Politecnico di Milano

Hardware architecture for embedded and edge AI

058140

The Relationship Between Internal Load and Sport Performance Outcomes

Oskar Haapalo 11109666
Ellen Lodin, 11102689
Adam Spreitz, 11109973

September 30, 2025

Abstract

This report explores how internal load and wellness metrics relate to match performance in professional volleyball players. Using machine learning on training load, jump, and wellness data, some predictive signals were found, especially from workload ratios and recovery scores. However, overall accuracy was low, highlighting the complexity of performance modeling and the need for richer, more individualized data.

Contents

1 Introduction

Performance in elite volleyball is influenced by how well athletes train, recover, and manage physical stress. Recent research shows that tracking both training load and players well being can help improve match performance and reduce injury risks.

This study explores the relationship between internal load, perceived recovery, and match outcomes in professional female volleyball players. The aim is to identify which wellness parameters are most strongly associated with match performance and how this information can be used to optimize training plans and athlete recovery. To achieve this, machine learning algorithms are applied to data from daily wellness reports and performance metrics.

1.1 Background

Training load, particularly jump activity and strength training, plays an important role in determining match performance. in the De Leeuw paper it is found that stronger lower-body workouts in the four weeks prior to a match were associated with better attack performance. However, high variability in jump counts or excessive upper body training was associated with worse outcomes. These findings emphasize the importance of carefully planning training loads.[de Leeuw et al., 2022]

Clemente et al observed that short-term increases in training load were associated with elevated fatigue, muscle soreness, and decreased sleep quality. Moreover, their study showed that acute load had stronger correlations with well-being than chronic load or training monotony. They also found that the acute-to-chronic workload ratio (ACWR) was highest during mid-season, indicating a period of imbalance between current and accumulated load, and that higher stress levels were observed toward the end of the season. These results suggest that careful monitoring of ACWR is critical for managing fatigue and avoiding non-functional overreaching. [Clemente et al., 2020]

Similar trends have been reported in soccer. Simonelli et al. demonstrated that fatigue and muscle soreness were the strongest predictors of subjective recovery of players. Moreover, these variables mediated much of the relationship between training load and perceived readiness, highlighting the importance of integrating wellness feedback into daily training decisions [Simonelli et al., 2025].

1.2 Problem Description

Previous research demonstrates that internal load factors, such as perceived fatigue, sleep quality, stress, and muscle soreness affect both total quality of recovery and overall training load. Furthermore we aim to explore how overall internal load are associated with in-game performance, and how this insight can be applied in practice. To narrow the scope of this project, we focus on the following research questions:

- Which subjective wellness parameters are most strongly associated with volleyball match performance?
- Is it possible to predict individual/role-specific performance based on wellness-and jump parameters.

By answering these questions, we aim to provide insights into how data driven wellness monitoring can support individualized training strategies and improve overall team performance.

2 Method

2.1 Provided data

The dataset analyzed in this project consider multiple Excel files containing both metadata and time-series performance data related to the volleyball athletes. The data encompass a total of 36 players, 16 from the women's team and 20 from the men's team. Throughout an entire competitive season, five distinct types of datasets were collected: player metadata, daily wellness reports, subjective training load and jump count for each training session or match, and individual match performance records.

The player metadata links each anonymized player ID to their designated role (noted in Italian). Table 1 presents a sample view of the dataset, showing information from five players drawn from both the men's (A1M) and women's (A1F) teams. Since jump data are only available for the women's team and due to the biological gender differences, this report is limited to only analyzing the female athletes.

Athlete	**Role**	**Squad**
player P4T5	Opposto	A1F
player TS3H	Palleggio	A1F
player 8SZL	Giovanilli	A1M
player OOEO	Opposto	A1M
player M9TM	Libero	A1F

Table 1: Athletes and Their Roles by Squad

The provided wellness data set includes subjective self-reported assessments of each player on a daily basis. These assessments cover the following well-being indicators: fatigue, sleep quality, muscle soreness (DOMS), stress levels, mood, and Total Quality Recovery (TQR). Table 2 illustrates a sample of the dataset, showing normalized wellness data and TQR recorded over five different days for the same player.

Athlete	Date	Fatigue	Sleep	Doms	Stress	Mood	TQR
09RI	2023-09-09	-0.047	-1.145	1.483	-0.675	-0.780	0.611
09RI	2023-09-10	1.538	-0.175	1.483	-0.675	-0.780	0.611
09RI	2023-09-11	-1.631	1.764	-0.072	-0.675	-0.780	2.024
09RI	2023-09-13	-0.047	1.764	1.483	-0.675	0.418	0.611
09RI	2023-09-14	-0.047	-0.175	1.483	-0.675	-0.780	0.611

Table 2: Normalized wellness data for athlete 09RI over selected dates

Furthermore, the training load data set captures the Rate of Perceived Exertion (RPE) reported by players for each training session or match. The Training Load (TL) is subsequently calculated as the product of the duration of the session (in minutes) and the corresponding RPE value. Table 3 illustrates a sample of the dataset, showing the total training load, recorded over five different days for a player with ID = '09RI'.

Athlete	Date	Raw TL	Normalized TL
09RI	2023-09-08	500.0	-0.444
09RI	2023-09-09	1700.0	2.390
09RI	2023-09-11	1300.0	1.445
09RI	2023-09-12	600.0	-0.208
09RI	2023-09-13	480.0	-0.491

Table 3: First five entries of raw and normalized training load for athlete 09RI

In addition, the jump load data set captures the **jumps** reported by players. Table 3 illustrates a sample of the dataset, showing the jump load, recorded for five different players during one game.

Athlete	Date	Jumps	Normalized Jumps
TS3H	2023-09-19	112.0	-0.212
62DH	2023-09-19	106.0	1.682
09RI	2023-09-19	0.0	-2.282
8EZC	2023-09-19	39.0	-0.786
XPN2	2023-09-19	114.0	1.050

Table 4: Jump counts and normalized values per athlete on 2023-09-19

The match performance data files offer detailed statistics for each game, including individual metrics for key volleyball fundamentals such as serves

(Battuta), attacks (Attacco), and receptions (Ricezione). An accompanying instruction document is provided to explain the meaning and structure of the various columns. As a general guideline, "=" represent a poor or unsuccessful action, while "#" indicate a perfect action. Table 5 illustrates the serve performance for five players on a random selected game day.

Athlete	=	-	!	+	#
player 09RI	66	96	58	57	35
player P4T5	17	29	17	26	6
player 8EZC	37	130	49	51	10
player NEMC	31	79	33	36	8
player TS3H	62	243	105	147	31

Table 5: Action counts and normalized performance metrics for five selected players

2.2 Data preprocessing

To prepare the dataset for modeling volleyball performance, we integrated wellness, training load, and match performance data. Athlete identifiers and dates were standardized to ensure consistent merging across sources. Key variables such as wellness scores, training load (TL), and jump counts were z-score normalized on a per-athlete basis to account for individual baselines.

In order to enable analysis and comparison between wellness data and individual match performance, each match action was first *quantized* and scored based on its quality and contribution to the game. This was achieved through a weighted scoring system, as shown in Table 6, where technical evaluations (e.g., '+', '-', '=') were mapped to numerical values. The weights were based on the framework proposed by [de Leeuw et al., 2022].
Furthermore each action type was weighted according to Table 8, capturing its overall importance to match outcomes, following the same study's methodology

To capture temporal dynamics, we engineered several features for each athlete:

- **Lagged values** (e.g., `TL_prev`), representing the previous day's measurements.

- **Exponentially Weighted Moving Averages (EWMAs)** over 3, 7, and 14 days, reflecting short- and medium-term trends.

Missing data was handled by flagging rest days and imputing zeros only in those cases. All z-scored and EWMA features were clipped to the range $[-3, 3]$ to mitigate the influence of outliers and stabilize training for neural networks.
Finally, for each match, we constructed a 7-day input window of historical data per athlete, resulting in a three-dimensional input tensor with shape $(\text{samples}, 7\text{ days}, n_{\text{features}})$. This ensured temporal causality and realistic predictive modeling using neural networks later on.

Table 7 shows the individual match performance across five different players and five different actions during one game day.

Action	=	-	!	+	/	#
Battuta	0	7	3	4	4	10
Ricezione	0	1	2	7	5	10
Attacco	0	0	4	6	6	10
Att dopo Ricez	0	5	6	10	10	10
Contrattacco	0	3	5	8	9	10
Muro	0	0	3	4	6	10
Difesa	0	2	5	7	3	10
Free ball	0	4	6	9	9	10
Alzata	0	5	7	9	10	10

Table 6: Action weights for different volleyball actions and outcomes

Athlete	**Date**	**Battuta**	**Ricez.**	**Attacco**	**Att dopo**	**Contratto**
09RI	2024-05-05	0.941	1.608	0.527	1.313	-0.374
P4T5	2024-05-05	2.495	0.000	-0.687	0.000	-0.728
8EZC	2024-05-05	-1.624	0.000	-1.733	-1.430	0.000
NEMC	2024-05-05	0.169	0.000	-0.513	0.011	0.000
TS3H	2024-05-05	0.489	0.000	0.641	0.437	-0.497

Table 7: Normalized action performance metrics per athlete on 2024-05-05

A total individual match performance score and its normalization are calculated by summing the weighted scores across all action types, as shown in Table 9. This is based on weights, see Table 8, proposed by [Marcelino et al., 2008].

Action	**Importance**
Ricezione	1.00
Alzata	0.95
Attacco	0.90
Muro	0.85
Contrattacco	0.80
Battuta	0.75
Att dopo Ricez	0.70
Difesa	0.65
Free ball	0.40

Table 8: Action importance for the game

Index	**Athlete**	**Weighted match performance**	**Perf_Norm**
0	3BL5	339.55	-3.17
1	3BL5	424.35	-0.48
2	3BL5	316.05	-3.32
3	3BL5	623.80	8.64
4	3BL5	379.30	-1.55

Table 9: Match performance and normalized values for payer 3BL5

Workload metrics proposed by [Clemente et al., 2020]: weekly acute load (wAL), chronic load (wCL), and the acute-to-chronic workload ratio (wACWL), are computed separately based on both training load and jump load data. For each wellness parameter, as well as for training load and jump count (including wAL, wCL, and wACWL).

All engineered features were subsequently merged into a unified dataset, indexed by `Athlete` and `Date`. The dataset was then chronologically sorted to ensure consistent temporal ordering. This structure facilitates downstream analysis and guarantees that any modeling approach respects the inherent time-series nature of the data.

2.3 Branching

Early in the process, we recognized that the analysis could be approached at multiple levels: at the team level, where predictions are based on aggregated team performance and features, and at the subgroup level, where performance is analyzed based on player roles or positions (e.g., attackers, liberos, setters). To accommodate these perspectives, we postponed data pruning and instead carried out dataset splitting and subgrouping during post-processing. As a result, the project branched into parallel tracks, one focusing on team-wide analysis for the women, and the other on individual and subgroup-level modeling.

2.4 Correlation Analysis

To explore relationships between input features and match performance, we conducted both linear and rank-based correlation analyses. First, we computed the Pearson correlation matrix between all features and the target variable. This provided insight into linear dependencies. We then constructed a hierarchical clustering dendrogram using a distance metric defined as $1 - |\mathrm{corr}|$, where `corr` is the absolute Pearson correlation. This allowed us to identify groups of highly interrelated features and visualize their structural similarity via a clustered heatmap.

In parallel, we calculated the Spearman rank correlation between each feature and the target to capture non-linear monotonic relationships. Features were then ranked by the absolute value of their Spearman coefficient, highlighting variables with strong predictive monotonic trends even if their linear correlation was weak. These analyses informed both feature selection and model interpretation in subsequent stages.

As a complementary analysis, a Principal Component Analysis (PCA) was also performed to explore the underlying structure of the feature space. PCA helps identify the main directions of variance in the data and can offer insights into potential dimensionality reduction and feature redundancy.

2.5 Machine learning approach

2.5.1 Baseline model

To establish baseline predictive performance, three machine learning models: ElasticNet, Random Forest, and XGBoost—were evaluated due to their suitability for high-dimensional datasets with numerous input features. Model performance was assessed using RMSE and R^2 metrics. Correlation analysis revealed that some features were highly similar, while others contributed minimally to prediction accuracy. Consequently, redundant features were removed and similar ones merged into a refined feature set to test whether this would enhance predictive capability.

2.5.2 Advanced model

Based on the initial baseline results from simpler regression models, we developed a more advanced predictive model. This involved implementing a Temporal Convolutional Network (TCN), which is well-suited for modeling time-series data such as sequences of daily wellness and workload metrics leading up to each match. To maximize the model's predictive performance and reduce estimation error, we incorporated hyperparameter tuning using Optuna, a Bayesian optimization framework. The tuning process explored combinations of architectural parameters—such as kernel size, dilation depth, number of hidden channels, and dropout rate—as well as training-related parameters like learning rate and weight decay. A form of randomized or grid-like search was used across these parameter spaces, guided by validation performance (R^2 score). This allowed us to systematically identify a well-performing model configuration while trying to avoid over- and underfitting.

2.5.3 Evaluation

To evaluate baseline regression models, RMSE and R^2 are used as the primary performance metrics, given that the target variable is continuous. RMSE (Root Mean Squared Error) measures the average prediction error in the same units as the target, providing an interpretable measure of prediction accuracy. R^2 (coefficient of determination) indicates the proportion of variance in the target explained by the model.

3 Results

3.1 Team-Level

3.1.1 Correlation Analysis

The feature correlation heatmap visualizes the pairwise Pearson correlation coefficients between all input variables. In figure 1 the color scale ranges from green to yellow, where:

- **Yellow** indicates a strong positive linear correlation ($r \to +1$), meaning that as one variable increases, the other tends to increase as well.
- **Green** represents a strong negative linear correlation ($r \to -1$), meaning that as one variable increases, the other tends to decrease.

A correlation coefficient of $r = 0$ (mid-range color) implies no linear relationship. This visualization helps identify groups of highly interrelated features, potential redundancy, and the strength of relationships between predictors and the performance target.
In Table 3, the ten features with the strongest positive and negative Pearson correlations to match performance are listed.

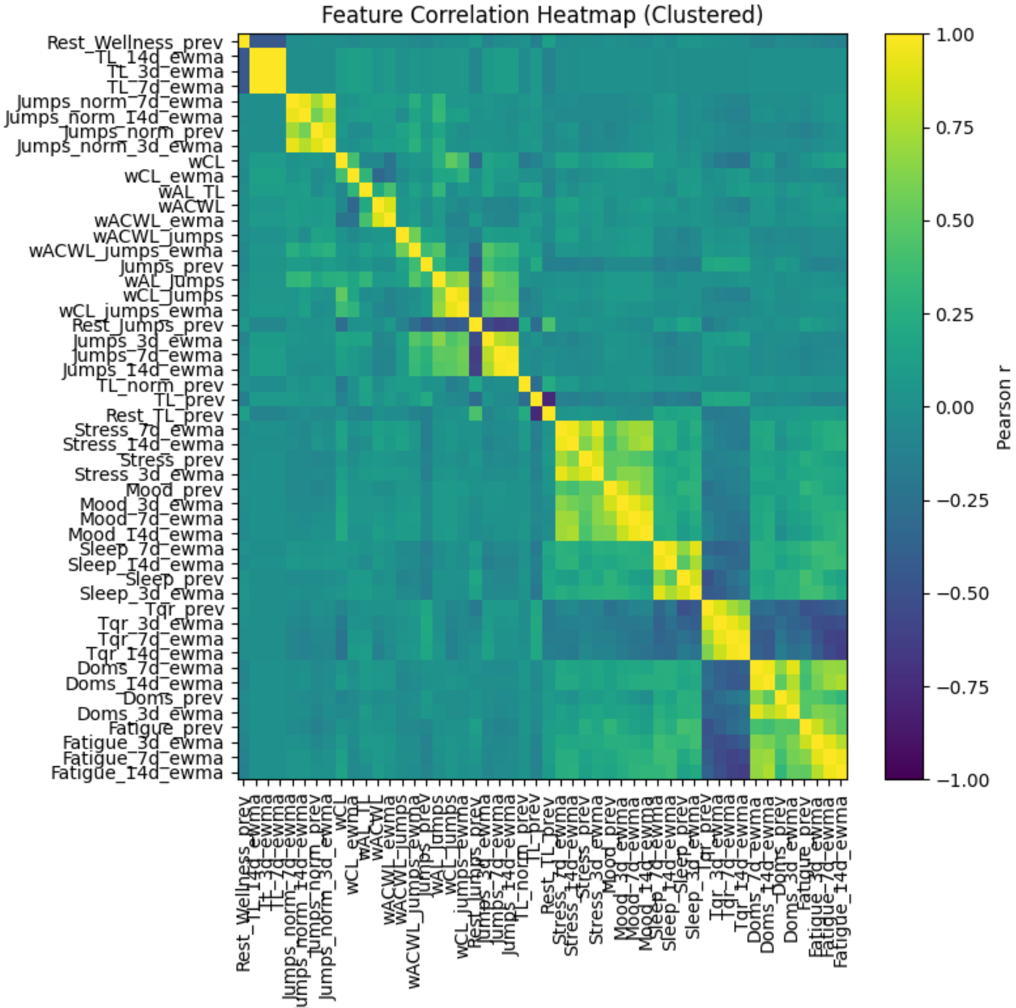

Figure 1: Pearson Correlation Heatmap: Yellow shows positive linear correlation whilst green indicates negative linear correlation.

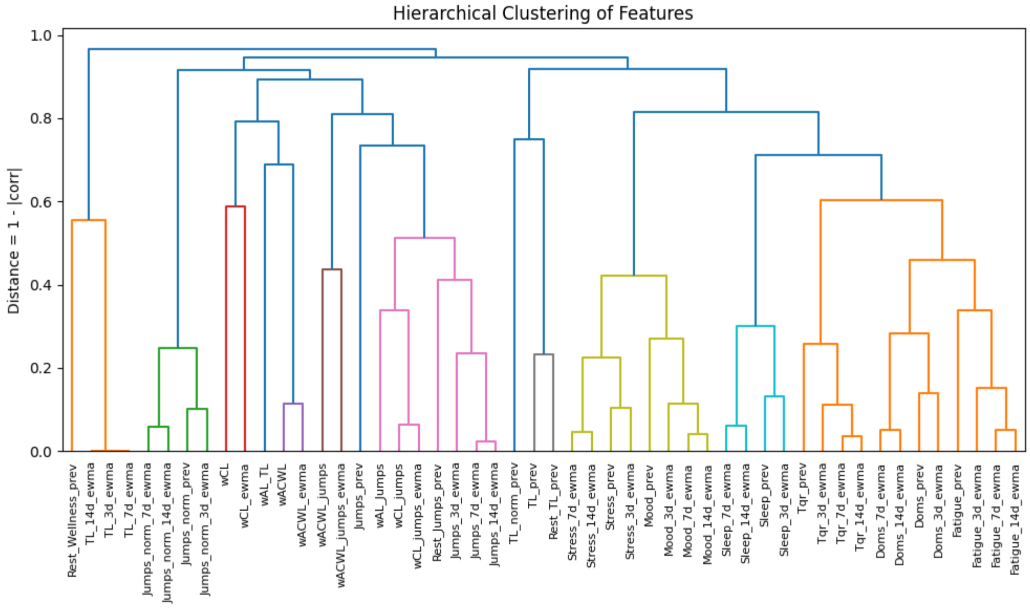

Figure 2: Hierarchical clustering of features using distance $= 1 - |\text{correlation}|$. Closely clustered features exhibit strong correlation (positive or negative).

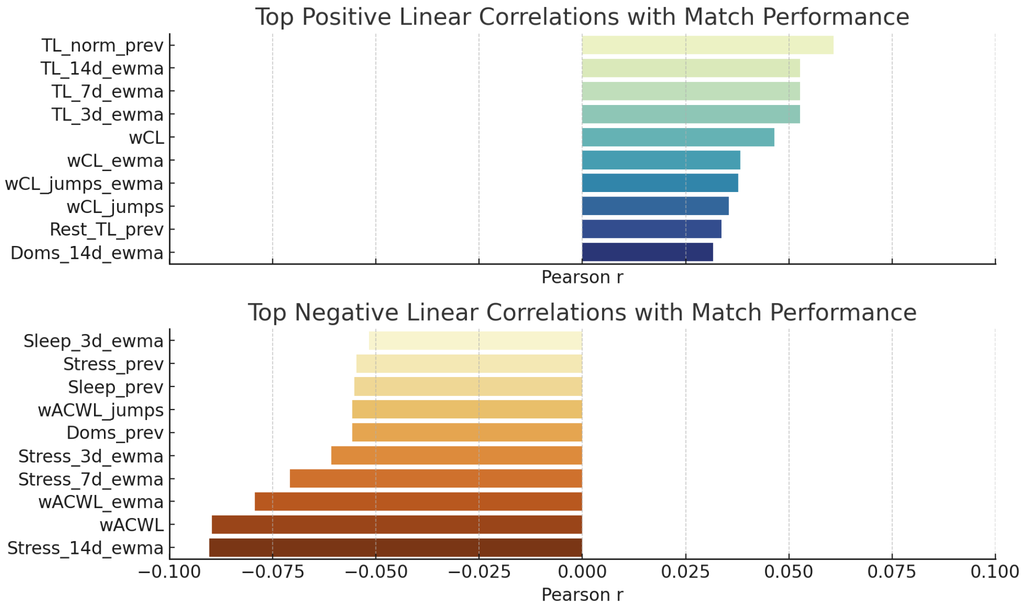

Figure 3: Feature correlation with match performance

Figure 4 shows the Pearson correlation matrix, which captures the linear relationships between features by measuring how strongly they vary together. Values close to +1 indicate strong positive linear correlation, while values near −1 indicate strong negative correlation. This matrix is useful for identifying redundant features, revealing clusters of interrelated variables, and understanding which inputs may influence model predictions similarly.

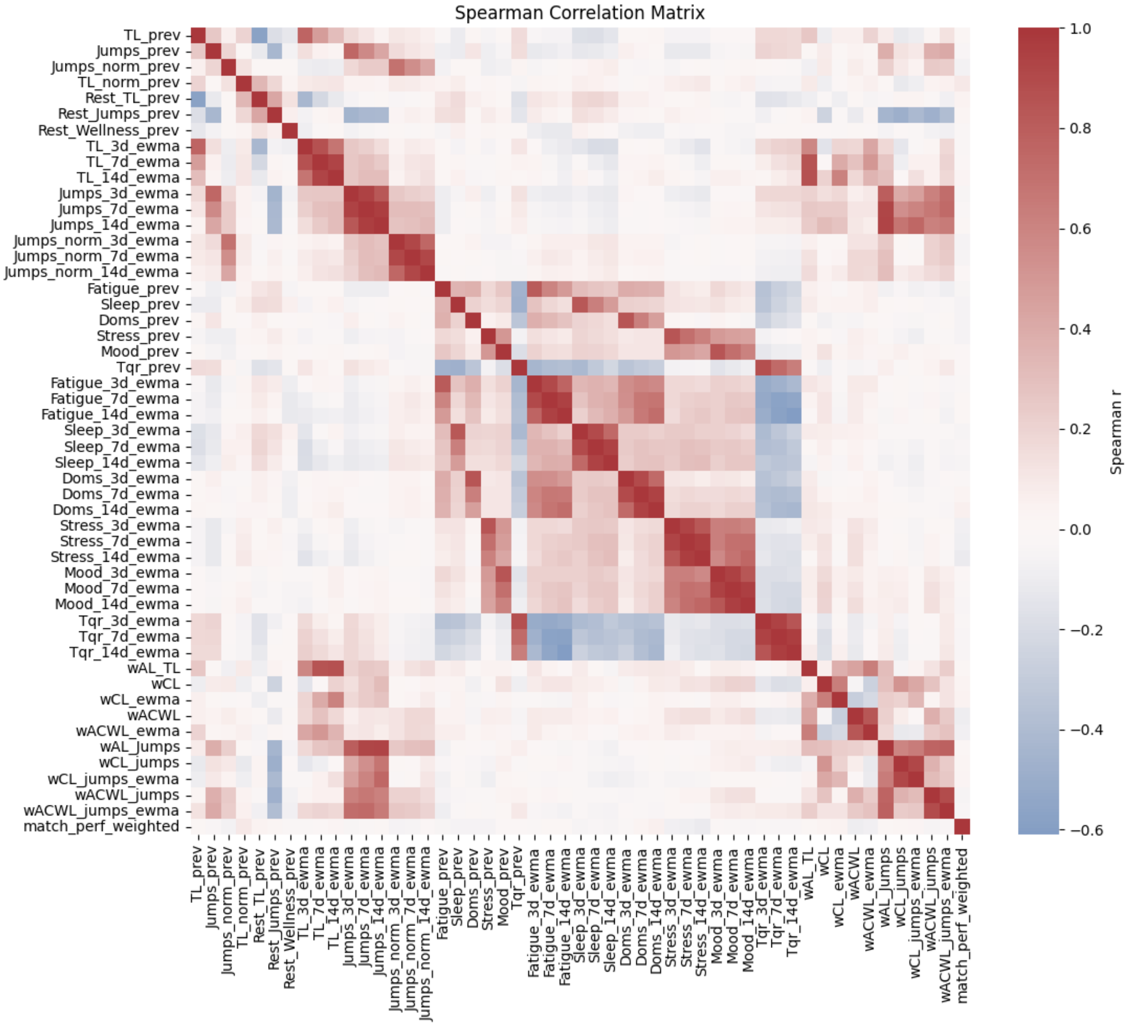

Figure 4: Spearman Correlation Heatmap

Principal Component Analysis (PCA) was used to reduce dimensionality and further explore the relationships between training data and match performance. As seen in figure 5, five main components explaining roughly 55% were found. However, a linear regression model using PC2 and PC1 yielded a very low R^2 of 0.001, indicating that these components explain close to nothing of the performance variance. This suggests that PCA alone is insufficient for prediction, and more advanced models are needed.

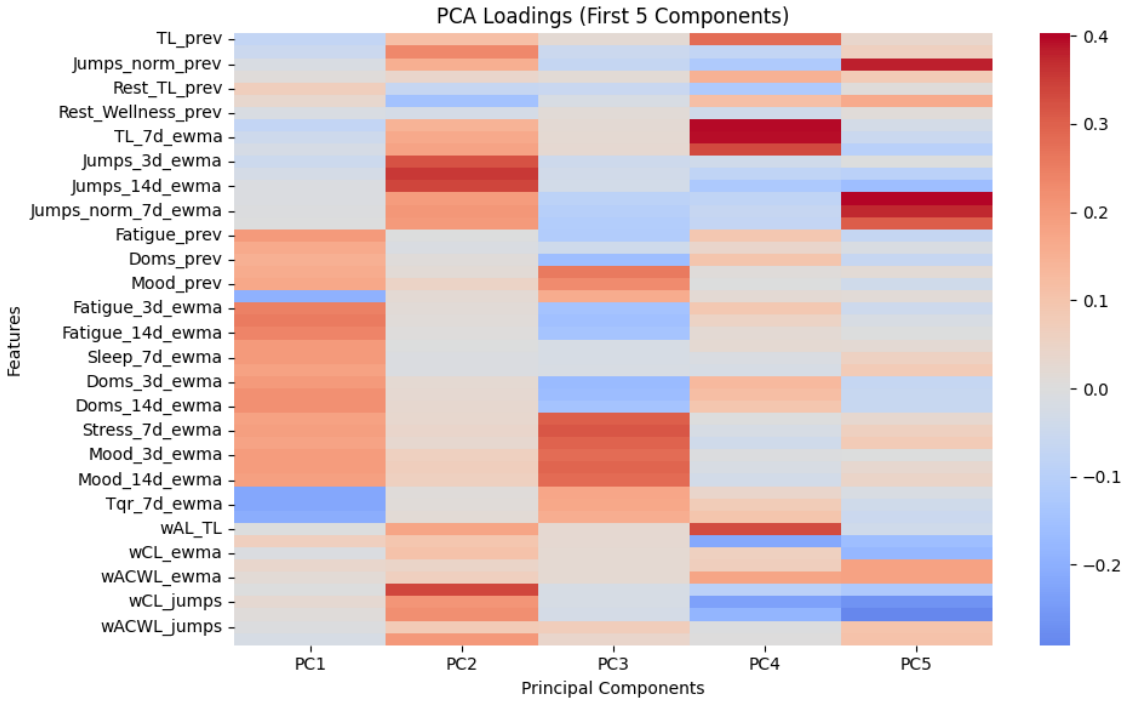

Figure 5: Principal Component Analysis

3.1.2 Baseline models

Table 10 compares the performance of three regression models: ElasticNet, Random Forest, and XGBoost on predicting match performance. It depicts each model's R^2 and RMSE, alongside the top 10 most influential features identified by each model. The ElasticNet model exhibited such strong sparsity that only a single feature retained a non-zero coefficient, suggesting that the relationships between our wellness and match performance are predominantly non-linear and cannot be captured by a simple linear combination of inputs.

Table 11 summarizes the performance metrics achieved after dimensionality reduction. In this version of the model, only a small subset of features was used—selected through domain knowledge and iterative experimentation aimed at maximizing the R^2 score. The final set of predictors included both continuous workload metrics and binary rest indicators, which proved especially important in capturing performance dynamics. Specifically, adding the

	ElasticNet	**Random Forest**	**XGBoost**
	R^2: -0.0036 RMSE: 0.57	R^2: -0.108 RMSE: 0.6	R^2: -0.030 RMSE: 0.58
1	Stress_14d_ewma: 0.0098	Fatigue_prev: 0.046	Fatigue_prev: 0.034
2		wAL_TL: 0.045	Sleep_prev: 0.0257
3		Stress_3d_ewma: 0.042	wACWL_ewma: 0.028
4		Stress_14d_ewma: 0.036	wACWL_jumps_ewma: 0.027
5		Sleep_14d_ewma: 0.035	Stress_3d_ewma: 0.026
6		wCL: 0.347	Tqr_prev: 0.026
7		Doms_14d_ewma: 0.032	Stress_14d_ewma: 0.025
8		Tqr_7d_ewma: 0.031	wACWL: 0.025
9		TL_norm_prev: 0.03	Doms_7d_ewma: 0.025
10		DOMS_3d_ewma: 0.029	Mood_14d_ewma: 0.025

Table 10: Comparison of ElasticNet, Random Forest, and XGBoost model performance and top predictive features. Showing very low predictive power

`Rest_Jumps_prev`, and `Rest_TL_prev` boosted performance by 8%. The selected features were: `wACWL_ewma`, `TL_norm_prev`, `Jumps_norm_prev`, `Tqr_prev`, `Rest_Jumps_prev`, and `Rest_TL_prev`.

Model	$\mathbf{R^2}$	**RMSE**
Reduced ElasticNet	0.0039	0.55
Reduced Random Forest	0.085	0.55
Reduced XGBoost	0.1287	0.53

Table 11: Performance of reduced models after feature selection.

3.1.3 Advanced model

After multiple rounds of Optuna hyperparameter tuning—exploring different window sizes, feature subsets, and trial budgets—we found that a look-back window of five days offered the best trade-off between signal capture and noise avoidance. With `window_size = 5`, the highest validation R^2 achieved during tuning was 0.092. The corresponding hyperparameter configuration and performance metrics are given in Table 12. When retrained on the full train+val split and evaluated on the held-out test set, the final model achieved an RMSE of 0.469 and an R^2 of 0.012.

Table 12: Best TCN Hyperparameters and Performance Metrics

Hyperparameter	**Value**
Window size	5
Hidden channels	2
Kernel size	3
Number of dilations	1
Dropout rate	0.5839
Learning rate	5.99×10^{-5}
Weight decay	1.11×10^{-5}
Best validation R^2	0.09199
Best validation MSE	0.25258
Test RMSE	0.469
Test R^2	0.012

Figure 6 shows the learning curve over 200 epochs using the best hyperparameters from the Optuna study.

Training R^2 starts around -0.15 and gradually increases to approximately $+0.05$ by epoch 200, indicating the model is slowly learning to explain about 5% of the variance in the training data. The pronounced jitter reflects the small dataset size and batch-wise updates.

Validation R^2 rises quickly from about -0.10 to near -0.05 by epoch 30, but thereafter steadily declines back toward -0.08. At no point does the validation curve exceed zero, confirming the model remains worse than a constant-mean predictor on unseen data.

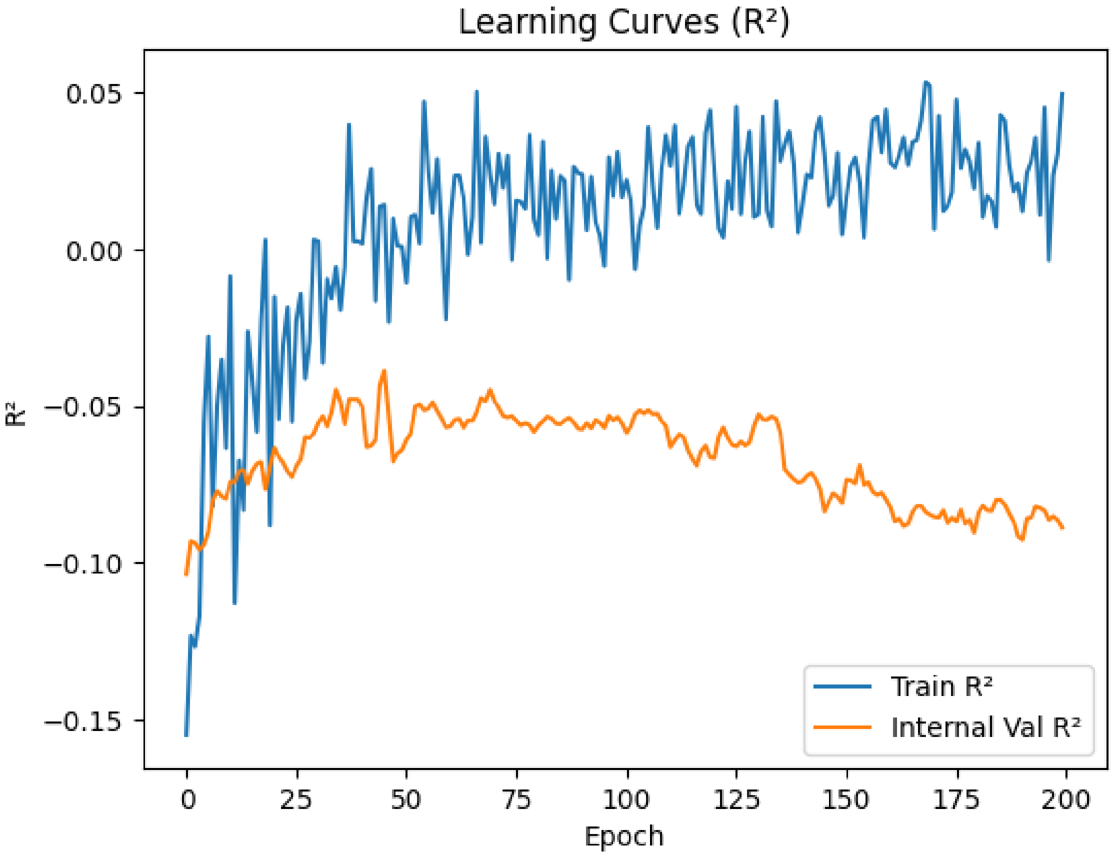

Figure 6: Learning Curve of Temporal Convolutional Network

3.2 Role-Based Subgroup Analysis

3.2.1 Correlation analysis and feature importance by role

Furthering the team-level correlation findings we sectioned all the individual players into their specific role (e.g, setter, blocker, libero). Since the impact of wellness and workload variables may vary by positional demands. The role-specific breakdown aimed to further our understanding of features predictive power within role-subgroups.

In efforts to reduce dimensionality and mitigate collinearity within the time-lagged wellness features we applied once again PCA to each EWMA group, capturing over 90% of the variance in all cases.

To explore role-specific drivers of performance, Pearson correlation analysis was repeated on a per-athlete basis within each role, excluding athletes with insufficient data. Individual results were then averaged to produce role-level feature-performance correlations. The results are visualized in Figure 7, highlighting how features relate to match performance depending on position.

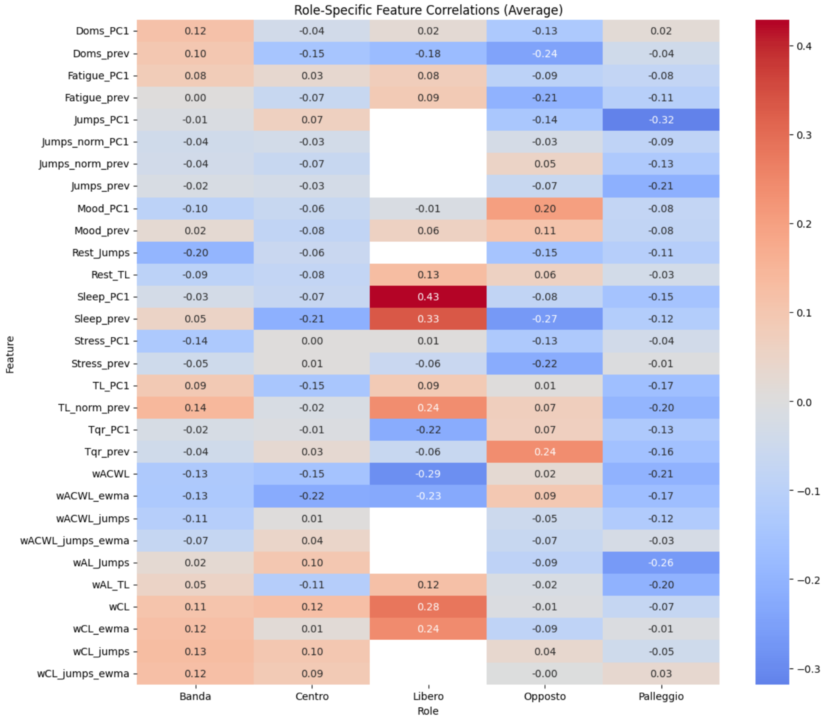

Figure 7: Heatmap of average Pearson per-athlete correlation coefficients between selected features and match performance, grouped by role

- **Libero – Stronger Correlations (with Caution):** Liberos exhibited relatively higher correlations across several features, particularly *Sleep_EWMA_PC1* (0.43) and *Sleep_prev* (0.33), suggesting a stronger link between sleep and performance. Additional positive signals were seen for *Rest_Jumps*, *wCL*, and *wCL_ewma*. However, the small sample size for this role may have inflated these values, warranting cautious interpretation.

- **Banda and Palleggio – Weak and Inconsistent Patterns:** Banda showed small positive correlations for training load features like *TL_norm_prev* (0.14). Palleggio showed predominantly negative associations, especially with *Jumps_EWMA_PC1* (-0.32) and *wAL_Jumps* (-0.26), possibly indicating greater sensitivity to jump-related fatigue.

- **Centro and Opposto – Mild Negative Trends:** These roles revealed slight negative correlations with features tied to fatigue and readiness, such as *Fatigue_prev*, *Stress_prev*, and *Tqr_prev.* For Opposto, *Tqr_prev* showed a weak positive correlation (0.24), hinting that higher perceived readiness might correspond with improved performance.

- **Overall – No Strong Universal Predictors:** No feature emerged as consistently predictive across all roles. This underscores the complexity of performance modeling in team sports and suggests that linear correlations may be insufficient to capture the nuanced, role-specific relationships in the data.

To explore more complex. nonlinear relationships beyond correlation, we trained Random Forest regression models for each athlete within their respective role. These models estimated the relative importance of each feature in predicting match performance. Feature importance were normalized and aggregated by role, and athletes whose models did not meet a minimum R^2 threshold were excluded. This approach provided insight on feature relevance across subgroups, and presented below in Figure 8.

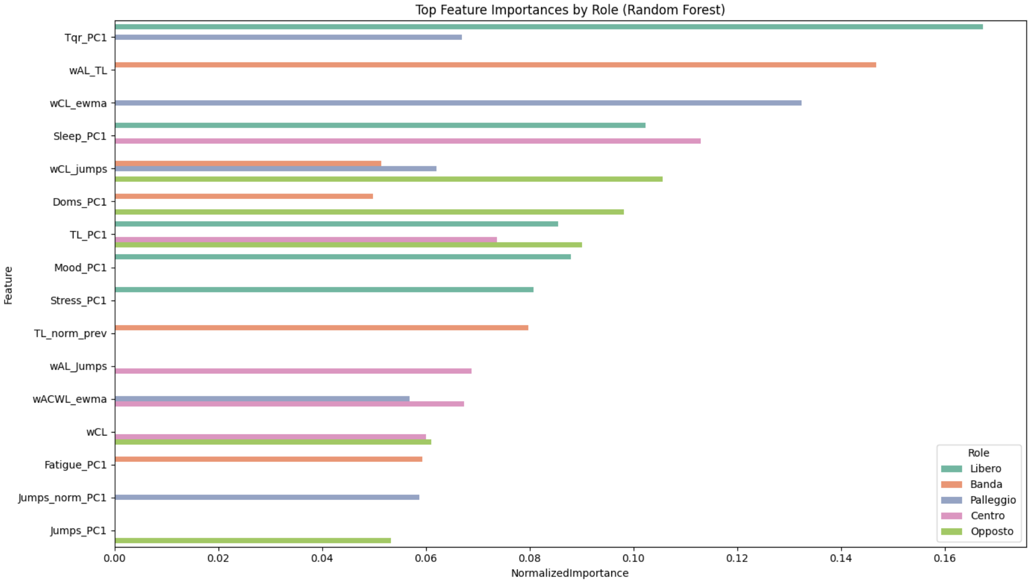

Figure 8: Top 5 most important features per player role as determined by Random Forest models trained individually for each athlete

The feature importance patterns observed further reinforce the role-specific nature of performance drivers:

- **Libero:** *Tqr_PC1* emerged as the most critical feature, suggesting that perceived readiness is a strong performance indicator for this role—aligning with earlier correlation findings around *Sleep* and recovery.

- **Banda:** Features like *wAL_TL*, *TL_norm_prev*, and *Doms_PC1* dominated, highlighting accumulated workload and muscle soreness as influential.

- **Palleggio:** Importance was more dispersed, with *wCL_ewma*, *Jumps_norm_PC1*, and *Tqr_PC1* standing out. This aligns with their mixed correlation signals, indicating a blend of chronic load and readiness effects.

- **Centro:** *TL_PC1* and *Mood_PC1* were most impactful, suggesting a link between emotional state, training load, and performance—consistent with mild negative correlation trends for fatigue and stress.

- **Opposto:** *wCL_jumps*, *TL_PC1*, and *Mood_PC1* were top contributors, pointing again to the importance of jump load and mental state—reinforcing earlier findings from both the correlation and importance perspectives.

These Random Forest results support and expand upon the correlation-based insights, illustrating how different roles rely on different underlying mechanisms for performance.

3.2.2 Baseline model

We replicated the regression experiment introduced in Table 11, but performed role wise. Two modeling approaches were tested: one using the same handpicked subset of features across all roles (previously identified through trial and error), and a second using a role-specific subset of top 5 features derived from feature importance scores presented in Figure 8.

Table 13 summarizes the average R^2 and RMSE values for each model applied per role using the two aforementioned strategies. The role-specific models using tailored features showed modest improvements in certain roles such as Libero and Centro, where Random Forest and XGBoost reached positive R^2 scores, indicating some predictive signal.

Overall, while role-specific predictors provided slightly more nuanced modeling potential in some subgroups, the generally low or negative R^2 values reflect the limited predictive power of the dataset, either due to small sample sizes within roles or inherent unpredictability in match performance given the chosen features.

Role	**ElasticNet R^2**	**Random Forest R^2**	**XGBoost R^2**
Banda	-0.011	-0.023	-0.027
Centro	0.097	0.184	0.082
Libero	-0.034	0.201	0.114
Opposto	-0.188	-0.236	0.123
Palleggio	-0.007	-0.869	-1.148

Table 13: Comparison of model performance (R^2) across roles and regression models

As shown in Table 14, this shared predictor set resulted in overall worse predictive performance, especially for positions like Opposto and Palleggio, where R^2 values dropped as low as -3.08 for XGBoost. These consistently negative R^2 scores across roles indicate that this reduced set failed to generalize well to role-specific demands, reinforcing the need for tailored feature selection. The discrepancy suggests that workload and wellness features influence performance differently depending on positional responsibilities. These findings once again emphasize the limitations of the dataset, either due to insufficient size or the limited explanatory power of wellness metrics alone for predicting match performance.

Role	**ElasticNet $\mathbf{R^2}$**	**Random Forest $\mathbf{R^2}$**	**XGBoost $\mathbf{R^2}$**
Banda	-0.006	-0.277	-0.449
Centro	0.000	-0.263	-0.381
Libero	-0.034	-0.254	-0.820
Opposto	-0.670	-2.137	-3.084
Palleggio	-0.007	-0.570	-0.925

Table 14: Model performance by role using a fixed set of shared predictors across all roles. R^2 scores are generally lower compared to the role-specific setup.

3.2.3 Advanced model

Given the poor performance of the Temporal Convolutional Network (TCN) on the full dataset, and the limited sample size available per role, we did not proceed with advanced deep learning models for the role-specific subsets.

4 Discussion

4.1 Baseline model evaluation

The ElasticNet model highlights the limitations of linear approaches, revealing their poor performance on this dataset. In contrast, tree-based models like Random Forest and Gradient Boosting show stronger results.

Both agree that the most important predictor of match performance is the acute-to-chronic workload ratio. Secondary contributions come from jump readiness and wellness indicators such as stress, DOMS, and TQR.

After feature selection, using a reduced set of features nearly doubled the predictive performance of the baseline models. XGBoost achieved the best results, with an $\mathbf{R^2}$ of approximately 27.5% and an RMSE of around 0.42, making it the most effective model and is chosen as the baseline model to reduce the approximation error.

Although the role-based approach demonstrated some promising predictive potential, highlighting how certain features may impact players differently depending on their position—the overall results suggest clear limitations. We conclude that either the dataset was too small to extract meaningful patterns, or that the selected wellness and workload features are inherently insufficient to reliably predict match performance. This underscores the complexity of performance modeling in team sports and points to the need for richer, more in-depth data to capture the multifaceted nature of athlete output.

4.2 Final model evaluation

4.2.1 Team-Wide

The exceptionally poor generalization of our Temporal Convolutional Network (TCN) highlights the challenge of applying complex deep-learning architectures to a very small, noisy dataset. Although TCNs excel at capturing temporal dependencies, in our case a short look-back window led to severe under-fitting, while extending the window by even two or three days rapidly induced over-fitting. In essence, there simply isn't enough high-quality data for the model to learn robust patterns, and it instead memorizes noise.

However, this does not imply that predicting match performance is impossible. Our reduced XGBoost model still achieved an R^2 of nearly 0.13, demonstrating that—given the right inductive biases—meaningful signal can be extracted.

There are several avenues for improvement:

- **Simpler architectures.** A lightweight feed-forward network (MLP), or even regularized linear models (Ridge/Lasso/ElasticNet), may outperform a TCN when data are scarce.
- **Alternative methods.** Techniques such as *Partial Least Squares (PLS)* regression, *Support Vector Regression (SVR)*, *Gaussian Process Regression*, or ensembles of decision trees (e.g. Random Forest, LightGBM) can be more robust on small datasets.
- **Subgroup modeling.** Training separate models for different player roles (e.g. attackers vs. liberos) could reduce heterogeneity and improve fit.
- **Data augmentation.** Synthetic time-series augmentation—such as window warping, jittering, or generative models (e.g. TimeGAN)—could expand the effective dataset and allow deeper models to generalize.
- **More data.** Ultimately, collecting additional seasons of wellness/logged training and match outcomes would most directly address the data scarcity, though it is expensive in time and resources.

4.3 Further Enhancements

Future work could include studying the men's team to investigate whether there are clear similarities or differences between male and female athletes in terms of wellness and performance patterns. Additional subjective parameters could also be valuable to explore, such as hormonal factors, not least the menstrual cycle in female athletes. It may also be valuable to study and integrate objective health metrics, such as heart rate data. Even subjective wellness factors used, such as sleep quality, can now be measured using modern smart technology.

Expanding the data set by combining data from multiple teams and seasons would increase robustness and generalizability. Additionally, the current approach using weighted sums for performance metrics could be further explored by testing alternative weighting strategies to potentially improve model performance and interpretability.

4.4 Market

Advances in our understanding of how internal load relates to on-court performance, especially in volleyball, are already reshaping the sports-tech landscape. Our study confirms that relying solely on subjective recovery metrics makes accurate performance forecasting very difficult; however, adding objective jump data captured by wearable IMUs markedly improves predictive power. This insight is likely to drive demand for sport-specific smart devices with built-in, personalized analytics and on-device learning. As these wearables mature, their ability to feed more reliable performance models will grow, boosting their practical value. We therefore expect a surge of interest in AI-driven coaching platforms that integrate real-time data analytics, recovery tracking, and performance forecasting.

SaaS (Software-as-a-Service) solutions for professional teams will increase and include machine learning models for injury prevention, workload management, and real-time training advice. As a result, longer athletic careers and more consistent performance will become more valuable, influencing player contracts, insurance policies, and transfer values. At the same time, investment is growing in mobile applications and digital tools to track subjective wellness. These trends may drive the growth of these platforms to scalable and affordable cloud-based performance platforms suitable even for lower level teams.

4.5 Sustainability

Understanding internal load impacts several of the Sustainable Development Goals (SDGs) of the United Nations, both directly and indirectly. For example, SDG 3 (Good Health and Well-being) is supported since it helps players stay healthy, motivated, and engaged throughout their careers. Continuous overload without recovery increases the risk of physical injuries and psychological burnout. By monitoring internal load, training programs can be better tailored to ensure better performance.

Furthermore, this approach contributes to SDG 4 (Quality Education) by promoting awareness and knowledge about sustainable athletic development among players, coaches, and stakeholders. It also supports SDG 5 (Gender Equality) and SDG 10 (Reduced Inequalities) by enabling individualized training and recovery strategies that ensure fair treatment regardless of gender, ethnicity and general background.

By considering safer and more secure athletic environments, internal load monitoring aligns with SDG 8 (Decent Work and Economic Growth), since it helps athletes maintain long-term, healthy careers without injuries. It also guides coaches in avoiding short-term decisions that could compromise athlete's long-term potential.

Additionally, the efficient use of training time and medical or coaching resources contributes to SDG 12 (Responsible Consumption and Production), that ensures that performance is achieved sustainably.

5 Conclusion

Our baseline experiments on the full feature set revealed severe multicollinearity and yielded modest predictive performance. By leveraging both correlation analysis and model-based feature importances, we pruned redundant variables and retrained a reduced XGBoost model, achieving a validation R^2 of approximately 0.13. This result demonstrates that integrating objective jump metrics with normalized training-load indicators can capture meaningful—but limited—predictive signal. In contrast, our advanced Temporal Convolutional Network (TCN), despite extensive Optuna-driven tuning and evaluation across multiple look-back windows, repeatedly oscillated between under- and over-fitting and produced test R^2 values near zero.

Subgroup analyses by player position (attackers, liberos, setters) highlighted that certain features—such as jump-derived workload metrics for attackers and recovery trends for liberos—bear relatively greater importance within those cohorts. However, even these specialized models achieved only modest gains, underscoring the inherent noise and complexity of match performance.

Despite our comprehensive modeling and feature-engineering efforts, accuracy remained modest, reflecting the multifactorial nature of sport outcomes. Unobserved factors—tactical strategies, player rotations, opponent quality, refereeing decisions, and stochastic events—inject substantial noise that wellness and load metrics alone cannot capture. Psychological states, team dynamics, and real-time in-game adjustments further clutters any signal derived from internal-load data.

Key Takeaways and Future Directions

- **Data Enrichment:** Incorporate contextual match variables (e.g., opponent ranking, home-court advantage, match importance) and extend data collection across additional seasons to bolster sample size.

- **Model Simplicity:** Prioritize parsimonious models (e.g., MLPs, Partial Least Squares, ensemble tree methods) over deep architectures when working with limited, noisy datasets.

- **Target Refinement:** Explore classification formulations (e.g., above/below median performance) or athlete-specific deviation models to mitigate label noise.

- **Subgroup Focus:** Develop and rigorously validate role-specific models using tailored feature sets to address cohort heterogeneity.
- **Data Augmentation:** Leverage synthetic time-series generation or data-warping techniques to expand the effective training set and improve model generalization.

In summary, while internal-load and wellness metrics provide valuable insights—particularly when combined with wearable-derived movement data—their standalone predictive utility for match performance is constrained by data scarcity and intrinsic noise. Future breakthroughs will depend on richer, more contextualized datasets, refined feature integration, and model choices calibrated to the realities of elite sports environments.

References

[Clemente et al., 2020] Clemente, F. M., Silva, A. F., Clark, C., Conte, D., Ribeiro, J., Mendes, B., and Lima, R. (2020). Analyzing the seasonal changes and relationships in training load and wellness in elite volleyball players. *International Journal of Sports Physiology and Performance*, 15(5):731–740.

[de Leeuw et al., 2022] de Leeuw, A.-W., van Baar, R., Knobbe, A., and van der Zwaard, S. (2022). Modeling match performance in elite volleyball players: Importance of jump load and strength training characteristics. *Sensors*, 22(20):7996.

[Marcelino et al., 2008] Marcelino, R., Mesquita, I., and Afonso, J. (2008). The weight of terminal actions in volleyball: Contributions of the spike, serve and block for the teams' rankings in the world league 2005. *Faculty of Sport, University of Porto.*

[Simonelli et al., 2025] Simonelli, C., Formenti, D., and Rossi, A. (2025). Subjective recovery in professional soccer players: A machine learning and mediation approach. *Journal of Sports Sciences.* Advance online publication.

A Data Set

The following files are provided as supplementary materials for this report:

- **anonymized_matches_F.xlsx**: Contains anonymized match performance data used for model training and evaluation.
- **anonymized_wellness_file.xlsx**: Contains wellness questionnaire responses aligned with the match data.
- **matches_files_explination.txt**: Provides detailed descriptions of columns in anonymized_matches_F.xlsx.

These files are provided separately and should be reviewed alongside the report for full context and reproducibility.