S, Sharma S, Raval MS, Taber C, Nolan J, ..., NSA, Kaya T Impact of sleep and training on game performance and injury in division-1 women's basketball amidst the pandemic. IEEE Access 2022;10, 15516–15527
5. Ribeiro C, Freitas AA. A data-driven missing value imputation approach for longitudinal datasets. Artif Intell Rev. 2021;54:6277–307.
6. D'Ambrosio A, Aria M, Siciliano R. Accurate tree-based missing data imputation and data fusion within the statistical learning paradigm. J Classif. 2012;29:227–58.
7. Pires IM, Hussain F, Garcia NM, Zdravevski E. Improving human activity monitoring by imputation of missing sensory data: Experimental study. Future Internet. 2020;12(9):155.
8. Benson LC, Stilling C, Owoeye OB, Emery CA. Evaluating methods for imputing missing data from longitudinal monitoring of athlete workload. Journal of Sports Science & Medicine. 2021;20(2):188.
9. Griffin A, Kenny IC, Comyns TM, Purtill H, Tiernan C, O'Shaughnessy E, et al. Training load monitoring in team sports: a practical approach to addressing missing data. J Sports Sci. 2021;39(19):2161–71.
10. Jiang N, Gruenwald L. Estimating missing data in data streams. In: Kotagiri R, Krishna PR, Mohania M, Nantajeewarawat E, editors. Advances in Databases: Concepts, Systems and Applications, vol. 4443. Lecture Notes in Computer Science. Berlin/Heidelberg, Germany: Springer; 2007. p. 981–7.
11. Shoaib M, Scholten H, Havinga PJM Towards physical activity recognition using smartphone sensors. In: Proceedings of the 2013 IEEE 10th International Conference on Ubiquitous Intelligence and Computing, Vietri sul Mere, Italy, 2013; pp. 80–87
12. Guo Q, Liu B, Chen CW A two-layer and multi-strategy framework for human activity recognition using smartphone. In: Proceedings of the 2016 IEEE International Conference on Communications (ICC), Kuala Lumpur, Malaysia, 2016; pp. 1–6
13. Ni D, Leonard JD, Guin A, Feng C. Multiple imputation scheme for overcoming the missing values and variability issues in its data. J Transp Eng. 2005;131:931–8.
14. Phung LS Deep learning methods for health data imputation and classification. PhD thesis, Doctoral Dissertation 2021
15. Phung S, Kumar A, Kim J A deep learning technique for imputing missing healthcare data. In: Proceedings of the 2019 IEEE Engineering in Medicine and Biology Society, 2019; pp. 6513–6516
16. Baigutanova A, Park S, Constantinides M, Lee SW, Quercia D, Cha M. A continuous real-world dataset comprising wearable-based heart rate variability alongside sleep diaries. Scientific Data. 2025;12(1):1474. https://doi.org/10.1038/s41597-025-05801-3.
17. Taber CB, Sharma S, et al., M.S.R.: A holistic approach to performance prediction in collegiate athletics: player, team, and conference perspectives. Scientific Reports 2024;14, 1162
18. Barrientos AF, Sen GLD, Dunson DB. Bayesian inferences on uncertain ranks and orderings: Application to ranking players and lineups. Bayesian Anal. 2022;1:1–3.
19. Hall MA Correlation-based feature selection of discrete and numeric class machine learning. Technical report 2000
20. Sharma SU, Divakaran S, Kaya T, Raval M A hybrid approach for interpretable game performance prediction in basketball. In: 2022 International Joint Conference on Neural Networks (IJCNN), 2022; pp. 01–08
21. Kroese DP, Botev Z, Taimre T, Vaisman R. Data Science and Machine Learning: Mathematical and Statistical Methods. ??? CRC Press; 2019.
22. Khobdeh SB, Yamaghani MR, Sareshkeh SK. Machine learning methods for data imputation in biomedical datasets: A comparative analysis. Comput Biol Med. 2020;121:103800.
23. Poggio C, Lechtenberg M, Goodyear P, Lee RL. Evaluating the effectiveness of imputation methods for missing athlete workload data: A systematic review. Sports Medicine - Open. 2020;6:33.
24. Munir T, Martinez M, Al-Jumaily A. Deep learning techniques for real-time missing data imputation: Applications in wearable health monitoring systems. IEEE Trans Biomed Circuits Syst. 2020;14(6):1125–33.
25. Serletis D, Whittaker J. Improving athlete readiness predictions through hybrid machine learning models: A case study in collegiate sports. Journal of Applied Sports Science. 2023;36(4):215–27.