Leveraging GPS Data and Attention-based BiLSTM for Injury Prediction in Professional Football

Aleksandra Sadurska
Adam Mickiewicz University
Faculty of Mathematics and Computer Science
Poznań, Poland *aleksandra.sadurska@amu.edu.pl*

Michał Zaręba
Adam Mickiewicz University
Faculty of Mathematics and Computer Science
Poznań, Poland *michal.zareba@amu.edu.pl*

Tomasz Piłka
Adam Mickiewicz University
Faculty of Mathematics and Computer Science
Poznań, Poland *tomasz.pilka@amu.edu.pl*

Tomasz Górecki
Adam Mickiewicz University
Faculty of Mathematics and Computer Science
Poznań, Poland *tomasz.gorecki@amu.edu.pl*

Abstract

Injuries pose a significant challenge in professional football, affecting player availability, team performance, and club finances. Accurate prediction of injury risk is crucial for implementing effective prevention strategies. This study develops a deep learning model to predict the likelihood of injury in professional football players using data collected through Catapult Sports tracking devices. The research is carried out in collaboration with KKS Lech Poznań, a Polish football club. The proposed model architecture combines bidirectional long- and short-term memory networks with an attention mechanism to learn from the time series data and predict player injury risk. The model is trained on sequences of data spanning 14 days before each recorded injury or non-injury event. To address class imbalance, a custom loss function was implemented that balances focal loss and the F_β score. The model's performance is evaluated on an independent test set, achieving a specificity of 0.90, an accuracy of 0.90, and a recall of 0.40.

Keywords: Sports Analytics, Injury Prediction, BiLSTM, Imbalanced Classes Problem

1. Introduction

In recent years, data analysis and machine learning have become deeply embedded across a broad spectrum of industries, including professional sports. The conversation has shifted from whether artificial intelligence (AI) should be adopted to how extensively it can be applied to solve complex, multifaceted problems. In the domain of professional football, clubs increasingly use predictive models to optimize ticket prices [11], enhance recruitment strategies by identifying high-value players, and utilize advanced metrics such as expected goals (xG) and expected assists (xA) to evaluate on-field performance [4], [13]. Despite these advances, several critical challenges remain underexplored or insufficiently addressed, chief among them being injury prediction.

Predicting non-contact injuries in elite football, which typically result from over- or undertraining, represents a multidimensional problem that extends beyond physiological data alone. Effective predictive models must integrate a variety of dimensions, including difficult-to-quantify factors such as psychological well-being, mental fatigue, and contextual stressors [6], [9].

This study proposes a novel, comprehensive injury prediction framework integrating advanced methods across all stages of the modeling pipeline. The approach begins with rigorous data preprocessing and feature selection based on LASSO regularization, ANOVA F-tests, and Random Forest importance measures. Athlete-specific multivariate time series are then constructed using a sliding window strategy to preserve temporal dependencies. For classification, a bidirectional long- and short-term memory (BiLSTM) network is combined with an attention mechanism, utilizing a custom loss function that combines focal loss and F_β score optimization to address the substantial class imbalance.

2. Related Work

Numerous studies have explored injury prediction using various analytical approaches, ranging from traditional statistical methods to advanced machine learning techniques [2], [7]. For example, Rossi et al. [9] utilized gradient boosting algorithms to predict injury risk based on workloads and player-specific variables. Their approach demonstrated that machine learning methods could effectively capture non-linear relationships between training loads and injury occurrence. Ruddy et al. [6] demonstrated how combining workload data with players' physical responses improves predictive accuracy, highlighting the importance of individualized models.

Recent advancements also emphasize the role of wearable sensor technology for injury prevention. Carey et al. [2] explored the use of GPS and accelerometer-based metrics to assess injury risks among elite athletes. Their work demonstrated that high-frequency data collection could reveal patterns that are invisible to traditional monitoring methods. Similarly, López-Valenciano et al. [7] systematically reviewed the effectiveness of machine learning algorithms, confirming their superior performance over conventional statistical methods for injury forecasting in football [8], [10].

Despite these promising approaches, injury prediction remains a complex analytical and practical challenge, primarily due to the intricate interplay of physiological, environmental, and behavioral factors involved. Most existing approaches focus on tabular data analysis without considering the temporal nature of athlete monitoring data, which may contain crucial sequential patterns predictive of injury events.

3. Dataset Description

The dataset employed in this study comprises real-world measurements obtained through collaboration with KKS Lech Poznań, a professional football club. Data collection encompasses a wide range of on-pitch activities, including daily training sessions, preseason preparations, friendly matches, domestic league matches, and international competitions. Measurements include all athletes assigned to both the first and second teams, thus covering players across all positions.

Player data is recorded using Catapult wearable GPS trackers.[1] This GPS equipment was equipped with an accelerometer, gyroscope, and magnetometer (3D), all of which sampled data at a frequency of 100 Hz. These devices are integrated into specially designed vests worn underneath a match and training apparel. These devices utilize

[1]Vector S7 4 GHz, Catapult Innovations, Melbourne, Australia, `https://www.catapultsports.com/`

GPS technology, accelerometers, and gyroscopes to capture multidimensional movement data, including velocity, direction, and distance covered.

The raw dataset, retrieved via the Catapult API, comprises 1,738 parameters; however, not all are explicitly related to football performance. The API delivers a range of metrics, from fundamental indicators such as total distance, active time, acceleration counts, and high-speed distance (across multiple thresholds), to advanced and complex metrics, including Total Player Load[2].

Another dataset included in this study consists of medical records manually maintained by the club's medical staff. This dataset documents all player-reported injuries from the 2019/20 season onward. Each record contains the date of injury, the player's identity, and detailed information about the injury, including the affected body part and the injury mechanism. These records are primarily used to annotate corresponding entries in the Catapult dataset, indicating whether a player sustained an injury on a particular day.

4. Methodology

4.1. Data Acquisition and Preprocessing

Predicting non-contact injuries resulting from over- or undertraining presents a highly complex and non-trivial challenge. Even when employing machine learning or deep learning techniques, the quality and structure of the input data remain critical to the performance of predictive models. The dataset used comprises measurements collected across multiple seasons from all players assigned to the team, totaling 1,738 parameters.

However, variations in team performance and training strategies between seasons, often influenced by changes in coaching staff or participation in international competitions, can introduce inconsistencies. Moreover, certain players may contribute limited value to the dataset due to their infrequent involvement, potentially acting as sources of noise. To mitigate these issues, a series of preprocessing steps was implemented.

Initially, columns with a percentage of missing values exceeding a predefined threshold (set at 60%) were removed, as they are unlikely to provide meaningful information for the study. This step also facilitated the practical application of the K-Nearest Neighbors (KNN) imputation method, which estimates missing values by identifying the K nearest instances based on a distance metric computed over the available features.

Following this, the dataset was scanned for NaN values, and any incomplete records were imputed. Subsequently, in consultation with the Head of the Research and Development Department at KKS Lech Poznań, a curated list of players was excluded from the training dataset. This list comprised individuals with minimal participation due to long-term injuries or transfers. Furthermore, only athletes appearing in the medical report, i.e., those who sustained at least one injury, were retained. Goalkeepers were excluded due to the unique nature of their role and workload.

The remaining data was aggregated by date and player to account for multiple activities performed by a player on the same day. The preprocessed dataset was annotated using the medical report and sorted chronologically by players and date, enabling the construction of time series for subsequent analysis. After the aforementioned data processing, the dataset contains 21,203 rows and 1122 numeric features.

[2]Calculated by adding up the acceleration in all directions, as captured by a three-axis accelerometer

Feature Selection

Feature selection is a crucial step in machine learning pipelines, particularly when working with high-dimensional and imbalanced datasets. In this study, we employed a hybrid feature selection strategy combining Random Forest importance measures [1], LASSO regularization [12], and ANOVA F-tests [5].

Random Forests use an ensemble approach to estimate the importance of features by measuring their contribution to the model's predictive accuracy. LASSO combines variable selection and regularisation by penalising the sum of the absolute values of the model coefficients, effectively reducing some coefficients to zero. ANOVA F-tests evaluate the statistical significance of individual features by analysing the variance between different classes.

By combining these techniques, we aimed to enhance the robustness and generalisability of the model. Ultimately, 100 numerical features were selected for the final dataset based on their high importance in at least two of the three methods.

Creating Sequences

To prepare data for training and evaluation of a BiLSTM neural network, athlete-monitoring data were systematically segmented into fixed-length sequences that explicitly capture temporal dependencies. Specifically, sequences of length $L = 14$ days were generated using a sliding-window approach, where each input sequence is defined as $X_t = \{x_{t-L+1}, x_{t-L+2}, \ldots, x_t\}$. Each sequence was associated with a binary target variable y_{t+1}, representing injury occurrence on the immediate subsequent day.

To maintain temporal consistency and prevent cross-season data leakage, sequence generation was confined to specific seasonal intervals. The data were then split into training and testing sets based on a cut-off date to ensure that evaluation was based only on future, unseen data, providing a more realistic performance assessment.

4.2. Addressing Class Imbalance

Dataset imbalance is a crucial problem in this research. After dataset preprocessing, for all 21,203 remaining records, only 136 are labelled as injury, so the final dataset contains 0.64% of the minority class.

Focal Loss

The focal loss function is a significant development in addressing class imbalance issues in deep learning classification problems. It is defined as follows:

$$FL(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t), \tag{1}$$

where p_t represents the model's estimated probability for the actual class, α_t is a balancing factor, and γ is the focusing parameter that modulates the rate at which easy examples are down-weighted. This formulation extends cross-entropy loss by incorporating a modulating factor $(1 - p_t)^{\gamma}$, which automatically reduces the contribution of well-classified examples and concentrates the optimization on challenging examples.

F_β *Loss*

The F_β loss presents a direct optimization method for the F_β metric, defined as

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}, \tag{2}$$

where β controls the relative importance between precision and recall. This implementation directly computes micro-averaged precision and recall from predicted probabilities and ground truth labels.

Custom Loss Function

Addressing the challenges of highly imbalanced datasets in sequential pattern recognition, this study presents a parameterized combinatorial loss framework that dynamically balances Focal and F_β components:

$$\mathcal{L}_{\text{custom}}(y_{\text{true}}, y_{\text{pred}}) = \lambda\, \mathcal{L}_{\text{focal}}(y_{\text{true}}, y_{\text{pred}}) + (1-\lambda)\, \mathcal{L}_{F_\beta}(y_{\text{true}}, y_{\text{pred}}), \tag{3}$$

where $\lambda \in [0, 1]$ controls the balance between focal loss and F_β loss. In this work, α, γ, β, and λ parameters are tuned automatically using the Optuna optimization framework. This dual-component architecture eliminates the need for explicit data resampling techniques and offers flexibility through hyperparameter optimization.

4.3. Baseline Model with XGBoost

To evaluate the effectiveness of sequential modeling for injury prediction, we developed a baseline approach using XGBoost [3]. The baseline model was trained on the same dataset and feature set as the sequential model, but treated the data as independent daily records rather than temporal sequences. Hyperparameter optimization was conducted using the Optuna framework with 100 trials (Table 1).

Table 1. Best hyperparameters selected by Optuna for XGBoost model.

Hyperparameter	**Value**
Max depth	4
Learning rate	0.0105
Number of estimators	500
Gamma	0.754
Min child weight	6
Subsample	0.794
Colsample by tree	0.714
Scale positive weight	200

4.4. Model Architecture

The injury prediction architecture is based on a deep BiLSTM network (see Table 2) enhanced with attention mechanisms and sophisticated pooling techniques. BiLSTM networks effectively capture long-term dependencies in sequential data by processing inputs in both the forward and backward temporal directions. This is essential for modelling the complex dynamics of player workload over time.

Incorporating an attention layer enables the model to focus on the most relevant temporal patterns, thereby reducing the risk of overfitting irrelevant or repetitive sequences. Feature extraction is further improved through the use of both global average pooling and max pooling operations. Additionally, dense layers with L2 regularisation and dropout are employed to guard against overfitting further. The optimal parameters selected by Optuna are detailed in Table 3.

Table 2. Summary of the BiLSTM-Attention model architecture.

Layer (Type)	Output Shape	Param #
InputLayer	(None, 14, 102)	
Bidirectional	(None, 14, 256)	236,544
Dropout	(None, 14, 256)	
Bidirectional	(None, 14, 64)	73,984
Dropout	(None, 14, 64)	
Attention	(None, 14, 64)	
GlobalAvgPool1D	(None, 64)	
GlobalMaxPool1D	(None, 64)	
Concatenate	(None, 192)	
Dense	(None, 64)	12,352
BatchNorm	(None, 64)	256
Dropout	(None, 64)	
Dense	(None, 32)	2,080
BatchNorm	(None, 32)	128
Dropout	(None, 32)	
Dense	(None, 1)	33
Total		**325,377**

Table 3. Best hyperparameters selected by Optuna for the BiLSTM-Attention model.

Hyperparameter	Value
LSTM units (Layer 1)	128
LSTM units (Layer 2)	32
Dropout rate (Layer 1)	0.20
Dropout rate (Layer 2)	0.30
Final dropout rate	0.40
Dense units (Layer 1)	64
Dense units (Layer 2)	32
L2 regularization factor	0.00663
Learning rate	0.00017
Focal loss: α	0.25
Focal loss: γ	9.0
Loss combination: λ	0.7
F_β: β	2
Positive class weight	20.0
Batch size	128

Evaluation with Sliding Window

To account for temporal uncertainty in injury prediction, model evaluation was performed using a sliding window tolerance approach. Let t_p denote the date of a predicted injury and t_a the date of an actual injury. A prediction is classified as a true positive if it occurs within a tolerance window of $\pm d$ days around the actual injury date, where $d = 5$. This evaluation strategy reflects the clinical relevance of near-miss predictions, providing a more robust assessment of model performance.

4.5. Ablation Study: Impact of Loss Function Design

To assess the influence of different loss functions on model performance, we conducted an ablation study comparing binary cross-entropy, focal loss, and the proposed custom combination loss. As shown in Fig. 1, the custom loss function led to the most balanced performance across all evaluation metrics.

While binary cross-entropy yielded the highest accuracy (0.9448), it entirely failed to detect injury cases, resulting in a recall and F1-score of 0.0. This confirms that standard loss functions are inadequate in scenarios with extreme class imbalance. Focal loss improved recall substantially to 0.50 and yielded the highest AUC (0.6667), confirming its ability to down-weight easily classified majority-class examples.

The custom combination loss demonstrated a more favorable balance between sensitivity and overall classification performance. It achieved a recall of 0.40 and the highest F1-score (0.0229), while maintaining a high AUC (0.6563) and significantly better accuracy (0.9049) than focal loss.

5. Results and Discussion

The evaluation of the proposed injury prediction model using a ±5-day tolerance window demonstrates promising results particularly given the extreme imbalance and inherent uncertainty of the dataset. The model achieved a specificity of 0.90 and an overall accuracy of 0.90, underscoring its ability to identify non-injury cases while maintaining

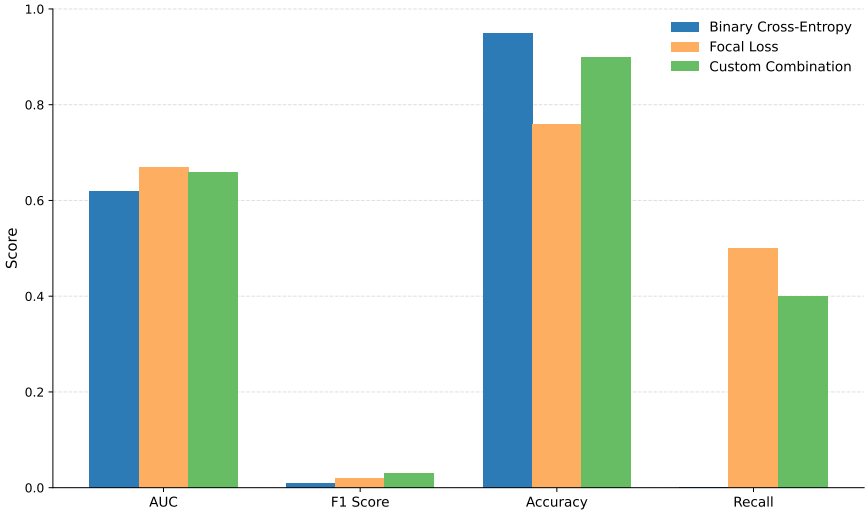

Fig. 1. Comparison of loss functions on test set performance.

robust generalization correctly. Despite the considerable challenge of predicting rare injury events, the model attained a recall of 0.40, successfully detecting 40% of injuries within the defined temporal window.

Timing analysis revealed that true positive predictions occurred, on average, within 4.25 days after the injury event, demonstrating strong alignment with clinically relevant timeframes. Notably, no predictions happened before the actual injury dates, indicating a conservative and stable model behavior.

Compared to the sequential BiLSTM-Attention model, the XGBoost baseline achieved noticeably lower predictive performance, as summarized in Table 4.

Table 4. Comparison of BiLSTM-Attention and XGBoost models on injury prediction task. Evaluation performed using a ±5-day tolerance window.

Metric	**BiLSTM-Attention**	**XGBoost Baseline**
Precision (Injury)	0.02	0.02
Recall (Injury)	0.40	0.17
Accuracy	0.90	0.80
Specificity (Non-Injury)	0.90	0.83

Experiments incorporating more sophisticated architectural components, such as multi-head attention mechanisms, resulted in a significant increase in false positive rates, suggesting that excessive model complexity may lead to overfitting to noisy sequential patterns in highly imbalanced datasets. This observation highlights a fundamental challenge: the onset of injury may be too complex to predict reliably using simple daily aggregation windows.

Future studies may benefit from incorporating more granular temporal features, such as cumulative workload metrics or player-specific fatigue models, to enhance predictive precision. Further research should explore the application of alternative machine learning approaches, including ensemble methods or transformer-based architectures specifically adapted for sparse event detection.

Acknowledgements

This research and the resulting article were made possible through the cooperation and support provided by the KKS Lech Poznań club.

References

[1] Breiman, L.: Random forests. Machine Learning 45(1), pp. 5–32 (2001)

[2] Carey, D., Ong, K.L., Whiteley, R., Crossley, K., Crow, J., Morris, M.: Predictive modelling of training loads and injury in australian football. International Journal of Computer Science in Sport 17 (06 2017)

[3] Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). pp. 785–794 (2016)

[4] Decroos, T., Bransen, L., Van Haaren, J., Davis, J.: Actions speak louder than goals: Valuing player actions in soccer. In: Proceedings of the 25th ACM SIGKDD. pp. 1851–1861 (2019)

[5] Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning Research 3, pp. 1157–1182 (2003)

[6] JD, R., AJ, S., N, M., MD, W., S, D., RG, T., J, H., MN, B., DA, O.: Predictive modeling of hamstring strain injuries in elite australian footballers. Medicine and Science in Sports and Exercise 50(5), pp. 906–914 (2018)

[7] López-Valenciano, A., Ayala, F., Puerta, J.M., Croix, M.D.S., Vera-Garcia, F.J., Hernández-Sánchez, S., Ruiz-Pérez, I., Myer, G.D.: A preventive model for muscle injuries: A novel approach based on learning algorithms. Medicine & Science in Sports & Exercise 50, pp. 915–927 (2018)

[8] Piłka, T., Grzelak, B., Sadurska, A., Górecki, T., Dyczkowski, K.: Predicting Injuries in Football Based on Data Collected from GPS-Based Wearable Sensors. Sensors 23(3) (2023)

[9] Rossi, A., Pappalardo, L., Cintia, P., Iaia, F.M., Fernàndez, J., Medina, D.: Effective injury forecasting in soccer with gps training data and machine learning. PLOS ONE 13(7) (07 2018)

[10] Sadurska, A., Piłka, T.K., Grzelak, B., Górecki, T., Dyczkowski, K., Zaręba, M.: Fusion of a fuzzy rule-based method and other decision-making models in injury prediction problem in football. In: 2023 IEEE International Conference on Fuzzy Systems (FUZZ) Proceedings. pp. 1–6 (2023)

[11] Shapiro, S.L., Drayer, J., Dwyer, B.: Variable ticket pricing in sport: A review of the academic literature. Sport Management Review 22(4), pp. 512–523 (2019)

[12] Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58(1), pp. 267–288 (1996)

[13] Zaręba, M., Piłka, T., Górecki, T., Grzelak, B., Dyczkowski, K.: Improving the evaluation of defensive player values with advanced machine learning techniques. In: Marcinkowski, B., Przybyłek, A., et al. (eds.) Harnessing Opportunities: Reshaping ISD in the post-COVID-19 and Generative AI Era (ISD2024 Proceedings). pp. 1–5 (2024)