

Article

Interpretable Decision Support for Next-Morning Soreness in Elite Women's Football

Tomasz Piłka^{1,*}, Martyna Ławniczak^{1,2,*}, Tomasz Górecki¹, Kaja Dziergas² and Bartłomiej Grzelak^{1,2}

¹ Faculty of Mathematics and Computer Science, Adam Mickiewicz University, 61-614 Poznań, Poland; tomasz.gorecki@amu.edu.pl (T.G.); bartlomiej.grzelak@lechpoznan.pl (B.G.)

² KKS Lech Poznań, 60-320 Poznań, Poland; kaja.dziergas@lechpoznan.pl

* Correspondence: tomasz.pilka@amu.edu.pl (T.P.); marlaw6@st.amu.edu.pl (M.Ł.)

Abstract

This paper presents a retrospective proof-of-concept development and preliminary evaluation of an interpretable decision-support system for daily fatigue-risk management in one elite women's football club. The system integrates morning wellness, GPS-derived external load, and a daily wellness-duration internal-load proxy at the player-day level. It combines a player-day integration layer, an interpretable predictive layer, and a recommendation layer that returns one of three staff-facing actions: *Reduce*, *Maintain*, or *Progress*. The predictive model outputs a calibrated probability of elevated next-morning self-reported soreness. The target is a subjective questionnaire outcome, not an injury, medical diagnosis, or objective marker of recovery. The decision-support layer maps this probability to a three-state recommendation, informed by a review threshold, operational guardrails, and staff oversight. Using retrospective monitoring data from two competitive seasons (2024/25 and 2025/26) in a single professional team, we evaluated the proposed approach using rolling-origin temporal validation, leave-one-player-out cross-validation, and between-season validation. To separate genuine predictive signal from the day-to-day persistence of soreness, we report a baseline ladder ranging from a trivial persistence rule to the full model, with bootstrap confidence intervals for performance differences. Under rolling-origin validation across 16 monthly folds, the final logistic regression model achieved a mean ROC-AUC of 0.759 ($SD = 0.089$). Critically, a model excluding current soreness still outperformed the persistence baseline (ROC-AUC 0.738 vs. 0.721), and the isolated contribution of current soreness was modest but reliable ($\Delta\text{ROC-AUC} = +0.044$, 95% CI [+0.027, +0.059]). Between-season validation (train: 2024/25; test: 2025/26) yielded an ROC-AUC of 0.801. The three-state recommendation layer separated outcomes monotonically, with observed next-morning soreness rates of 0.054 for *Progress*, 0.217 for *Maintain*, and 0.326 for *Reduce* ($p < 0.001$ for the *Progress*-versus-*Maintain* contrast). These preliminary findings support the feasibility of the proposed approach within the club studied. However, because the model requires complete wellness, GPS, and proxy data, it operates only on the fully monitored on-pitch stratum (3386 of 17,703 player-days); the reported performance therefore applies to this stratum rather than to a typical player-day, and prospective evaluation and external validation by independent teams are required before broader implementation can be considered.



Academic Editor: Grigorios Beligiannis

Received: 1 July 2026

Revised: 22 July 2026

Accepted: 23 July 2026

Published: 3 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

Keywords: decision-support systems; interpretable machine learning; athlete-monitoring data; soreness prediction; workload monitoring; women's football

1. Introduction

Elite football organizations increasingly collect daily athlete-monitoring data, including morning wellness questionnaires, GPS-derived external-load measures, and indicators of internal load. The practical challenge is to integrate these incomplete, heterogeneous, and time-sensitive data into operational decisions about whether to reduce, maintain, or progress a player's planned training load.

Short-term fatigue and recovery reflect interactions between current readiness, recent workload, and individual responses to training. GPS-derived variables quantify external demand, internal-load measures describe the athlete's response to that demand, and wellness questionnaires provide practical information on fatigue, soreness, sleep quality, and mood [1–4]. Workload indicators such as the Acute:Chronic Workload Ratio may provide additional context but should not be treated as standalone representations of fatigue or deterministic decision rules [5]. Biomechanical research also indicates that fatigue-related risk may involve cumulative tissue loading and changes in movement mechanics that aggregate workload indicators do not directly capture [6,7].

Although these methodological components are well established, existing research provides limited guidance on how they should be combined into a complete operational decision-support workflow [8–10]. Player-day representations are commonly used in longitudinal athlete-monitoring studies, but their integration with interpretable prediction and explicit staff-facing recommendations has received less attention.

To address this gap, this study presents and retrospectively evaluates a proof-of-concept interpretable decision-support system for daily fatigue-risk management in elite women's football. The system integrates morning wellness, GPS-derived external load, and a wellness-duration internal-load proxy at the player-day level. It estimates the calibrated probability of elevated next-morning self-reported soreness and maps this estimate, together with predefined operational guardrails, to one of three recommendations: *Reduce*, *Maintain*, or *Progress*.

The predicted outcome is a subjective questionnaire measure intended as a pragmatic early-warning signal for staff review. It should not be interpreted as an injury outcome, medical diagnosis, or objective measure of recovery. Similarly, the wellness-duration proxy is not equivalent to a conventional post-session rating of perceived exertion because it combines pre-session wellness with session duration, rather than relying on an independent post-session CR-10 rating.

The research question guiding this study was as follows: *How can an interpretable decision-support information system integrate heterogeneous daily athlete-monitoring data to support early warning of elevated next-morning self-reported soreness in elite women's football?*

Based on the study objectives, we evaluated two primary hypotheses:

H1. *The integrated model would retain predictive discrimination for elevated next-morning soreness under temporal, between-season, and leave-one-player-out validation.*

H2. *Integration of wellness, GPS-derived external load, and the internal-load proxy would provide higher discrimination than models based on any single data source.*

Comparisons between logistic regression and gradient boosting, the persistence baseline ladder, feature-level associations, threshold sensitivity, and the three-state recommendation analysis were treated as exploratory. Because the study was retrospective and not preregistered, the hypotheses represent directional analytical expectations rather than preregistered confirmatory tests.

The primary contribution of this study is not a new predictive algorithm. Logistic regression, wellness measures, GPS-derived variables, and workload descriptors are al-

ready established methodological components. Rather, the contribution lies in integrating them into an operational decision-support architecture that: (1) aligns subjective readiness, external load, and proxy-derived internal load within a unified player-day representation; (2) produces calibrated predictions of next-morning soreness; (3) translates these predictions and predefined guardrails into explicit staff-facing recommendation states; and (4) evaluates the resulting architecture using temporal, between-season, leave-one-player-out, persistence-baseline, and source-ablation analyses. The system is intended to support, rather than replace, practitioner judgment [8,11].

2. Related Work

2.1. Athlete Monitoring, Recovery, and Fatigue Modeling

Research on athlete monitoring commonly distinguishes between external load, that is, the work performed by the athlete, and internal load, that is, the individual psychophysiological response to that work. In football, GPS-derived variables such as total distance, high-speed running, sprint distance, accelerations, decelerations, and player Load are widely used to quantify external demand, whereas session rating of perceived exertion (sRPE) remains one of the most practical and widely adopted indicators of internal load in applied settings [1]. Self-reported wellness measures provide complementary information about fatigue, soreness, sleep quality, and mood and have proven feasible for routine monitoring in elite team-sport environments [2–4]. Recent research in elite youth soccer further showed that different GPS-derived external load variables predicted distinct components of perceived exertion, reinforcing the multidimensional relationship between external and internal training load [12]. Taken together, the literature supports the use of multiple monitoring streams rather than reliance on any single indicator.

A second line of work focuses on workload progression, accumulated stress, and recovery dynamics. Indicators such as acute load, chronic load, monotony, strain, and the Acute:Chronic Workload Ratio (ACWR) have been widely used to describe recent exposure and contextualize training stress [5]. However, these conventional ratio-based summaries have also been the subject of substantial methodological criticism, particularly regarding their statistical validity, interpretability, and operational use as decision rules. This issue is discussed in detail in Section 2.2. Moreover, short-term recovery states are shaped not only by accumulated load but also by current readiness, session structure, and the specific mechanical and neuromuscular demands imposed by football training. For example, small-, medium-, and large-sided games may generate substantially different locomotor and neuromuscular profiles even when total duration is comparable [13].

This problem is particularly relevant in elite women's football, where evidence derived from routine training-load monitoring remains comparatively sparse and heterogeneous [14]. In addition, female-specific contextual factors, including menstrual-cycle-related variation, may influence sleep, perceived fatigue, soreness, and readiness, thereby affecting the interpretation of wellness data and recovery status [15]. Despite growing interest in athlete monitoring, existing studies offer limited guidance on how heterogeneous information streams should be organized into a coherent player-day representation to support day-to-day operational decisions.

2.2. Workload Ratios and Machine Learning for Injury and Fatigue Risk

Considerable attention has been devoted to workload-based indicators, particularly the ACWR. Although originally introduced as a practical framework for relating recent training exposure to longer-term workload history [5], the concept has become increasingly controversial. One important criticism concerns mathematical coupling: because the acute workload is also a component of the chronic workload in conventional ACWR calculations,

the resulting ratio may induce spurious correlations and inflate apparent associations with outcomes [16]. Broader methodological critiques have further questioned the conceptual basis, statistical robustness, and evidential support for the ACWR as a valid predictor of injury or fatigue-related outcomes [17,18].

These critiques do not imply that recent workload history is irrelevant. Rather, they suggest that ratio-based workload descriptors should not be interpreted as standalone causal indicators or deterministic thresholds for decision-making. In applied monitoring contexts, workload ratios are more defensible when treated as contextual descriptors within a broader multivariate representation of player status. In the present study, session-based Time-Dependent ACWR (TD-ACWR) is therefore used as one component of an interpretable feature set rather than as an isolated risk rule. By defining acute and chronic windows in terms of recent training events rather than fixed calendar periods, the approach aims to reduce noise from irregular microcycles while avoiding overreliance on any single ratio-derived indicator.

In parallel, machine learning has emerged as a promising approach to modeling injury and fatigue risk as multifactorial, partly nonlinear phenomena. Beyond sport, machine-learning-based fatigue detection has also been investigated in other safety-critical domains, including EEG-based pilot monitoring, illustrating the broader potential of computational systems for fatigue-related early warning [19].

In football, Rossi et al. [20] demonstrated that GPS-derived training data could be used for injury forecasting, showing that predictive models may identify elevated-risk situations with practically relevant sensitivity, although precision remains constrained by the low prevalence and complexity of injury events. Similarly, Piłka et al. [21] showed that GPS-based wearable sensor data could support machine-learning-based prediction of football injuries, reinforcing the relevance of external load monitoring as a predictive data source in football-specific risk modeling. More recently, Majumdar et al. [22] reviewed machine-learning approaches for understanding and predicting injuries in football, emphasizing both their potential and their persistent limitations, including small datasets, limited external validation, insufficient interpretability, and challenges in practical deployment.

However, much of the literature remains prediction-centered. Existing studies most often evaluate whether an adverse event can be forecast but provide less guidance on how model outputs should be integrated into routine performance workflows. In particular, fewer studies address how heterogeneous monitoring streams should be aligned at the player-day level, how probabilistic outputs should be calibrated, and how predictions should be translated into actionable recommendations under human supervision. This leaves an important gap between predictive modeling and operational decision support.

2.3. Interpretable Analytics, Calibration, and Human-in-the-Loop Decision Support

A related body of work concerns interpretable machine learning and decision-support systems. Explainable artificial intelligence methods, including feature-attribution approaches such as SHAP, can improve transparency by showing how individual variables contribute to model behavior [23]. Nevertheless, in applied performance environments, predictive performance alone is insufficient. Staff-facing systems must generate outputs that are understandable, actionable, and compatible with existing decision-making workflows [8]. This requirement is consistent with human-in-the-loop machine learning, where analytical systems are designed to support expert judgment rather than replace it [11]. Related research on IoT-enabled human-machine systems similarly emphasizes that system performance depends on the dynamic interaction between automated components and human operators, rather than on algorithmic output alone [24].

Sports-specific decision-support systems differ substantially in their intended decisions, analytical scope, and level of empirical maturity. Ofoghi et al. [25] developed a software-supported framework that combined statistical, probabilistic, and machine-learning methods to assist with athlete selection and tactical planning in track cycling. Although this work demonstrated how analytical outputs can be connected to concrete coaching decisions, it addressed a narrowly defined competition-level task and relied on historical performance results rather than longitudinal readiness and workload monitoring. In elite men's basketball, Burger et al. [26] described athlete-monitoring systems as centralized platforms for integrating, processing, and visualizing multiple data streams to inform daily load and recovery management. However, predictive analytics and artificial intelligence were presented primarily as future extensions of these systems rather than as empirically validated probability-to-action pathways. More recent approaches have emphasized time-series analysis and broader analytical architectures. Schelling et al. [27] proposed a sports decision-support framework based on time-series segmentation and changepoint detection, illustrated by identifying the most demanding scenarios in basketball. This approach provides temporally contextualized performance information, but its primary output is retrospective characterization of demand rather than individual next-day risk stratification. Mănescu [28] presented a broader framework integrating data acquisition, predictive analytics, visualization, and alerts across football, basketball, and athletics. Its cross-domain architecture was comprehensive, but the reported demonstrations relied on synthetic scenarios and therefore provided limited evidence regarding routine use with longitudinal data from elite athletes.

From a prognostic modeling perspective, evaluation should extend beyond discrimination metrics such as ROC-AUC. Calibration is particularly important when predicted probabilities are used to support decisions, because poorly calibrated models may provide misleading risk estimates despite acceptable ranking performance [29]. In the present study, practical performance is further characterized through threshold sensitivity analysis, including precision, recall, F1, and alert burden across alternative operating thresholds.

These considerations align with the information systems perspective, where the objective is not merely to optimize predictive accuracy but to design sociotechnical systems that integrate data, analytics, interpretability, and human decision processes [8–10]. Existing research offers valuable building blocks for athlete monitoring, workload modeling, interpretable analytics, and human-centered decision support. However, relatively little work has integrated these components into a coherent player-day decision-support information system for elite football environments.

Taken together, previous systems range from competition-level tactical support and centralized monitoring dashboards to time-series analysis and simulated cross-domain analytical frameworks. The present study does not claim to introduce the first decision-support system in elite sport. Its more specific contribution is to connect three elements that have rarely been evaluated jointly using longitudinal real-world club data: player-day alignment of wellness, GPS-derived external load, and proxy-derived internal-load information; calibrated prediction of a next-morning subjective outcome under temporal and cross-player validation; and translation of that probability into explicit staff-facing recommendation states using a fixed threshold, predefined guardrails, and human oversight. At the same time, the proposed interface and human-in-the-loop workflow were not prospectively evaluated, so the study should be interpreted as an empirical evaluation of a decision-support architecture rather than validation of a fully deployed operational system.

3. System Design

The proposed system is a proof-of-concept interpretable decision-support information system designed for daily fatigue risk management in elite women's football. Its main objective is to transform heterogeneous athlete-monitoring data into a single, player-centered daily representation that supports timely, actionable training decisions. Rather than treating wellness, external load, and proxy-derived internal-load information as distinct analytical domains, the system integrates them into a single operational feedback process. The design follows three layers: a player-day integration layer, an interpretable predictive layer, and a recommendation layer that returns staff-facing actions under human oversight.

At the input level, the system combines three categories of daily signals. The first captures the athlete's subjective readiness before training through morning self-report. The second represents external demand through GPS-derived workload measures. The third is a wellness-duration internal-load proxy calculated from morning wellness and session duration. Because it does not use an independent post-session CR-10 rating, this measure should not be interpreted as conventional sRPE or as a direct measure of the athlete's post-session internal response [1,30]. To retain temporal context, the integration layer further derives rolling and change-based workload descriptors of the proxy and external load. The specific variables, transformations, and rolling windows used to operationalize each category are detailed in Section 4.

Figure 1 illustrates the overall architecture. At the conceptual level, the system follows a closed daily fatigue loop: *morning wellness* → *training stimulus* → *wellness-duration internal-load proxy* → *next-morning soreness*. The core unit of integration is the *player-day*, defined as one athlete observed on a single calendar day. This representation combines subjective readiness, external demand, and proxy-derived internal load information into a coherent temporal record for monitoring and prediction.

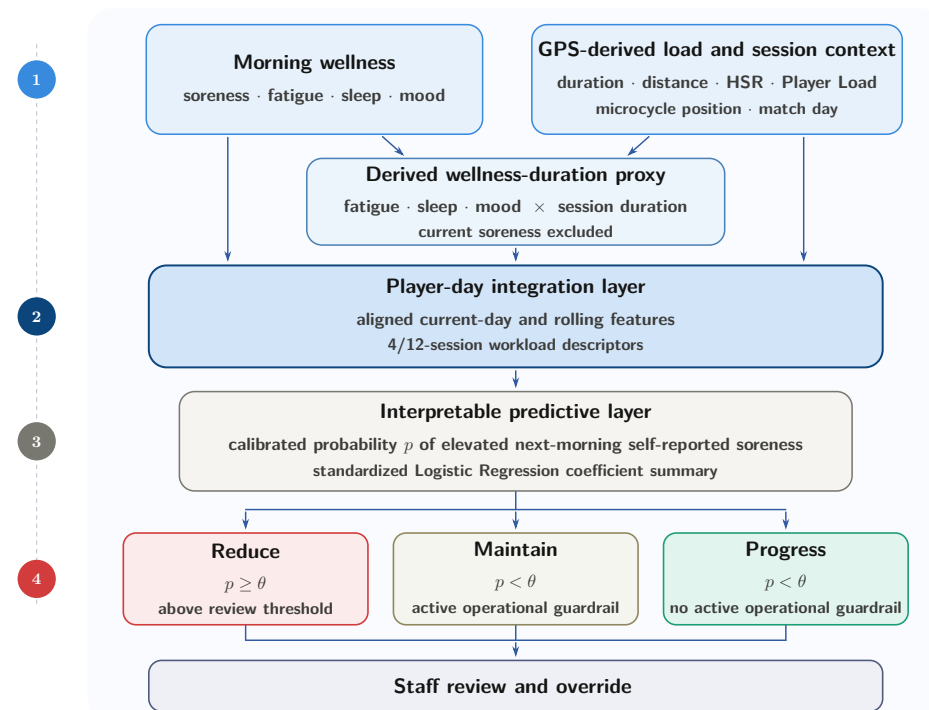


Figure 1. Architecture of the proposed proof-of-concept decision-support information system. The predictive layer outputs a single calibrated probability of elevated next-morning soreness, and the recommendation layer converts this score, together with operational guardrails and staff oversight, into one of three staff-facing outputs: *Reduce*, *Maintain*, or *Progress*.

On top of this integration layer, the system applies a comparatively transparent predictive model. Standardized logistic regression coefficients indicate the direction and relative magnitude of associations between predictors and estimated risk. This provides model-level transparency but does not establish that the resulting information is readily understood or operationally useful to practitioners.

The recommendation layer converts the calibrated probability into one of three operational recommendation states. *Reduce* is returned when the predicted probability meets or exceeds the review threshold. Below this threshold, *Maintain* is assigned when at least one conservative guardrail is active, namely elevated current soreness or fatigue, or a marked increase in recent workload. *Progress* is returned only when the predicted probability is below the threshold, and no guardrail is active. Formally, let $p_{i,t}$ denote the calibrated probability of elevated soreness on the morning of day $t + 1$ for player i , estimated using information available for day t . Let $G_{i,t} \in 0, 1$ indicate whether at least one operational guardrail is active. The recommendation state is defined as

$$R_{i,t} = \begin{cases} \text{Reduce}, & p_{i,t} \geq \tau, \\ \text{Maintain}, & p_{i,t} < \tau \text{ and } G_{i,t} = 1, \\ \text{Progress}, & p_{i,t} < \tau \text{ and } G_{i,t} = 0, \end{cases} \quad (1)$$

where τ denotes the probability threshold triggering high-priority staff review.

Unlike the probability threshold τ , which was selected empirically using data from the analyzed club as described in Section 5.5, the operational guardrail cutoffs were specified a priori and were not optimized against the retrospective outcome. Scores of ≥ 4 on the club's 5-point wellness scale were treated as clearly elevated soreness or fatigue. A session-based TD-ACWR ≥ 1.5 for the internal-load proxy (srpe_td_acwr in Table 1) was used as a conservative operational flag for a marked increase in recent workload, informed by prior workload literature [5].

Table 1. Internal-load-proxy and external-load predictors included in the player-day model.

Predictor	Definition	Window
<i>Internal-load-proxy variables</i>		
srpe_td_acwr	Session-based TD-ACWR of the wellness-duration internal-load proxy.	4/12 sessions
srpe_td_acute4	Mean internal-load proxy over the four most recent sessions.	Last 4 sessions
srpe_td_chronic12	Mean internal-load proxy over the preceding twelve sessions.	Previous 12 sessions
srpe_td_relchange	Relative change in the internal-load proxy, calculated as (acute – chronic)/chronic.	4/12 sessions
rpe_td_acwr	Session-based TD-ACWR of the wellness-derived perceived-intensity composite.	4/12 sessions
<i>External-load variables</i>		
dist_td_acwr	Session-based TD-ACWR of total distance.	4/12 sessions
dist_td_acute4	Mean total distance over the four most recent sessions.	Last 4 sessions
dist_td_chronic12	Mean total distance over the preceding twelve sessions.	Previous 12 sessions
dist_td_relchange	Relative change in total distance, calculated as (acute – chronic)/chronic.	4/12 sessions
pl_td_acwr	Session-based TD-ACWR of player load.	4/12 sessions
hi_td_acwr	Session-based TD-ACWR of high-intensity running distance.	4/12 sessions
dur_td_acwr	Session-based TD-ACWR of session duration.	4/12 sessions

Note. TD-ACWR denotes the session-based Time-Dependent Acute:chronic Workload Ratio, calculated as the acute mean divided by the chronic mean. Acute and chronic windows were defined in terms of discrete training sessions rather than calendar days. The internal-load proxy was calculated as the morning wellness composite (fatigue, sleep, and mood; excluding soreness), rescaled to a Borg-like 0–10 range, multiplied by GPS-recorded session duration. The variable names retain the original analytical-code labels (srpe and rpe), but all such variables in this study were derived from the wellness-duration proxy and do not represent conventional post-session sRPE. Wellness and contextual predictors are presented in Table 3.

Both the probability threshold and the guardrail cutoffs should be regarded as club-specific operational rules rather than validated clinical thresholds or universally transferable safety boundaries. Their definitions and practical consequences depend on the wellness scale, monitoring procedures, training structure, staff capacity, and the analyzed

club's risk tolerance. Accordingly, these rules should be recalibrated and prospectively validated before being applied in other teams or monitoring environments.

Accordingly, the guardrail indicator is defined as

$$G_{i,t} = \mathbb{I}\left(S_{i,t} \geq 4 \vee F_{i,t} \geq 4 \vee \text{TD-ACWR}_{i,t}^{(L)} \geq 1.5\right), \quad (2)$$

where $\mathbb{I}(\cdot)$ is the indicator function, $S_{i,t}$ and $F_{i,t}$ denote current soreness and fatigue, respectively, and $\text{TD-ACWR}_{i,t}^{(L)}$ denotes the TD-ACWR of the internal-load proxy. An illustrative player-day example showing the *Reduce* branch and the associated staff-review workflow is provided in Section 5.5.

Importantly, the names of the three recommendation states—Reduce, Maintain, and Progress—denote suggested directions for training load adjustment for the upcoming session, rather than absolute diagnostic evaluations of the athlete's current physiological state. Because the sub-threshold transition to *Progress* is strictly constrained by the safety guardrails, a recommendation of *Maintain* does not necessarily imply a physical deficit; rather, it often represents a conservative, precautionary ceiling triggered by a protective boundary (such as an abrupt workload spike or incomplete data). By framing these outputs strictly as decision-support suggestions for load modification rather than absolute status evaluations, the system aims to provide clear, actionable guidance for the coaching team while preventing cognitive misalignment regarding the athletes' actual physiological readiness.

The predictive model therefore estimates risk, whereas the full system translates risk and contextual safeguards into a staff-facing recommendation. These outputs are intended to structure practitioner review rather than automate coaching decisions. The final player-day card presents the risk estimate, key explanatory cues, and the proposed action tier.

Regarding timing, a player-day record for day t is completed after that day's session, so that same-day GPS-derived external load and session duration are available. The model then estimates the probability of elevated soreness on the morning of day $t + 1$, and the resulting recommendation concerns the planned load for the *next* session (day $t + 1$) rather than the already-completed session on day t . Any recommendation is provisional and is intended to be updated once the athlete's morning wellness on day $t + 1$ has been recorded.

4. Data and Methods

4.1. Study Context, Sample, and Data Sources

Data were collected from Lech Poznań UAM, a professional women's football team competing in the top tier of the Polish national league, across two consecutive competitive seasons (2024/25 and 2025/26). All records were integrated at the player-day level, with athlete identity and calendar date serving as primary keys, aligning morning wellness, same-day external load, and daily internal load into a single analytical unit, consistent with the player-day integration layer described in Section 3.

The integrated dataset comprised 17,703 player-day observations across both seasons (8395 in 2024/25 and 9308 in 2025/26). Morning wellness questionnaires were available for 70% of player-days, GPS-derived external-load data for 26%, and the daily wellness-duration internal-load proxy for 22%. The lower GPS coverage relative to wellness reflects that wellness self-reports were collected daily for the full squad, whereas GPS data were recorded only for outfield players during on-pitch sessions. When multiple sessions occurred on the same calendar day, their GPS records were aggregated into a single daily observation in order to preserve consistency with the morning-to-next-morning fatigue cycle. Days without verified formal training exposure were retained as rest days

rather than discarded, ensuring that periods of reduced load remained represented in the temporal record.

Three data streams were combined. Morning wellness questionnaires captured self-reported fatigue, soreness, sleep quality, mood, and menstrual cycle context on a 5-point ordinal scale. External load was obtained from 10 Hz GPS devices and comprised total distance, high-speed running (HSR), sprint distance, player load, and acceleration–deceleration efforts. A proxy for internal load was constructed as a daily wellness-duration product by combining the morning wellness composite (fatigue, sleep, and mood; excluding soreness) with GPS-recorded session duration. This measure is referred to throughout as the wellness-duration internal-load proxy and is not considered equivalent to conventional post-session sRPE, as detailed in Section 4.2. Each player-day was further assigned a position within the training microcycle relative to the nearest official match, expressed in standard football notation (e.g., MD, MD-1, MD-2, MD+1), where negative values denote days preceding a match and positive values denote days following it. This labeling preserved the weekly tactical and recovery context of training and identified the days most relevant to load-adjustment decisions.

Because the predictive model required the internal-load proxy (which in turn required both wellness and GPS session duration), the primary modeling sample comprised player-days with a complete source set (the “full-source stratum”). The wellness-duration proxy was available on 22% of player-days (approximately 3900 observations); after retaining only player-days with a valid next-day soreness label and excluding the first 28 days of each athlete’s series within each season to stabilize rolling indicators, the full-source stratum comprised 3386 player-days from 24 players, containing 350 elevated-soreness events (10.3% prevalence). Events were present in both seasons, with the majority in 2025/26 (134 events in 2024/25 and 216 in 2025/26, reflecting higher GPS coverage in the second season). With 19 candidate predictors, this corresponded to 18.4 events per variable, comfortably above the conventional minimum for regularized logistic regression and providing a reasonable event count for a regularized, low-dimensional model compared with a single-season analysis. Single- and dual-source variants were additionally evaluated on broader strata in the source-ablation analysis (Section 5.1). Across both seasons, 31 unique players were observed, including 26 members of the 2025/26 squad. Of these, 24 contributed complete-source observations to the primary modeling sample, while the per-player LOPO summary was restricted to the 20 players with at least two elevated-soreness events.

Restricting the primary analysis to the complete-source stratum introduced systematic sample selection. The 3386 modeled player-days disproportionately represent structured on-pitch training sessions and match days, during which GPS monitoring was routinely available. In contrast, the excluded observations more frequently represent scheduled rest days, off-pitch recovery, gym-only sessions, injury-rehabilitation periods, travel, and days with incomplete monitoring. The analyzed sample therefore differs from the full set of 17,703 player-days not only in data completeness but also in training context, physical exposure, and the likely prevalence of outcomes.

This selection is consistent with the intended use of the system for load-management decisions during active on-pitch preparation. Nevertheless, it limits generalizability. The fitted relationships describe fully monitored, predominantly active training states and may not apply to rest, rehabilitation, gym-based, or otherwise unmonitored days. Consequently, the reported findings should be interpreted as conditional on the complete-source stratum rather than as representative of a typical player-day across the full competitive cycle.

Table 2 summarizes the study context, data-source coverage, and final modeling sample. The modeled sample showed clear microcycle stratification: elevated next-morning

soreness was most prevalent on match days and least prevalent on MD-1, supporting the inclusion of microcycle position as a contextual predictor. Detailed microcycle-specific sample sizes and outcome prevalence are provided in Appendix A Table A1.

Table 2. Study context, data-source coverage, and final modeling sample across two competitive seasons.

Category	Metric	Value
<i>Study context</i>		
	Players in the 2025/26 squad	26
	Unique players across both seasons	31
	Monitoring period	July 2024–June 2026
<i>Data-source coverage</i>		
	Player-day observations	17,703
	2024/25 season	8395
	2025/26 season	9308
	Wellness availability	70% of player-days
	GPS availability	26% of player-days
	Internal-load proxy availability	22% of player-days
<i>Final modeling sample: full-source stratum</i>		
	Player-day observations	3386
	Players represented	24
	Elevated next-morning soreness cases (≥ 4)	350 (10.3%)
	2024/25 season	134 events
	2025/26 season	216 events
	Events per predictor	18.4

Note. The full-source stratum includes player-days with wellness, GPS-derived external-load data, the internal-load proxy, and a valid next-morning soreness outcome. Events per predictor were calculated for 19 candidate predictors. Detailed microcycle-specific sample sizes and outcome prevalence are presented in Appendix A Table A1.

4.2. Preprocessing, Feature Construction, and Prediction Target

Following integration, the data were cleaned by harmonizing athlete identifiers—reconciling the “Surname Forename” convention of the wellness export with the “Forename Surname” convention of the GPS export, including diminutive forms, through an explicit name-mapping dictionary—aligning timestamps, and removing duplicate records. Rolling descriptors were computed strictly within each (player, season) block to prevent leakage across the off-season boundary. Days without verified training exposure were retained as rest days, with load-related variables set to zero only for confirmed no-session days. Missing values arising from non-response or unavailable devices were not recoded to zero; instead, source-availability indicators were added, and missing entries were imputed within the modeling pipeline using training-set medians for numerical predictors and the most frequent category for categorical predictors. Performing imputation within the cross-validation pipeline, rather than beforehand, prevented validation-fold information from leaking into the training process. These provisions handled item-level gaps in an otherwise available source (for example, a single missing wellness item on a day with valid GPS data). Player-days lacking an entire required source were not imputed because the predictive model required wellness, GPS, and the internal-load proxy; such player-days fell outside the full-source stratum and received no prediction rather than an imputed one. Consequently, within the analyzed full-source sample, the source-availability indicators and the “incomplete monitoring” guardrail (Section 5.1) refer to residual item-level gaps; whole-source absence was handled upstream by the stratum definition, and the corresponding deployment-time fallback for GPS-missing days is discussed in Section 6.2.

The daily internal-load proxy was constructed to approximate a session rating of perceived exertion (sRPE) while remaining aligned with the player-day record. Three morning

wellness dimensions—fatigue, sleep, and mood—were rescaled to a Borg-like 0–10 range and combined into a single perceived-intensity composite; current soreness was deliberately excluded from this composite so that the internal-load features did not reintroduce same-day soreness into the no-current-soreness baselines used in the persistence-versus-learning analysis (Section 5.2). The composite was then multiplied by daily GPS session duration to yield an sRPE-style internal-load value [1,30]. We emphasize that this is a wellness-duration proxy rather than a true sRPE: it substitutes a morning wellness composite for the conventional post-session category-ratio-10 (CR-10) rating, which was not available in this dataset. The proxy therefore provides a daily internal-load signal in the absence of a traditional post-session CR-10 report, a limitation revisited in Section 6.2. Throughout, references to the “internal-load proxy” or “sRPE-style proxy” denote this wellness-duration construct.

To capture temporal workload dynamics, the feature set combined raw daily measures with engineered descriptors reflecting short- and medium-term accumulation. The final classifier relied primarily on session-based rolling descriptors: acute load computed over the previous four sessions, chronic load over the previous twelve sessions, a session-based Time-Dependent Acute:Chronic Workload Ratio (TD-ACWR), and short-term relative change. To accommodate irregular training schedules, the Acute:Chronic Ratio was defined over discrete training events rather than fixed calendar windows. Whereas calendar-based formulations (e.g., a 7:28-day window) are sensitive to gaps introduced by fixture congestion or seasonal breaks, the TD-ACWR computes the ratio over a fixed sequence of sessions—the four most recent sessions for the acute window and the preceding twelve for the chronic window—thereby preserving sensitivity to the actual ordering of physical stress. Calendar-window descriptors (7-day acute load, 28-day chronic load, daily change, monotony, and strain) were computed and available as an alternative operationalization [5]; the primary analysis used the session-based descriptors, and calendar-window results are not reported in this proof-of-concept. Session-based ratios were preferred for the primary analysis because they afforded substantially greater coverage in an intermittently monitored applied setting, whereas stricter calendar-window definitions would have introduced excessive missingness. We note that ratio-based workload indices are themselves subject to known methodological criticism; the rationale for retaining a session-based formulation as one interpretable feature among several, rather than as a standalone decision rule, is discussed in Section 2.2.

The prediction target was next-morning muscle soreness, recorded on a five-point ordinal wellness scale. For the primary analysis, this was dichotomized into an operational binary target denoting elevated soreness (≥ 4) on the next calendar day, a formulation that maps directly onto early-warning decision support and daily load adjustment. Formally, let $S_{i,t}$ denote the self-reported soreness score for player i on the morning of day t . The binary prediction outcome was defined as

$$Y_{i,t+1} = \mathbb{I}(S_{i,t+1} \geq 4). \quad (3)$$

Thus, $Y_{i,t+1} = 1$ denotes elevated next-morning soreness, whereas $Y_{i,t+1} = 0$ denotes a soreness score below 4.

Because the underlying construct is ordinal, the full 1–5 soreness scale was retained for a planned secondary robustness check; the present proof-of-concept report only the binary early-warning formulation, and the ordinal analysis is deferred to future reporting. The 19 predictors are presented in two tables for readability. Table 3 summarizes the wellness and contextual variables, whereas Table 1 presents the internal- and external-load variables and their window definitions.

Table 3. Wellness and contextual predictors included in the player-day model.

Predictor	Definition	Time Reference
<i>Morning wellness variables</i>		
soreness_today	Self-reported muscle soreness measured on a 1–5 ordinal scale.	Current morning
fatigue_today	Self-reported fatigue measured on a 1–5 ordinal scale.	Current morning
sleep_today	Self-reported sleep quality measured on a 1–5 ordinal scale.	Current morning
mood_today	Self-reported mood measured on a 1–5 ordinal scale.	Current morning
<i>Session-context variables</i>		
session_duration	Total duration of GPS-recorded sessions on the current day, in minutes.	Current day
microcycle_offset	Signed position relative to the nearest match (e.g., MD = 0, MD-1 = −1, MD+1 = +1).	Current day
is_match_day	Binary indicator identifying a match day.	Current day

Note. Wellness variables were recorded before training. The predicted outcome was elevated muscle soreness the following morning. Internal- and external-load predictors are presented separately in Table 1.

4.3. Models and Temporal Evaluation Protocol

For the primary classification task, a regularized Logistic Regression (LR) model was compared with a histogram-based Gradient Boosting (GB) model. The model comparison was deliberately limited to logistic regression and Gradient Boosting rather than designed as an exhaustive benchmarking study. Logistic regression was selected as the primary model because it provides a comparatively transparent representation of associations with predictors, supports probability calibration, and can be implemented with limited computational and operational complexity. Gradient Boosting was selected as the sole comparator because it is a well-established, conceptually distinct nonlinear approach for tabular data. Unlike logistic regression, it can capture nonlinear relationships, interactions, and threshold effects without requiring their explicit specification. It therefore provided a focused test of whether additional nonlinear flexibility produced a meaningful predictive advantage over the simpler model.

Other commonly reported algorithms, including random forests, support vector machines, neural networks, and ensemble methods, were not evaluated because the objective of this proof-of-concept study was not to identify the best-performing classifier from a broad candidate set. Extensive algorithm selection would also increase the risk of model-selection optimism, given the moderate number of complete-source observations and the elevated number of elevated-soreness events. The comparison was instead designed to determine whether a representative nonlinear model clearly outperformed logistic regression under the same temporal validation framework. Because Gradient Boosting did not show a consistent advantage, logistic regression was retained and subsequently connected to the recommendation layer, which mapped its calibrated probabilities onto the *Reduce*, *Maintain*, and *Progress* actions using the fixed review threshold and operational guardrails defined in Section 3. Broader comparisons involving additional algorithms should be conducted in future studies using larger and externally collected datasets.

For the logistic regression model, the conditional probability of elevated next-morning soreness was defined as

$$p_{i,t} = \mathbb{P}(Y_{i,t+1} = 1 \mid \mathbf{x}_{i,t}), \quad (4)$$

with

$$\text{logit}(p_{i,t}) = \log\left(\frac{p_{i,t}}{1 - p_{i,t}}\right) = \beta_0 + \mathbf{x}_{i,t}^T \boldsymbol{\beta}, \quad (5)$$

where $\mathbf{x}_{i,t}$ is the vector of predictors available for player i on day t , β_0 is the intercept, and $\boldsymbol{\beta}$ is the vector of regularized regression coefficients. The resulting probability $p_{i,t}$ was subsequently calibrated within the training pipeline and used by the recommendation layer.

To preserve temporal validity, preprocessing, encoding, scaling, imputation, model fitting, and probability calibration were encapsulated within unified pipelines. The system

was evaluated under a multi-pronged validation strategy designed to address the distinct threats to validity in player-monitoring data:

1. Rolling-origin temporal cross-validation (16 monthly folds across both seasons): Each fold was trained on all prior months and evaluated on the subsequent month. This provides a statistically meaningful estimate of fold-to-fold variability, unlike a two-fold split.
2. Leave-one-player-out (LOPO) cross-validation: Each player was held out in turn, so that no within-player baseline could be exploited at evaluation. This directly addresses the concern that performance might reflect player-specific reporting tendencies rather than a transferable signal.
3. Between-season validation: Training on 2024/25 and testing on the entirely unseen 2025/26 season, providing a deployment-like estimate of generalization to a new competitive period and partially renewed squad.
4. Baseline ladder with isolation of soreness persistence (Section 5.2), quantifying how much performance reflects learning beyond the autocorrelation of soreness.

Predictive performance was assessed using ROC-AUC, Average Precision (AP), and the Brier score, capturing discrimination, ranking quality under class imbalance, and calibration, respectively. The operational review threshold was selected within the training data by maximizing F1 and then applied once to the held-out data, rather than being tuned on the evaluation set. Standardized logistic regression coefficients were used to describe model-level associations; their comprehensibility and usefulness to end users were not evaluated.

H1 was evaluated across the three validation schemes, whereas H2 was evaluated using source-ablation analysis, with ROC-AUC as the primary measure of discrimination.

4.4. Ethics, Consent, and Data Protection

All participants were members of a professional football squad, and the monitoring data analyzed in this study (wellness self-reports and GPS-derived load) were collected by the club as part of routine performance management, while the wellness-duration internal-load proxy was derived analytically. The study was conducted in accordance with the principles of the Declaration of Helsinki.

No approval from a university research-ethics committee was sought or required for this work. The data were not generated through any research intervention: they originate from routine, contractually based performance monitoring carried out by the club in the ordinary course of professional team operation, and were analyzed retrospectively in pseudonymized form. The club acts as the data controller and is responsible for data collection, for informing the players about the monitoring program, and for pseudonymizing the records before they are made available for analysis. As part of their professional engagement with the club, players are informed about the routine performance-monitoring program and consent to it under their contractual and club-regulatory arrangements. The university research team acted solely as a data processor on already-pseudonymized data and at no point had access to directly identifying information; analytical outputs were returned only to the club for internal performance-management use.

All data were pseudonymized by the club prior to being shared for analysis: athlete names were replaced with non-identifying codes, and no individual is identifiable in any reported result or figure. Menstrual-cycle context, collected by the club, constitutes a special category of personal data under the EU General Data Protection Regulation (GDPR); it was handled solely in pseudonymized form for the club's monitoring purposes and was *not* used as input to the predictive model. Data were stored on access-controlled club

infrastructure and processed under the lawful basis of legitimate interest in the context of routine professional performance management.

5. Evaluation Results

5.1. Integration Adds Value

The internal consistency of the integrated player-day representation was first supported by within-player associations between the daily internal-load proxy and mechanically driven GPS metrics. As expected, the strongest median Spearman correlation was observed for session duration ($r_s = 0.812$), followed by total distance ($r_s = 0.545$) and player load ($r_s = 0.525$). Moderate associations were also observed for acceleration efforts ($r_s = 0.392$), high-intensity distance ($r_s = 0.378$), and deceleration efforts ($r_s = 0.368$). These relationships support the coherence of the player-day integration layer, although they do not by themselves establish whether integration improves decision support. Because session duration is a multiplicative component of the wellness-duration proxy, its strong association with session duration is expected by construction and should not be interpreted as independent evidence of construct validity. Associations with total distance, player load, and other GPS-derived measures provide limited convergent evidence within the studied environment, but they do not establish equivalence with conventional post-session sRPE.

The principal assessment of integration effects was therefore the source-ablation analysis presented in Table 4. Using rolling-origin temporal validation across 16 folds, the full multi-source system delivered the strongest discrimination overall (ROC-AUC = 0.759 ± 0.089) but was not consistently dominant across all performance metrics. The wellness + GPS variant achieved the highest Average Precision (0.371 ± 0.131), while the wellness-only model obtained the lowest Brier score (0.0794). Because the Brier score combines calibration and refinement, and because the single- and dual-source variants were evaluated on their respective broader strata with differing outcome prevalence (see caption), this indicates the lowest overall probabilistic error rather than demonstrably superior calibration; between-variant comparisons of the Brier score in particular should be read with that caveat. All single-source variants performed worse in terms of discrimination, with GPS-only being the weakest (ROC-AUC = 0.586 ± 0.101), while wellness-only (0.712 ± 0.118) and proxy-only (the wellness-duration proxy; 0.721 ± 0.087) showed intermediate performance. Because these estimates are based on 16 folds, the reported standard deviations reflect meaningful temporal variability. The overlap between the strongest configurations suggests that while full integration improves discrimination, performance differences across top-performing variants are modest and metric-dependent.

Taken together, these findings suggest that integrating multiple monitoring streams improves the overall predictive signal compared with using any single source alone, despite relatively modest differences among the strongest model variants.

5.2. Persistence Versus Learning: Isolating the Contribution of Current Soreness

Because next-morning soreness is autocorrelated with current soreness, and current soreness is among the model's inputs, a central question is how much predictive performance reflects genuine learning rather than the mechanical persistence of soreness. We addressed this directly with a baseline ladder evaluated under leave-one-player-out cross-validation, so that no within-player baseline was available at evaluation (Figure 2):

- L0—persistence rule: Flags elevated soreness tomorrow whenever soreness today ≥ 4 . ROC-AUC = 0.721.
- L1—two-feature model: LR on current soreness and current fatigue only. ROC-AUC = 0.707, AP = 0.329.
- L2—full feature set excluding current soreness: ROC-AUC = 0.738, AP = 0.235.

- L3—full model: ROC-AUC = 0.782, AP = 0.319.

Two findings are central. First, a model that excludes current soreness entirely (L2) still outperforms the persistence rule (ROC-AUC 0.738 vs. 0.721), suggesting that recent external and internal load and current readiness carry predictive signal independent of soreness persistence. Second, the isolated contribution of current soreness is reliable but modest: adding it (L3 vs. L2) improves discrimination by $\Delta\text{ROC-AUC} = +0.044$ (95% bootstrap CI [+0.027, +0.059]) and ranking by $\Delta\text{AP} = +0.084$ ([+0.057, +0.113]). The confidence intervals exclude zero, so current soreness is genuinely informative, but it accounts for roughly four ROC-AUC points, not the bulk of the model's performance.

Table 4. Performance of source-ablation variants under rolling-origin temporal validation (16 monthly folds, logistic regression). Dual-source variants cover all three pairwise combinations of the wellness, GPS, and internal-load (sRPE-style) streams. Each variant was evaluated on its own maximal available stratum, so sample sizes and outcome prevalence differ across rows; between-row comparisons (especially of the Brier score, which is prevalence-dependent) are therefore approximate rather than a matched-sample ablation.

Configuration	ROC-AUC $\uparrow$	Avg. Precision $\uparrow$	Brier Score $\downarrow$
Wellness-only	0.712 $\pm$ 0.118	0.327 $\pm$ 0.141	0.0794
GPS-only	0.586 $\pm$ 0.101	0.180 $\pm$ 0.107	0.0897
Proxy-only	0.721 $\pm$ 0.087	0.239 $\pm$ 0.112	0.0862
GPS + Proxy	0.704 $\pm$ 0.100	0.229 $\pm$ 0.115	0.0885
Wellness + GPS	0.747 $\pm$ 0.101	0.371 $\pm$ 0.131	0.0802
Wellness + Proxy	0.727 $\pm$ 0.093	0.339 $\pm$ 0.168	0.0809
Full integration	0.759 $\pm$ 0.089	0.350 $\pm$ 0.142	0.0805

Note. Bold text indicates the best performance for each metric. Upward arrows ($\uparrow$) denote that higher values are better; downward arrows ($\downarrow$) denote that lower values are better.

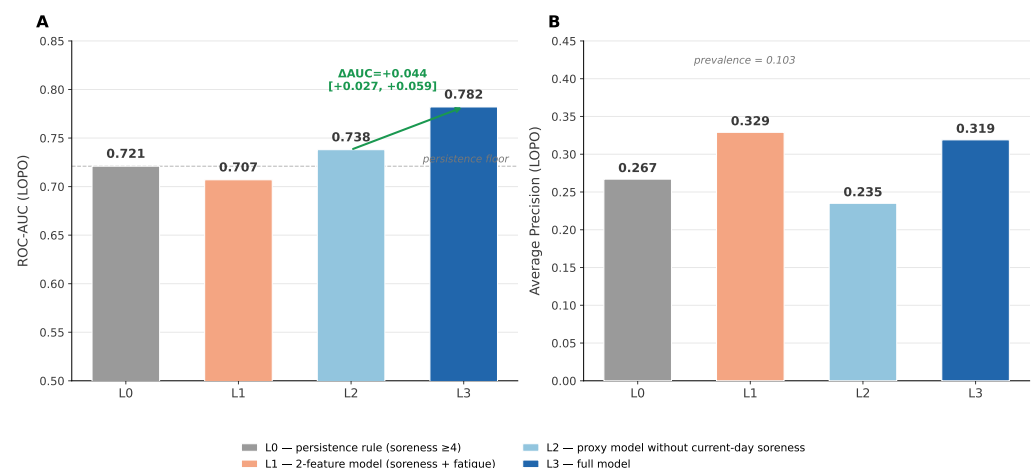


Figure 2. Baseline ladder under leave-one-player-out cross-validation, isolating the contribution of current soreness. (A, left) Discrimination (ROC-AUC) and (B, right) ranking quality (Average Precision) across increasing model complexity, from a trivial persistence rule (L0) to the full model (L3). The model excluding current soreness (L2) already exceeds the persistence floor; the isolated contribution of current soreness (L3 vs. L2) is reliable but modest.

The contrast between ROC-AUC and Average Precision across the baseline ladder is instructive in itself. The two-feature model (L1) attains higher Average Precision than the no-soreness model (L2), despite a lower ROC-AUC, reflecting that current soreness concentrates precision on the highest-risk cases (where it most closely resembles the outcome), whereas the broader load-based feature set improves overall ranking across the full

risk range. Reporting both metrics, rather than ROC-AUC alone, is therefore essential at this prevalence.

This analysis reframes the headline result: rather than claiming a single high AUC, we show transparently that the system learns a real, if moderate, signal beyond soreness persistence—a more defensible and operationally honest basis for decision support.

5.3. Temporal Stability and Between-Season Generalization

Temporal validity was assessed under rolling-origin cross-validation across 16 monthly folds spanning both seasons (Figure 3A). The final classifier retained usable performance throughout, with ROC-AUC = 0.759 ± 0.089 , Average Precision = 0.350 ± 0.142 , and Brier score = 0.080 ± 0.023 . Logistic regression outperformed gradient boosting in the majority of folds (GB: ROC-AUC = 0.746 ± 0.085), consistent with the view that it did not provide a clear performance advantage at this sample size and prevalence. Because variability is now estimated from 16 folds, the standard deviations are meaningful, unlike a two-fold estimate.

As a season-holdout test, the model was trained on the entire 2024/25 season and evaluated on the unseen 2025/26 season (Figure 3B). This between-season test yielded ROC-AUC = 0.801 (Average Precision = 0.401, Brier = 0.077). Consistent with the fold-to-fold variability of the rolling-origin analysis (SD = 0.089), individual monthly folds contained few elevated-soreness events and therefore yielded noisier month-level estimates; evaluated instead on a full season with 216 events, the model retained useful discrimination in a new competitive season within the same club and a partially renewed squad. A model excluding current soreness also retained discrimination in the held-out season within the same club (ROC-AUC = 0.770), reinforcing that the cross-season signal does not depend solely on soreness persistence.

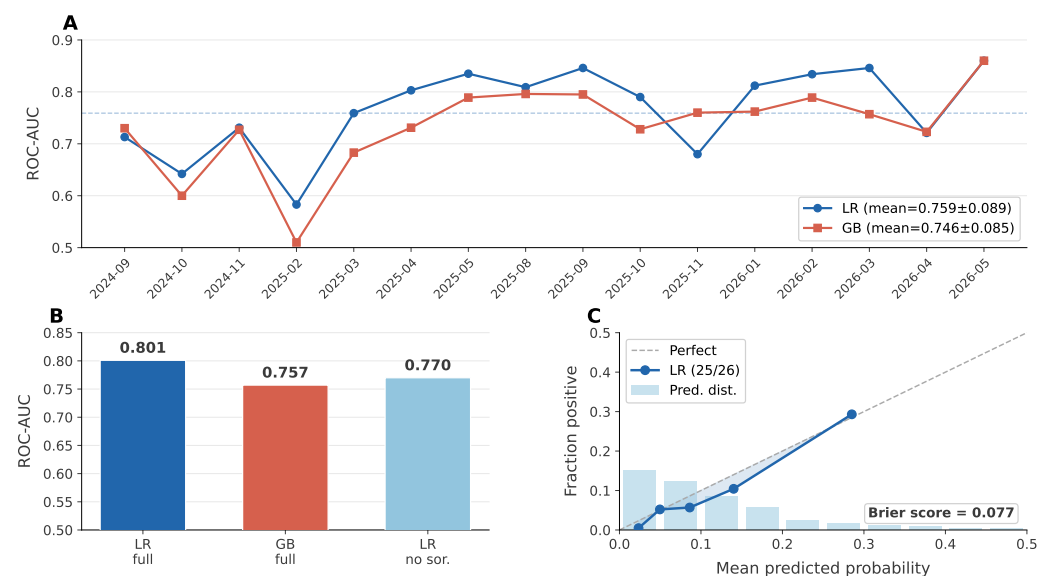


Figure 3. Temporal and cross-season validation. (A, top) Rolling-origin ROC-AUC across 16 monthly folds for logistic regression and gradient boosting. (B, bottom left) Cross-season validation, with models trained on the 2024/25 season and evaluated on the unseen 2025/26 season, for the full logistic regression model, gradient boosting, and a logistic regression model excluding current-day soreness. (C, bottom right) Calibration curve and predicted-probability distribution for the logistic regression model evaluated on the 2025/26 season; the corresponding Brier score is reported within the panel.

Leave-one-player-out cross-validation provided a complementary view in which no within-player baseline could be exploited: pooled ROC-AUC = 0.782, with a per-player

median of 0.708 (IQR: 0.678–0.813) across the 20 players with at least two events. The combination of rolling-origin, between-season, and leave-one-player-out results supports temporal stability and cross-player transfer within the studied club. However, these analyses do not constitute external validation and cannot establish applicability to elite women's football more broadly.

5.4. Operational Usefulness and Empirical Evaluation of the Three-State Recommendation Logic

For operational use, discrimination alone is insufficient; the estimated probabilities must also be interpretable at the threshold level, as calibration is essential whenever model outputs are translated into decisions [29]. Calibration on the unseen 2025/26 season (Figure 3C) showed acceptable agreement between predicted and observed risk across the lower-to-middle probability range (Brier = 0.077), with sparse data in the highest bins, supporting the translation of model output into transparent action tiers rather than reliance on probability values alone.

Numerically, calibration error on the unseen 2025/26 season was low: expected calibration error (ECE) = 0.036 and maximum calibration error (MCE) = 0.089 across ten quantile bins. We report quantile-binned values because equal-width binning places very few observations in the highest-probability bins, inflating the maximum-error estimate.

The review threshold was selected once during the 2024/25 training season by maximizing F1, yielding $p \geq 0.20$. We then assessed this threshold under two complementary schemes, which we kept explicitly separate.

F1 was adopted as the threshold-selection criterion because the operational objective was to balance missed elevated-soreness cases and excessive false alerts. However, false-positive and false-negative classifications have different practical consequences. A false positive may lead staff to conduct an unnecessary review or apply an overly conservative load adjustment to a player who would not have reported elevated soreness the following morning. Repeated false alerts could increase staff workload, reduce training exposure, disrupt session planning, and weaken confidence in the system.

A false negative occurs when a player who subsequently reports elevated soreness is not assigned to the high-priority *Reduce* state. In practice, this may result in the planned load being maintained or progressed without additional review, potentially overlooking a short-term recovery concern. Nevertheless, because the predicted outcome is self-reported soreness rather than injury or medical deterioration, a false negative should not be interpreted as a missed diagnosis of injury.

The relative costs of these errors depend on the club's training objectives, staffing capacity, competition schedule, and tolerance for missed alerts. Maximizing F1 implicitly assigns similar importance to precision and recall and therefore represents one operational compromise rather than a universally optimal decision rule. Clubs prioritizing the avoidance of missed elevated-soreness cases may select a lower threshold to increase sensitivity, whereas clubs seeking to limit unnecessary interventions may use a higher threshold. Human review and operational guardrails are intended to mitigate the consequences of both types of classification errors rather than to allow the model output to automatically determine training decisions.

Scheme 1—between-season threshold evaluation. Applying the fixed threshold to the held-out 2025/26 season produced precision = 0.350, recall = 0.542, and an alert burden of 15.8 player-days per 100. This estimates the operating point a practitioner would obtain by fixing the threshold for one season and applying it to the next.

Scheme 2—out-of-fold evaluation of the three-state layer. To characterize the full three-state output across both seasons rather than a single test block, we applied the same fixed threshold ($p \geq 0.20$) to out-of-fold predictions generated under rolling-origin validation. Because rolling-origin validation could not produce predictions for the earliest months (due to the lack of prior training data), this analysis covered 3127 of the 3386 full-source player-days; the 259 excluded observations fell in July and August 2024, before sufficient training history had accumulated. The decision-support layer returns one of three actions rather than a single probability. *Reduce* was returned when $p \geq 0.20$; below threshold, *Maintain* was returned when an operational guardrail was active (current soreness ≥ 4 , current fatigue ≥ 4 , internal-load-proxy TD-ACWR ($srpe_td_acwr$) ≥ 1.5 , or incomplete monitoring), and *Progress* otherwise. The three states separated outcomes monotonically: player-days labeled *Progress* had an observed next-morning soreness rate of 0.054 (95% Wilson CI [0.046, 0.064]), compared with 0.217 [0.152, 0.299] for *Maintain* and 0.326 [0.285, 0.369] for *Reduce* (Figure 4 and Table 5). Because *Maintain* and *Progress* are the two sub-threshold states distinguished only by the guardrails, we tested this specific contrast directly: a chi-square test on the 2×2 table of state (*Progress* vs. *Maintain*) by next-morning outcome was statistically significant ($\chi^2 = 49.8$, $df = 1$, $p < 0.001$), indicating that *Progress* days had a lower observed outcome rate than *Maintain* days rather than reflecting an arbitrary split. Among sub-threshold cases, guardrails were activated in 4.5% (driven primarily by elevated current fatigue and soreness), capping these at *Maintain* rather than *Progress*. This demonstrates that the three-state logic provides useful risk stratification but should be interpreted as an early-warning and review-prioritization mechanism rather than as an automated prescription for load adjustment. The probability-to-action mapping, therefore, functions as decision support rather than as an automatic prescription rule.

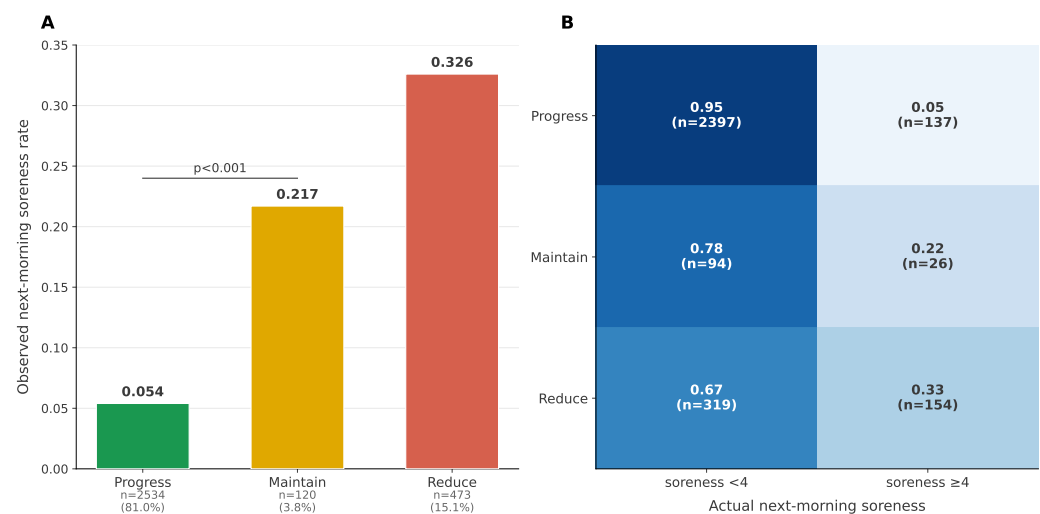


Figure 4. Empirical evaluation of the three-state recommendation layer on out-of-fold predictions. **(A)** Observed next-morning soreness rate by recommendation state, with the *Progress* versus *Maintain* contrast tested by χ^2 . **(B)** Row-normalized confusion of the DSS state against the actual next-morning outcome.

Table 6 makes this trade-off explicit by reporting precision, recall, F1, and alert burden across alternative thresholds under both validation schemes.

Table 5. Three-state operational recommendation logic for the final calibrated classifier, with empirical outcomes from out-of-fold predictions (Scheme 2; $n = 3127$ player-days with available rolling-origin predictions).

Value	Rule	Soreness Rate	95% CI	n (%)	Operational Use
Reduce	$p \geq 0.20$	0.326	[0.285, 0.369]	473 (15.1%)	High-priority manual review before training; consider load reduction if supported by context.
Maintain	$p < 0.20$ and active guardrail	0.217	[0.152, 0.299]	120 (3.8%)	Keep the planned session, or apply a conservative adjustment; default to the output if ambiguous.
Progress	$p < 0.20$ and no guardrail	0.054	[0.046, 0.064]	2534 (81.0%)	Allow planned progression only after routine staff confirmation.

Note. Bold text in the first column highlights the operational recommendation states. The model outputs one calibrated probability p of elevated next-morning soreness. The DSS then maps it, via post-processing, to one of three operational values. The threshold $p \geq 0.20$ was selected by maximizing F1 inside the 2024/25 training season. Observed soreness rates are reported with 95% Wilson score confidence intervals, which remain valid for small samples and proportions near the bounds (relevant given the small *Maintain* group, $n = 120$). Shares are reported over the 3127 player-days for which rolling-origin out-of-fold predictions were available (Scheme 2); the remaining 259 of the 3386 full-source player-days fall in the earliest months (July–August 2024) and lack out-of-fold predictions. For the between-season operating point obtained by fixing this threshold on 2024/25 and applying it to 2025/26 (Scheme 1), precision = 0.350 and recall = 0.542. *Maintain* and *Progress* are distinguished by operational guardrails, not by a second probability threshold. The *Progress* versus *Maintain* contrast (a 2×2 chi-square test of these two sub-threshold states by next-morning outcome) is statistically significant ($\chi^2 = 49.8$, $df = 1$, $p < 0.001$).

Table 6. Threshold sensitivity of the operating point. Precision, recall, and alert burden are reported across candidate thresholds under both validation schemes: between-season (training 2024/25, test 2025/26; 216 elevated-soreness events among 2112 player-days) and out-of-fold rolling-origin (both seasons; 317 events among 3127 player-days). The selected threshold ($p \geq 0.20$) maximized F1 in the 2024/25 training season. Alert burden is the number of flagged player-days per 100.

Threshold	Precision $\uparrow$	Recall $\uparrow$	F1 $\uparrow$	Alerts/100
<i>Scheme 1—between-season (test 2025/26)</i>				
$p \geq 0.150$	0.254	0.671	0.368	27.0
$p \geq 0.175$	0.285	0.593	0.385	21.3
$p \geq 0.200$	0.350	0.542	0.425	15.8
$p \geq 0.250$	0.441	0.463	0.451	10.7
$p \geq 0.300$	0.510	0.366	0.426	7.3
<i>Scheme 2—out-of-fold (both seasons)</i>				
$p \geq 0.150$	0.259	0.628	0.367	24.6
$p \geq 0.175$	0.296	0.539	0.383	18.5
$p \geq 0.200$	0.326	0.486	0.390	15.1
$p \geq 0.250$	0.384	0.385	0.384	10.2
$p \geq 0.300$	0.455	0.290	0.355	6.5

Note. Bold text in the metric columns indicates the best value for each metric, while bold text in the Threshold column highlights the selected operating threshold ($p \geq 0.20$). Italics denote validation scheme headers. Upward arrows ($\uparrow$) indicate that higher values are better. Precision and recall behave monotonically across thresholds in both validation schemes: precision increases, whereas recall and alert burden decrease. This indicates a stable operating curve. The operating point can therefore be adjusted to satisfy an institution-specific constraint. For example, a higher sensitivity target favors $p \geq 0.150$, whereas a lower sustainable review burden favors $p \geq 0.250$. At the selected threshold ($p \geq 0.20$), the out-of-fold operating point corresponds to the *Reduce* branch reported in Table 5, comprising 473 flagged player-days. The full-source sample contained 350 events among 3386 player-days. Rolling-origin out-of-fold predictions were available for 3127 player-days containing 317 events because the earliest 259 observations, including 33 events, had no preceding training window. The between-season test was restricted to the 2025/26 season and contained 216 events among 2112 player-days.

To assess clinical-style utility beyond discrimination, we performed a decision curve analysis on the out-of-fold predictions (Figure 5). Across the full range of threshold probabilities (0.05–0.50), the decision-support model achieved a higher net benefit than both the “alert-all” and “alert-none” reference strategies. At the operating threshold ($p \geq 0.20$) the model’s net benefit was 0.024, compared with −0.123 for “alert-all” and

0 for “alert-none”, confirming that the system adds decision value rather than merely discriminating.

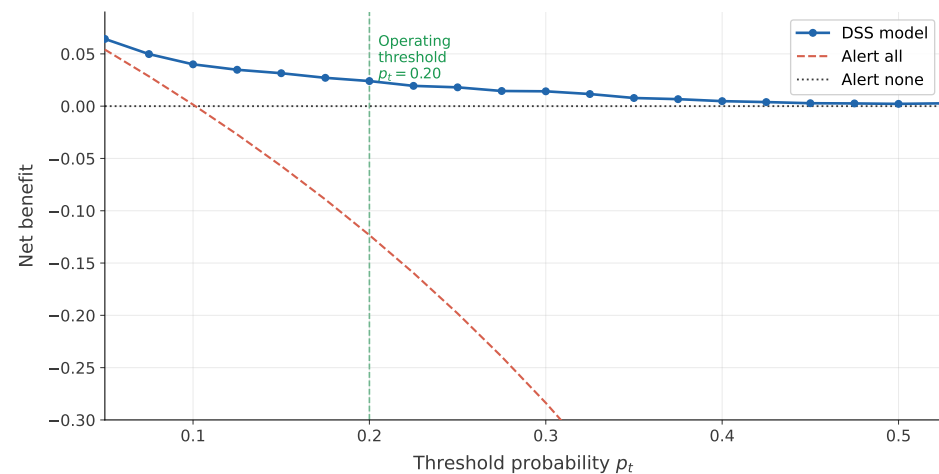


Figure 5. Decision curve analysis on out-of-fold predictions across both seasons. The decision-support model (blue) yields higher net benefit than the “alert-all” (orange) and “alert-none” (dotted) reference strategies across the operational threshold range. The vertical line marks the operating threshold ($p = 0.20$).

5.5. Proposed Human-in-the-Loop Workflow

Figure 6 presents the final player-day output designed to communicate model results to staff. The card foregrounds the calibrated probability of elevated next-morning soreness, the upper review threshold, compact morning-wellness cues, concise session-context indicators, and the resulting branch of the three-state recommendation logic. The card therefore functions as a structured review interface rather than a standalone decision engine. The example shown is a single illustrative player-day (a pre-match MD-3 observation) in which a high predicted risk ($p = 0.76$) coincides with a readiness deficit (fatigue 5/5, sleep 3/5, soreness 5/5) and a substantial recent internal load (internal-load-proxy acute-4 = 541 AU·min), triggering a *Reduce* recommendation with explicit provision for staff override. To avoid revealing the outcome at the time of decision, the observed next-morning soreness is deliberately omitted from the card.

The proposed workflow is supported conceptually by the model coefficients shown in Figure 7. The strongest positive signals were current soreness (+0.63), match-day status (+0.55), the acute four-session total-distance load (+0.46), current fatigue (+0.41), the acute four-session internal-load proxy (+0.34), and session duration (+0.30). On the protective side, the strongest negative associations were a lower session-based duration TD-ACWR (−0.60) and higher chronic distance (−0.51)—consistent with the interpretation that well-conditioned athletes carrying higher chronic load tolerate daily stimulus better. The recommendation layer preserved explicit staff override and applied conservative guardrails before assigning *Progress*. The system is therefore proposed as a human-in-the-loop early-warning mechanism that structures review and load-adjustment decisions while preserving practitioner control.

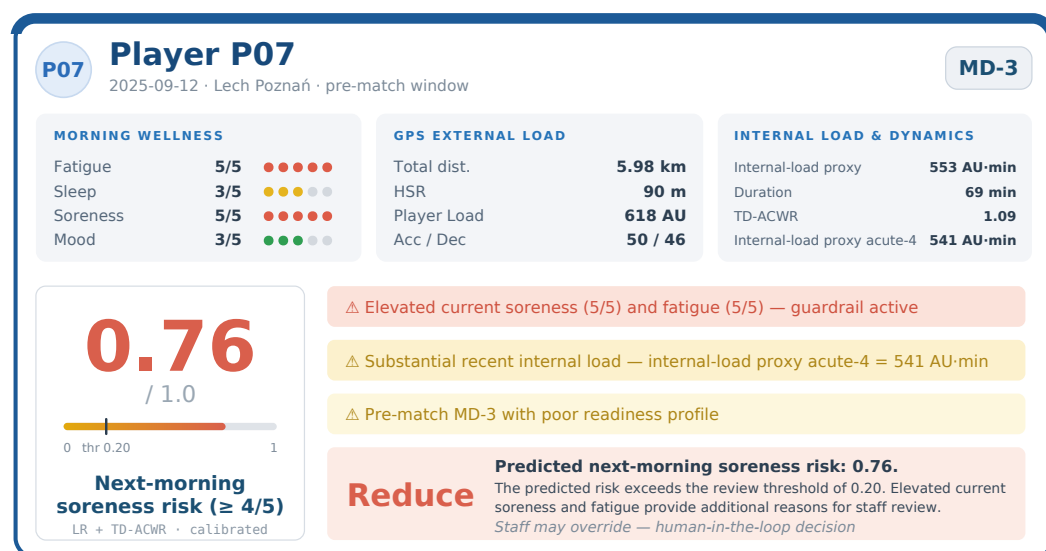


Figure 6. Example anonymized player-day card illustrating the output of the decision-support layer. The interface combines the calibrated probability of elevated next-morning soreness with the fixed review threshold of $p = 0.20$ (left), morning-wellness indicators, GPS-derived external-load measures, and wellness-duration internal-load proxy dynamics (top), automatically generated warning flags (center), and the resulting staff-facing recommendation with a human-in-the-loop override option (bottom). In this illustrative *Reduce* case, the predicted risk of 0.76 exceeds the review threshold. The example is based on retrospective data and is presented for illustration only; the observed next-morning outcome would not be available or displayed at the time of decision-making.

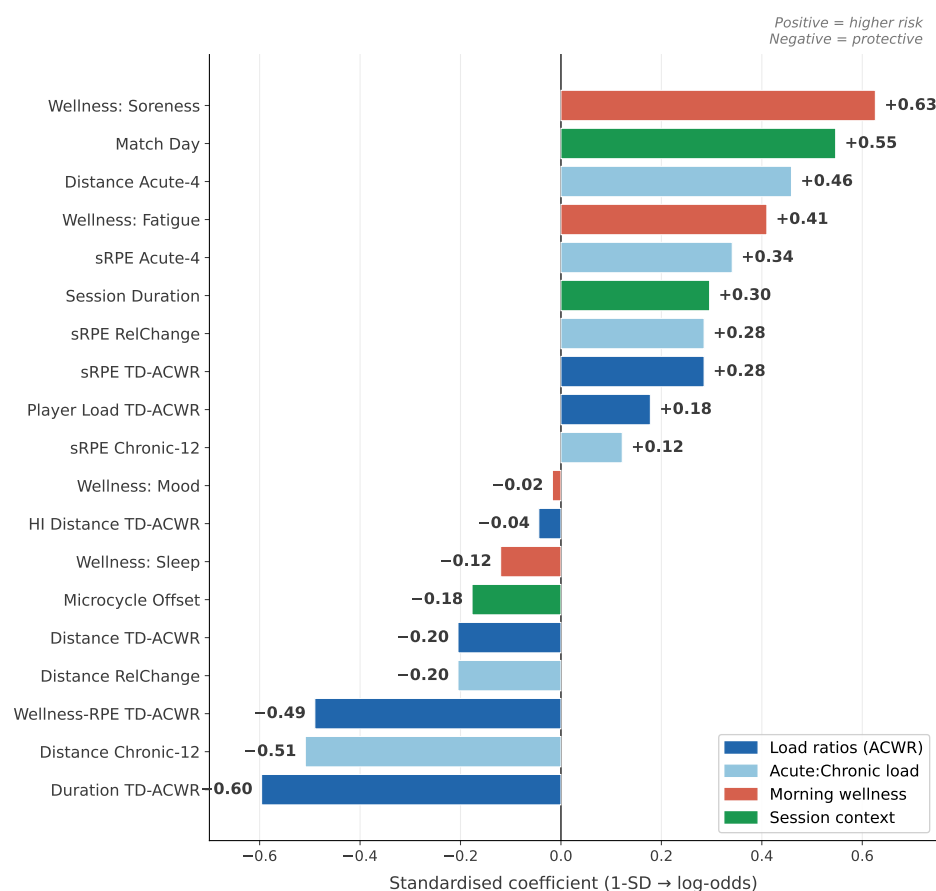


Figure 7. Standardized logistic regression coefficients (full sample). Positive values indicate a higher estimated risk of elevated next-morning soreness; negative values indicate a protective association. A one-standard-deviation change in a predictor corresponds to the plotted change in log-odds.

6. Discussion

6.1. Implications for Decision-Support Information Systems

The study does not propose a new fatigue-prediction method. Its principal contribution is system-level: established wellness measures, GPS-derived variables, workload descriptors, and a comparatively transparent predictive model are integrated into a common player-day representation and connected to an operational recommendation layer. The predictive model should therefore be viewed as one component of the proposed decision-support architecture rather than as the primary methodological innovation.

From an information systems perspective, the contribution lies in integrating heterogeneous monitoring signals into a single daily operational representation and translating them into calibrated, staff-facing recommendations [9,10]. The results provide preliminary evidence that this architecture can support within-club risk stratification and staff review. They should not be interpreted as demonstrating the superiority of a novel predictive algorithm or the universal optimality of the selected model and feature set.

The baseline-ladder analysis provides an important qualification of the role of current soreness. The model excluding current soreness retained competitive discrimination ($\text{ROC-AUC} = 0.738$), whereas adding this variable improved ROC-AUC by 0.044 and Average Precision by 0.084. This suggests that recent workload history and other readiness indicators capture a substantial, though not identical, portion of the information associated with next-morning soreness. Current soreness should therefore be interpreted as an incremental, rather than dominant, predictor: it improves the identification of higher-risk cases, but most of the model's discrimination does not depend on soreness persistence alone. In practice, soreness should be considered alongside longitudinal workload and readiness information rather than used as a standalone decision rule.

The internal-load proxy also requires cautious interpretation. Unlike conventional post-session sRPE, the wellness-duration measure combines pre-session status with session duration and therefore does not directly measure the perceived exertion generated by the completed session. The two measures may diverge when good morning wellness is followed by an unusually demanding session or when poor morning wellness precedes a light session. Because several predictors are derived from this proxy, such measurement differences may affect both estimated associations and predictive performance. Similar results cannot be assumed if conventional post-session sRPE were used; direct comparison of the two operationalizations is therefore required in future studies.

The source-ablation results further indicate that the value of integration depends on the performance objective. Full integration achieved the highest ROC-AUC, whereas wellness + GPS achieved the highest Average Precision and wellness-only produced the lowest Brier score. Thus, no configuration was uniformly superior across discrimination, ranking quality, and calibration. The fully integrated model remains the preferred proof-of-concept configuration because it preserves all three intended monitoring domains—current readiness, external exposure, and proxy-derived internal-load information—and provides the complete information required by the proposed recommendation and guardrail framework. This preference reflects system coverage and operational coherence rather than uniform statistical superiority; alternative configurations may be preferable when data availability or the primary performance objective differs.

An additional practical consideration concerns routine deployment under incomplete monitoring. Although the fully integrated model remains the preferred operational configuration, the source-ablation results suggest that the system would retain useful discrimination when GPS data are unavailable, as the wellness-only model achieved an ROC-AUC of 0.712, whereas GPS-only performance was substantially lower ($\text{ROC-AUC} = 0.586$). This indicates that wellness monitoring constitutes the most informative individual data source

in the present setting, while GPS-derived variables provide complementary predictive value. These conclusions should nevertheless be interpreted with caution because they are based on retrospective source-ablation analyses rather than on prospective evaluation under realistic patterns of missing data.

In operational use, the system translates calibrated risk estimates and guardrails into three staff-facing recommendations: *Reduce*, *Maintain*, or *Progress*. These recommendations are intended to prioritize practitioner review rather than automate load-adjustment decisions.

The observed discrimination suggests potential usefulness for preliminary within-club risk stratification and prioritization of player-days for staff review but not for autonomous load-adjustment decisions. The system is therefore best viewed as an early-warning tool that supports, rather than replaces, practitioner judgment. At the selected threshold, approximately 15–16 player-days per 100 would be flagged, with a precision of approximately 0.33–0.35, meaning that a substantial proportion of alerts would not be followed by elevated next-morning soreness. Such false alerts may be acceptable when review is brief, and the cost of missing a high-risk case is considered greater than the cost of an additional staff assessment. False-positive alerts may increase staff workload and encourage unnecessary caution, whereas false-negative classifications may fail to prioritize a player-day that would otherwise merit closer review. Because the system is intended for review prioritization rather than automated prescription, either type of error should be interpreted in conjunction with practitioner judgment and the broader training context. However, acceptable alert burden will depend on local staffing and risk tolerance. As shown in Table 6, increasing the threshold can reduce the review burden to approximately 10 alerts per 100 player-days, at the cost of lower sensitivity.

6.2. Limitations and Future Work

Several limitations should be acknowledged. First, all data originated from a single elite women's football club. Although the analysis spans two seasons and includes rolling-origin, between-season, and leave-one-player-out validation, these procedures assess temporal stability and cross-player transfer only within the same organizational context and do not constitute external validation. Transferability may be affected by differences in squad characteristics, training methodology, periodization, monitoring systems, reporting practices, match schedules, and staff workflows. The probability threshold and guardrail cutoffs were also developed within this specific organizational and monitoring context. Their apparent utility in separating recommendation states does not establish that the same numerical values would be appropriate in other clubs. External implementation would require local recalibration and prospective validation of the resulting alert burden, sensitivity, and practical consequences for training decisions. Applicability to semi-professional and amateur settings is also uncertain because of differences in training frequency, monitoring adherence, non-football demands, and staff resources. Similarly, transferability to male football populations cannot be assumed, as workload profiles, physical demands, and wellness-reporting patterns may differ. The findings should therefore be interpreted as evidence of feasibility within the studied club rather than as evidence of general applicability to elite women's football. Independent multi-club, multi-level, and sex-specific validation is required, potentially followed by local recalibration or model updating.

Second, model development and validation were restricted to the complete-source stratum. GPS data were available for only 26% of all player-days, while the wellness-duration internal-load proxy was available for 22%. Consequently, the final modeling sample of 3386 observations represented approximately 19.1% of the initial 17,703 player-days. The included observations disproportionately represented structured on-pitch training and

match exposure, whereas rest days, off-pitch recovery, gym sessions, rehabilitation periods, travel, and incompletely monitored days were more frequently excluded.

This selection may have affected the reported model performance because the prevalence of elevated soreness, the distributions of the predictors, and the relationships between workload and soreness may differ between fully monitored active-training days and excluded player-days. Restricting the analysis to a more homogeneous and systematically monitored subset could increase apparent discrimination and calibration by reducing contextual variability. Conversely, concentrating the analysis on demanding training and match states could make prediction more difficult and reduce performance. The direction and magnitude of this effect cannot be established from the available data.

Importantly, rolling-origin, between-season, and leave-one-player-out validation all preserved the same complete-source inclusion mechanism. These procedures therefore assess temporal and cross-player generalization within the selected stratum, but they do not demonstrate that the model would retain the same performance across the full competitive cycle. In particular, ROC-AUC, Average Precision, the Brier score, the selected probability threshold, and the resulting alert burden should all be interpreted as conditional on the fully monitored sample. Differences in outcome prevalence or data-collection practices could alter calibration and require threshold recalibration even if discrimination remained similar.

Future studies should prospectively evaluate the system across a broader spectrum of player-days and formally compare the characteristics and outcome prevalence of included and excluded observations. A multi-tiered architecture should also be considered, in which a reduced model, such as the wellness-only variant, provides a fallback prediction when GPS or proxy data are unavailable. This would allow the system's operational coverage and performance to be evaluated under realistic patterns of incomplete monitoring.

Third, internal load was represented by a wellness-duration proxy rather than a conventional post-session CR-10 sRPE measure. The proxy combines pre-session wellness, which may reflect fatigue accumulated over previous days, with session duration. It may therefore underestimate internal load when an athlete reports good morning wellness before an unexpectedly demanding session and overestimate it when poor morning wellness is followed by a light session. Although its associations with GPS-derived measures support some coherence within the studied training environment, these associations do not establish equivalence with true sRPE. The proxy should therefore be regarded as a pragmatic surrogate requiring direct validation against independent post-session ratings.

Fourth, soreness and fatigue are subjective constructs influenced by psychological, contextual, and reporting factors that were not fully observed. Individual athletes may also interpret the ordinal scale differently, introducing measurement error that leave-one-player-out validation only partly addresses. Moreover, the model predicts a subjective next-morning soreness report rather than injury, medical diagnosis, or an objective recovery marker. Its outputs should therefore be interpreted as early-warning information for staff review, not as evidence of injury risk or a basis for automated clinical or training decisions. As demonstrated by the baseline-ladder analysis, part of the predictive performance also reflects soreness persistence; the additional signal learned from workload and readiness variables is reliable but moderate.

Fifth, interpretability was assessed only at the model level and was not evaluated empirically with end users. The use of regularized logistic regression and standardized coefficients provides structural transparency by indicating the direction and relative magnitude of predictor associations. However, these characteristics do not establish that coaches, medical staff, or strength and conditioning practitioners can correctly understand, trust, and use the model outputs in routine decision-making.

The study did not evaluate whether practitioners could accurately interpret the predicted probabilities, explanatory cues, recommendation states, or operational guardrails. It also did not assess interface usability, practitioner acceptance, adherence to recommendations, the effects of staff overrides, or changes in training decisions. Accordingly, the term “interpretable” in this study refers to the relative transparency of the model architecture and the proposed presentation of its outputs, rather than to empirically demonstrated end-user interpretability. Prospective user-centered studies should assess explanation comprehension, usability, workflow integration, practitioners’ trust, and the system’s effect on actual decisions.

The model comparison was restricted to logistic regression and gradient boosting and should not be interpreted as a comprehensive benchmark of available machine-learning methods. Other algorithms may achieve different or superior predictive performance. The selected comparison reflected the study’s emphasis on transparency, implementation simplicity, and integration into a decision-support workflow rather than on maximizing predictive accuracy.

Finally, the recommendation layer was evaluated retrospectively using out-of-fold predictions. Prospective studies are needed to examine alert burden in routine practice, practitioner responses, recommendation overrides, resulting load modifications, and subsequent athlete outcomes before operational deployment can be considered.

In summary, future work should prioritize independent multi-club validation, direct comparison of the internal-load proxy with conventional post-session sRPE, evaluation under more complete GPS monitoring, and the prospective user-centered assessment outlined above.

7. Conclusions

This paper presented a proof-of-concept interpretable decision-support information system to support staff review of elevated next-morning self-reported soreness in elite women’s football, integrating wellness, GPS-derived external load, and a daily wellness-duration internal load proxy at the player-day level. The predictive component estimated a single calibrated probability of elevated next-morning soreness, while the full system returned a three-state recommendation (*Reduce*, *Maintain*, or *Progress*) through probability-to-action mapping, operational guardrails, and staff oversight.

Across two seasons and under rolling-origin, leave-one-player-out, and between-season validation, the logistic regression model achieved ROC-AUC values of 0.759 ± 0.089 (rolling-origin) and 0.801 (between-season). A transparent baseline-ladder analysis showed that the model retained a predictive, reliable signal beyond the persistence of soreness, and that the three-state recommendation layer separated observed outcomes monotonically (*Progress* 0.054 vs. *Maintain* 0.217 vs. *Reduce* 0.326; $p < 0.001$). Within the studied elite women’s football club, these findings suggest that daily fatigue management may be approached jointly as a prediction task and an information-systems design problem, provided that integration, interpretability, calibration, honest baselining, and workflow fit are treated as joint design requirements. External validation across clubs, training methodologies, competitive levels, and male football populations is required before broader applicability can be claimed.

Author Contributions: Conceptualization, T.P., M.L. and K.D.; methodology, T.P., M.L. and T.G.; software, T.P. and M.L.; validation, T.G., M.L. and B.G.; formal analysis, T.P., T.G. and M.L.; investigation, T.P. and M.L.; resources, K.D. and B.G.; data curation, M.L., K.D. and B.G.; writing—original draft preparation, T.P., T.G. and M.L.; writing—review and editing, T.P., T.G., K.D. and B.G.; visualization, T.P. and T.G.; supervision, T.P. and T.G.; project administration, T.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were not required for this study. The data were not generated through any research intervention but originated from routine, contractually based performance monitoring conducted by the club (KKS Lech Poznań) in the ordinary course of professional team operation, and were analyzed retrospectively in pseudonymized form. The club acts as the data controller and is responsible for data collection, informing players about the monitoring program, and pseudonymizing records prior to analysis; the university research team acted solely as a processor of already-pseudonymized data, with no access to directly identifying information. The study was conducted in accordance with the principles of the Declaration of Helsinki.

Informed Consent Statement: As part of their professional engagement with the club, players are informed about the routine performance-monitoring program and consent to it under their contractual and club-regulatory arrangements. The data were analyzed retrospectively in a pseudonymized form, and no individual is identifiable in any reported results or figures.

Data Availability Statement: The data analyzed in this study are not publicly available. They consist of pseudonymized athlete-monitoring records owned by the club (KKS Lech Poznań) and are subject to privacy and contractual restrictions, including the presence of special-category personal data under the GDPR. Access to the data is therefore restricted and requires the club's permission.

Acknowledgments: The authors thank KKS Lech Poznań for providing access to the pseudonymized athlete-monitoring data and for supporting this work.

Conflicts of Interest: Authors Martyna Ławniczak, Kaja Dziergas, and Bartłomiej Grzelak were employed by KKS Lech Poznań. The remaining authors, Tomasz Piłka and Tomasz Górecki, declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ACWR	Acute:Chronic Workload Ratio
CR-10	Category-Ratio 10 Scale
DSS	Decision-Support System
GPS	Global Positioning System
HSR	High-Speed Running
MD	Match Day
RPE	Rating of Perceived Exertion
ROC-AUC	Receiver Operating Characteristic Area Under the Curve
sRPE	Session Rating of Perceived Exertion
TD-ACWR	Time-Dependent Acute:Chronic Workload Ratio

Appendix A. Microcycle Composition of the Modeling Sample

Table A1. Microcycle-specific sample sizes and prevalence of elevated next-morning soreness in the full-source modeling sample.

Microcycle Category	Player-Days (n)	Elevated Soreness (%)
Recovery 1 (MD+1)	44	13.6
Recovery 2–5 (MD+2 to MD+5)	697	6.6
Mid-week (MD-6 to MD-4)	820	9.6

Table A1. Cont.

Microcycle Category	Player-Days (n)	Elevated Soreness (%)
Pre-match 2–3 (MD-3, MD-2)	750	12.7
Pre-match 1 (MD-1)	578	4.2
Match day (MD)	497	20.1
Total	3386	10.3

Note. Elevated soreness was defined as next-morning muscle soreness ≥ 4 on the five-point wellness scale. Microcycle categories sum to the full-source modeling sample.

References

1. Impellizzeri, F.M.; Rampinini, E.; Coutts, A.J.; Sassi, A.; Marcora, S.M. Use of RPE-Based Training Load in Soccer. *Med. Sci. Sport. Exerc.* **2004**, *36*, 1042–1047. [\[CrossRef\]](#) [\[PubMed\]](#)
2. Halson, S.L. Monitoring Training Load to Understand Fatigue in Athletes. *Sport. Med.* **2014**, *44*, S139–S147. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Thorpe, R.T.; Atkinson, G.; Drust, B.; Gregson, W. Monitoring Fatigue Status in Elite Team-Sport Athletes: Implications for Practice. *Int. J. Sport. Physiol. Perform.* **2017**, *12*, S2–27–S2–34. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Saw, A.E.; Main, L.C.; Gustin, P.B. Monitoring the athlete training response: Subjective self-reported measures trump commonly used objective measures: A systematic review. *Br. J. Sport. Med.* **2016**, *50*, 281–291. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Gabbett, T.J. The Training-Injury Prevention Paradox: Should Athletes Be Training Smarter and Harder? *Br. J. Sport. Med.* **2016**, *50*, 273–280. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Xu, D.; Zhou, H.; Jie, T.; Zhou, Z.; Yuan, Y.; Jemni, M.; Quan, W.; Gao, Z.; Xiang, L.; Gusztav, F.; et al. Data-Driven Deep Learning for Predicting Ligament Fatigue Failure Risk Mechanisms. *Int. J. Mech. Sci.* **2025**, *301*, 110519. [\[CrossRef\]](#)
7. Xu, D.; Zhou, H.; Quan, W.; Gusztav, F.; Wang, M.; Baker, J.S.; Gu, Y. Accurately and Effectively Predict the ACL Force: Utilizing Biomechanical Landing Pattern before and after-Fatigue. *Comput. Methods Programs Biomed.* **2023**, *241*, 107761. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Power, D.J. *Decision Support Systems: Concepts and Resources for Managers*; Quorum Books: Westport, CT, USA, 2002.
9. Hevner, A.R.; March, S.T.; Park, J.; Ram, S. Design Science in Information Systems Research. *MIS Q.* **2004**, *28*, 75–105. [\[CrossRef\]](#)
10. Peffers, K.; Tuunanen, T.; Rothenberger, M.A.; Chatterjee, S. A Design Science Research Methodology for Information Systems Research. *J. Manag. Inf. Syst.* **2007**, *24*, 45–77. [\[CrossRef\]](#)
11. Mosqueira-Rey, E.; Hernández-Pereira, E.; Alonso-Ríos, D.; Bobes-Bascarán, J.; Fernández-Leal, Á. Human-in-the-loop Machine Learning: A State of the Art. *Artif. Intell. Rev.* **2023**, *56*, 3005–3054. [\[CrossRef\]](#)
12. Śliwowski, R.; Górecki, T.; Włodarczyk, M.; Paillard, T.; Zwierko, T.; Marynowicz, J. Predicting Differential Ratings of Perceived Exertion (dRPE) in Youth Soccer Using Decision Tree Models. *Sci. Rep.* **2026**, *16*, 21220. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Beato, M.; Vicens-Bordas, J.; Peña, J.; Costin, A.J. Training Load Comparison Between Small, Medium, and Large-Sided Games in Professional Football. *Front. Sport. Act. Living* **2023**, *5*, 1165242. [\[CrossRef\]](#) [\[PubMed\]](#)
14. Costa, J.A.; Rago, V.; Brito, P.; Figueiredo, P.; Sousa, A.; Abade, E.; Brito, J. Training in women soccer players: A systematic review on training load monitoring. *Front. Psychol.* **2022**, *13*, 943857. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Carmichael, M.; Thomson, R.; Moran, L.; Wycherley, T. The Impact of Menstrual Cycle Phase on Athletes' Performance: A Narrative Review. *Int. J. Environ. Res. Public Health* **2021**, *18*, 1667. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Lolli, L.; Batterham, A.M.; Hawkins, R.; Kelly, D.M.; Strudwick, A.J.; Thorpe, R.; Gregson, W.; Atkinson, G. Mathematical coupling causes spurious correlation within the conventional acute-to-chronic workload ratio calculations. *Br. J. Sport. Med.* **2019**, *53*, 921–922. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Impellizzeri, F.M.; Tenan, M.S.; Kempton, T.; Novak, A.; Coutts, A.J. Acute:Chronic Workload Ratio: Conceptual Issues and Fundamental Pitfalls. *Int. J. Sport. Physiol. Perform.* **2021**, *16*, 907–913. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Wang, C.; Vargas, J.; Stokes, T.; Steele, R.; Shrier, I. Analyzing Activity and Injury: Lessons Learned from the Acute:Chronic Workload Ratio. *Sport. Med.* **2020**, *50*, 1243–1254. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Chen, L.; Wen, S.; Tang, Y.; Ma, Y.; Wang, H.; Geda, M.W. Brain Power Mapping with Optimized Deep Learning for EEG-Based Pilot Fatigue Detection. *Biomed. Signal Process. Control* **2026**, *114*, 109284. [\[CrossRef\]](#)
20. Rossi, A.; Pappalardo, L.; Cintia, P.; Iaia, F.M.; Fernández, J.; Medina, D. Effective injury forecasting in soccer with GPS training data and machine learning. *PLoS ONE* **2018**, *13*, e0201264. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Piłka, T.; Grzelak, B.; Sadurska, A.; Górecki, T.; Dyczkowski, K. Predicting Injuries in Football Based on Data Collected from GPS-Based Wearable Sensors. *Sensors* **2023**, *23*, 1227. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Majumdar, A.; Bakirov, R.; Hodges, D.; Scott, S.; Rees, T. Machine Learning for Understanding and Predicting Injuries in Football. *Sport. Med. Open* **2022**, *8*, 73. [\[CrossRef\]](#) [\[PubMed\]](#)

23. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. *Proc. Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4765–4774.
24. Lu, Y.; Wang, S.; Chen, R.; Dui, H.; Gao, J.; Zhang, C. Human-Centric Reliability Prediction for IoT-Enabled Human-Machine Systems Considering Component Hierarchical Dependency. *Int. J.-Hum.-Comput. Interact.* **2026**. [[CrossRef](#)]
25. Ofoghi, B.; Zeleznikow, J.; MacMahon, C.; Dwyer, D. Supporting Athlete Selection and Strategic Planning in Track Cycling Omnium: A Statistical and Machine Learning Approach. *Inf. Sci.* **2013**, *233*, 200–213. [[CrossRef](#)]
26. Burger, J.; Henze, A.S.; Voit, T.; Latzel, R.; Moser, O. Athlete Monitoring Systems in Elite Men's Basketball: Challenges, Recommendations, and Future Perspectives. *Transl. Sport. Med.* **2024**, *2024*, 6326566. [[CrossRef](#)] [[PubMed](#)]
27. Schelling, X.; Spencer, B.; Azalbert, V.; Alonso-Perez-Chao, E.; Sosa, C.; Robertson, S. Decision Support Systems for Time Series in Sport: Literature Review and Applied Example of Change-point-Based Most Demanding Scenario Analysis in Basketball. *Appl. Sci.* **2025**, *15*, 10575. [[CrossRef](#)]
28. Mănescu, D.C. Big Data Analytics Framework for Decision-Making in Sports Performance Optimization. *Data* **2025**, *10*, 116. [[CrossRef](#)]
29. Van Calster, B.; McLernon, D.J.; van Smeden, M.; Wynants, L.; Steyerberg, E.W. Calibration: The Achilles heel of predictive analytics. *BMC Med.* **2019**, *17*, 230. [[CrossRef](#)] [[PubMed](#)]
30. Borg, G. *Borg's Perceived Exertion and Pain Scales*; Human Kinetics: Champaign, IL, USA, 1998.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.