

Predição de Lesões no Futebol Profissional com ACWR e Aprendizado de Máquina

Gabriel Padrão¹, Matheus Melo¹, Ana G. Araújo¹, Juliano Spinetti²,
Lucas Giusti¹, Diego Brandão¹, Jorge Soares¹

¹ Cefet/RJ – Rio de Janeiro, RJ – Brasil

² Fluminense Football Club – Rio de Janeiro, RJ – Brasil

{gabriel.padrao, lucas.giusti}@aluno.cefet-rj.br

{diego.brandao, jorge.soares}@cefet-rj.br

Abstract. *The increasing density of professional football calendars has led to a higher incidence of non-contact injuries, affecting both athletic performance and the financial health of elite clubs. In this context, this study explores the use of machine learning models to predict such injuries using external-load data collected via GPS devices and the Acute:Chronic Workload Ratio (ACWR) as the central indicator. The best-performing model achieved an F1 score of 93.3%.*

Resumo. *A intensificação do calendário no futebol profissional tem aumentado a incidência de lesões, com impactos no desempenho esportivo e prejuízos econômicos, especialmente para clubes de elite. Diante disso, este trabalho visa explorar modelos de aprendizado de máquina capazes de prever tais lesões, utilizando dados de dispositivos GPS e o indicador de proporção de carga de trabalho aguda e crônica. O modelo de aprendizado de máquina com o melhor resultado desse estudo conseguiu prever lesões com um F1 score de 93,3%.*

1. Introdução

O futebol profissional tem enfrentado calendários cada vez mais intensos, com menor tempo de recuperação entre partidas, o que eleva a carga física imposta aos atletas e, consequentemente, o risco de lesões sem contato [Malone et al., 2015; Bengtsson et al., 2013]. Essas lesões impactam não apenas o desempenho esportivo, mas também a sustentabilidade financeira dos clubes, especialmente quando envolvem jogadores-chave [Salces et al., 2014; Rossi et al., 2018].

A predição de lesões é um problema desafiador, dada a variabilidade individual dos atletas e a influência de fatores fisiológicos e contextuais [Pilkka et al., 2023]. Ainda assim, evidências indicam que aumentos abruptos na carga de treinamento estão associados a maior incidência de lesões, o que destaca a importância do monitoramento contínuo da carga [T. J. Gabbett, 2016].

Nesse contexto, dispositivos GPS têm sido amplamente utilizados para mensurar a carga externa, fornecendo dados detalhados sobre deslocamento, velocidade e intensidade do esforço [Di Credico et al., 2021]. A métrica Acute:Chronic Workload Ratio (ACWR), definida como a razão entre a carga recente (aguda) e a carga acumulada (crônica), permite quantificar variações na carga em relação ao nível de adaptação do atleta, sendo um

indicador relevante para identificação de picos de sobrecarga [Dwyer e Tim J. Gabbett, 2012; Jiang et al., 2022].

Diante disso, este trabalho procura responder à seguinte questão de pesquisa: *”Como diferentes configurações da métrica ACWR e conjuntos de variáveis de carga externa influenciam o desempenho de modelos de aprendizado de máquina na predição de lesões sem contato em futebol profissional?”*

Como objetivo derivado, o presente artigo investiga o uso de modelos de aprendizado de máquina para prever lesões sem contato no futebol profissional, com base em dados de GPS e na métrica ACWR. Com isso, deseja-se avaliar o impacto dessa métrica na capacidade preditiva dos modelos, contribuindo para o desenvolvimento de ferramentas que auxiliem na gestão da carga e na prevenção de lesões. Além disso, considerando o desbalanceamento típico desse problema, adotam-se métricas de avaliação mais adequadas, como *F1 score* e AUC-ROC.

Este trabalho possui mais quatro seções. A Seção 2 apresenta os trabalhos relacionados. A Seção 3 descreve a metodologia com foco no uso do ACWR. A Seção 4 descreve os resultados obtidos. Por fim, a Seção 5 apresenta as conclusões e trabalhos futuros.

2. Trabalhos Relacionados

Diversos estudos têm investigado o uso de aprendizado de máquina na predição de lesões esportivas. Os trabalhos relacionados foram identificados por meio de buscas na base Scopus, utilizando a *string*: ((“soccer” OR “football”) AND (“injury prediction” OR “injury risk prediction” OR “acute chronic workload ratio” OR “ACWR”) AND NOT (“Australian Football” OR “American Football” OR “NFL” OR “semiprofessional”)). A busca foi realizada em 04/06/2025, retornando 188 trabalhos, dos quais foram selecionados cinco com maior aderência ao escopo deste estudo.

A revisão sistemática de Eetvelde et al., 2021 sintetiza o uso de técnicas de aprendizado de máquina na predição de lesões, destacando desafios relacionados à qualidade dos dados e à generalização dos modelos.

Em Rossi et al., 2018, dados de GPS foram utilizados para prever lesões sem contato por meio da métrica ACWR, com modelos como Decision Tree, Random Forest e Logistic Regression. A Decision Tree apresentou melhor desempenho, com cerca de 80% de detecção das lesões.

O estudo de Melo et al., 2024 propõe uma abordagem semelhante, porém sem o uso do ACWR, apresentando desempenho inferior ao deste trabalho, com AUC-ROC de 79,8% e *F1 score* de 5,5%. Enquanto Vallance et al., 2020 combinou dados objetivos (GPS) e subjetivos (RPE – Rate of Perceived Exertion), avaliando diversos algoritmos. Os melhores resultados foram obtidos com a integração de dados de carga, histórico de lesões e características dos atletas.

Mais recentemente, Piłka et al., 2023 explorou a combinação de regras baseadas em conhecimento especializado com aprendizado de máquina. O modelo *fuzzy* destacou-se pela interpretabilidade, enquanto o XGBoost apresentou melhor desempenho preditivo.

Apesar dos avanços, ainda existem lacunas quanto à padronização das janelas do

ACWR e à integração de múltiplas variáveis de carga. Este trabalho busca contribuir nesse contexto ao avaliar diferentes configurações da métrica ACWR em conjunto com modelos de aprendizado de máquina.

3. Metodologia

Esta seção detalha a metodologia adotada para desenvolver e validar modelos preditivos de lesões sem contato em jogadores de futebol profissional, com base no indicador ACWR. O fluxo geral divide-se em quatro etapas: (i) Conjunto de Dados e Limpeza, (ii) Acute:Chronic Workload Ratio (ACWR), (iii) Engenharia de Features e (iv) Modelagem e Otimização, conforme ilustrado na Figura 1.



Figura 1. Metodologia de modelos preditivos de lesões

3.1. Conjunto de Dados e Limpeza

Este estudo utiliza dados confidenciais de um clube da primeira divisão do futebol brasileiro, referentes a atletas profissionais masculinos entre 2017 e 2023, devidamente anonimizados. As informações de carga de treino e de partidas foram coletadas por dispositivos GPS acoplados a coletes, armazenadas na nuvem e acessadas via API Open Field¹, com o uso de um token criptografado fornecido pelo clube.

Paralelamente, o clube forneceu dados da equipe de fisiologia com o registro de 185 episódios de lesão entre 2017 e 2023. As variáveis incluíam: (i) Data da lesão; (ii) Data de liberação médica; (iii) Dias afastado; (iv) Nome do atleta; (v) Segmento corporal afetado; (vi) Tipo de lesão (aguda ou crônica); (vii) Subtipo; e (viii) Grau de severidade, conforme os dias afastado.

Foram removidos registros incompletos e selecionadas apenas lesões entre 2021 e 2023 dos tipos: (i) Muscular, (ii) Articular ou (iii) Ligamentar. Após essa filtragem, restaram 49 registros válidos, compatíveis com os objetivos da análise. O DataFrame resultante da junção entre os dados da API [omitido] e o cadastro de lesões, por meio das datas dos eventos, contém 65.906 registros e 1.701 variáveis. No pré-processamento, eliminaram-se variáveis compostas por valores nulos ou zero, e as observações faltantes foram imputadas pela média aritmética de cada variável.

3.2. Acute: Chronic Workload Ratio (ACWR)

O ACWR é calculado pela razão entre a carga aguda, associada ao esforço recente, e a carga crônica, acumulada em período mais extenso de atividades [Rossi et al., 2018]. Essa medida monitora variações da carga de trabalho em relação ao habitual do atleta, e não apenas à carga total [T. J. Gabbett, 2016; Andrade et al., 2020]. Aumentos súbitos e excessivos elevam o risco de sobrecarga e lesões [Bowen et al., 2020], permitindo uma gestão mais precisa e individualizada do volume e da intensidade conforme o limite fisiológico de cada atleta.

¹Omitido

A interpretação do ACWR orienta decisões práticas. Valores elevados (acima de 1.5) indicam carga aguda muito superior à crônica, elevando o risco de lesões sem contato [Bowen et al., 2020]. Já valores moderados (próximos de 1.0) refletem uma progressão equilibrada, ideal para adaptação e performance. Em contrapartida, valores baixos ($ACWR < 0.85$) também têm sido associados a maior risco de lesão sem contato, atribuído à subexposição do atleta a níveis adequados de carga, o que reduz sua tolerância fisiológica ao esforço [Carey et al., 2016; Delecroix et al., 2018].

3.3. Engenharia de Features

A seleção inicial de variáveis foi guiada pela literatura especializada, priorizando métricas de carga externa relevantes para o monitoramento do desempenho físico, como tempo de atividade, zonas de velocidade, acelerações, desacelerações, saltos e mudanças de direção, com o objetivo de capturar diferentes dimensões do esforço dos atletas.

A Tabela 1 apresenta as principais variáveis extraídas de dispositivos GPS, conforme indicado na literatura [Rossi et al., 2018; Piłka et al., 2023]. Todas foram agregadas no nível de microciclo, representando valores acumulados ou médios ao longo do intervalo entre duas partidas consecutivas.

Tabela 1. Variáveis de carga e desempenho obtidas por GPS.

Variáveis	Descrição	Variáveis	Descrição
field_time	Tempo em minutos em campo.	dist._>_25,2km/h_m/min	Distância (m/min) > 25,2 km/h.
total_distance	Distância total (m) percorrida.	acel+desacel_ima_alta	Acel./desacel. alta intensidade.
metrage_per_minute	Distância total (m/min).	acel+desacel_ima_alta_(min)	Acel./desacel. por minuto.
velocity_band1_total_distance	Distância (0–1,1 km/h).	acel+desacel_qtd._>2ms	Acel./desacel. >2 m/s ² .
velocity_band2_total_distance	Distância (1,1–7,2 km/h).	acel+desacel_qtd._>3ms	Acel./desacel. >3 m/s ² .
velocity_band3_total_distance	Distância (7,2–14,4 km/h).	total_player_load	Carga mecânica total.
velocity_band4_total_distance	Distância (14,4–19,8 km/h).	rhie_bout_count	Esforços intensos repetidos.
velocity_band5_total_distance	Distância (19,8–25,2 km/h).	rhie5_/min	Esforços intensos por minuto.
dist._>_19,8km/h	Distância > 19,8 km/h.	mudanças_direção_totais	Mudanças de direção.
dist._>_19,8km/h_m/min	Distância (m/min) > 19,8 km/h.	qtd_total_saltos	Saltos verticais.
		max_speed	Velocidade máxima.
velocity_band6_total_distance	Distância > 25,2 km/h.	max_effort_acceleration	Aceleração máxima.

A partir dessas variáveis, foram definidos sete conjuntos de *features*, denotados por $feat_i$, com $i \in \{1, \dots, 7\}$, cada um representando diferentes combinações de atributos originais e derivados. Para cada conjunto, consideraram-se duas definições da variável alvo: (i) uma versão binária ($feat_{i,bin}$), indicando a ocorrência de lesão, e (ii) uma versão baseada na reincidência ($feat_{i,soma}$), correspondente ao número total de lesões. Assim, foram avaliadas 14 configurações distintas.

Os dados foram agregados em microciclos (intervalos entre partidas), após a remoção de valores anômalos. Devido à variabilidade do calendário, a duração do microciclo não é fixa, podendo variar de poucos dias a mais de uma semana.

Em seguida, calculou-se o ACWR para as variáveis de carga, considerando diferentes janelas temporais para as componentes aguda e crônica, conforme a literatura, resultando em um conjunto final com 9.227 registros e 29 variáveis. A variável alvo foi definida no nível de microciclo, indicando a ocorrência de lesão sem contato no período subsequente. Dessa forma, o problema consiste em prever o risco de lesão no próximo microciclo a partir das cargas observadas nos períodos atual e anterior.

Entre os conjuntos avaliados, o melhor desempenho foi obtido com o *feat7_bin*, composto por variáveis relacionadas à carga de treino e jogo, como tempo em campo, zonas de velocidade, acelerações, desacelerações, mudanças de direção e saltos, além de atributos derivados, como a duração do microciclo. A diferença em relação ao *feat7_soma* reside exclusivamente na forma de representação da variável alvo.

3.4. Modelagem e otimização

A função principal foi projetada como um *pipeline* integrado, abrangendo do pré-processamento à modelagem. Inicialmente, o DataFrame, composto por combinações de variáveis de carga, o ACWR calculado e a variável-alvo (*Lesão*), é dividido em treino (80%) e teste (20%), utilizando validação por *holdout*, conforme a literatura [Rossi et al., 2018].

No pré-processamento, aplicou-se normalização com *StandardScaler* para evitar efeitos de escala entre variáveis. A redução de dimensionalidade por PCA² foi avaliada para mitigar multicolinearidade, mas não apresentou ganhos e, portanto, foi excluída dos resultados finais.

Diante do desbalanceamento entre casos de lesão e não lesão, foram testadas estratégias de balanceamento no pipeline, incluindo *undersampling* (*RandomUnderSampler*), *oversampling* (ADASYN) e ajuste de pesos de classe (*class_weight*), buscando equilibrar desempenho e simplicidade.

Em seguida, ajustou-se o limiar de classificação (*threshold*) para equilibrar falsos positivos e falsos negativos. O desempenho foi avaliado por precisão, *recall*, *F1 score* e AUC-ROC. Devido ao desbalanceamento dos dados, a acurácia não foi priorizada, pois pode induzir interpretações equivocadas em cenários com predominância da classe negativa. Assim, *F1 score* e AUC-ROC foram enfatizados por representarem melhor o equilíbrio entre precisão e *recall* e a capacidade discriminativa dos modelos na identificação de lesões.

Em seguida à definição dos algoritmos, realiza-se o ajuste do limiar de classificação (*threshold*), etapa importante para alinhar o modelo ao objetivo de equilibrar os erros de falso-positivo e de falso-negativo. Por fim, o desempenho é avaliado considerando precisão, *recall*, *F1 score* e AUC-ROC. Devido ao desbalanceamento do conjunto de dados, a acurácia não foi considerada métrica principal de avaliação, pois pode induzir interpretações equivocadas em cenários com predominância da classe negativa. Assim, priorizou-se o uso do *F1 score* e da AUC-ROC, por melhor capturarem o equilíbrio entre precisão e *recall* e a capacidade discriminativa dos modelos na identificação de lesões.

4. Resultados

Após avaliar diferentes janelas para cargas aguda e crônica, foram gerados 1.315 modelos. A configuração 10/40 dias apresentou o melhor desempenho, diferindo da abordagem mais comum na literatura, baseada em 7/28 dias.

O XGBoost destacou-se nos cinco melhores cenários, alcançando *F1 score* máximo de 93,3%. Os conjuntos *feat7_bin* e *feat7_soma* obtiveram os melhores resultados, reunindo métricas de carga associadas a tempo de treino e jogo, acelerações,

²do inglês *Principal Component Analysis*

desacelerações, mudanças de direção e saltos. A diferença entre eles está na representação da variável-alvo: *feat7_bin* usa indicador binário de lesão, enquanto *feat7_soma* considera o total de reincidências.

Para tratar o desbalanceamento, foram testadas proporções de reamostragem 70/30, 60/40 e 50/50. A configuração 50/50 apresentou melhor desempenho, com ganhos em *recall* e *F1 score*, indicando maior capacidade de identificar casos positivos. Todos os experimentos incluíram otimização de hiperparâmetros e ajuste de *threshold*. O PCA foi avaliado, mas não melhorou o desempenho, sugerindo que as variáveis selecionadas já capturam a informação relevante.

A Tabela 2 apresenta os melhores resultados. Em geral, o desempenho dos modelos depende fortemente da engenharia de atributos e da parametrização das métricas de carga. A superioridade da janela 10/40 pode estar associada à melhor captura do histórico de carga dos atletas, com menor influência de variações de curto prazo, aspecto relevante em contextos de alta variabilidade, como o futebol profissional.

Tabela 2. Melhores resultados obtidos

Variáveis	Acurácia	Precisão	Recall	F1 score	AUC-ROC	Sampling
<i>feat7_bin</i>	99,8%	100%	87,5%	93,3%	93,8%	50%
<i>feat7_bin</i>	99,7%	93,3%	87,5%	90,3%	93,7%	70%
<i>feat7_soma</i>	99,7%	93,3%	87,5%	90,3%	93,7%	60%
<i>feat7_soma</i>	99,6%	87,5%	87,5%	87,5%	93,6%	70%
<i>feat7_bin</i>	99,6%	87,5%	87,5%	87,5%	93,6%	60%

5. Considerações Finais

Este trabalho evidenciou o potencial do aprendizado de máquina, com ênfase na métrica ACWR, para a predição de lesões no futebol profissional. O XGBoost destacou-se como o algoritmo de melhor desempenho, e os modelos obtiveram resultados expressivos, configurando uma ferramenta promissora para a gestão da carga e prevenção de lesões.

A análise confirmou que, com pré-processamento e validação adequados, múltiplas variáveis de carga externa podem fornecer informações preditivas relevantes, mesmo a partir de dados de um único clube, embora isso possa limitar a generalização dos modelos. A combinação de atributos derivados de GPS com algoritmos bem ajustados alcançou desempenho comparável ao da literatura. Além de validar o ACWR como feature central, o estudo indica que modelos preditivos podem apoiar decisões sobre carga e recuperação, reduzindo lesões e otimizando o desempenho dos atletas.

Os resultados reforçam a importância de abordagens multidisciplinares que integrem ciência de dados, fisiologia e preparação física. Este trabalho abre caminho para estudos futuros com novas variáveis, contextos competitivos e técnicas avançadas, como redes neurais, sempre respeitando a ética e a privacidade dos atletas.

Uso de Inteligência Artificial

As ferramentas de Inteligência Artificial generativa foram utilizadas apenas para revisão linguística, com foco na clareza textual. Os autores analisaram todas as sugestões e permanecem totalmente responsáveis pelo conteúdo do artigo.

Referências

- Andrade, R. et al. [2020]. “Is the Acute: Chronic Workload Ratio (ACWR) Associated with Risk of Time-Loss Injury in Professional Team Sports? A Systematic Review of Methodology, Variables and Injury Risk in Practical Situations”. Em: *Sports Medicine* 50.9, pp. 1613–1635.
- Bengtsson, H. et al. [2013]. “Muscle injury rates in professional football increase with fixture congestion: An 11-year follow-up of the UEFA Champions League injury study”. Em: *British Journal of Sports Medicine* 47.12, pp. 743–747.
- Bowen, L. et al. [2020]. “Spikes in acute:chronic workload ratio (ACWR) associated with a 5–7 times greater injury rate in English Premier League football players: a comprehensive 3-year study”. Em: *British Journal of Sports Medicine*, pp. 731–738.
- Carey, D. et al. [2016]. “Training loads and injury risk in Australian football-differing acute: Chronic workload ratios influence match injury risk”. Em: *British Journal of Sports Medicine*.
- Delecroix, B. et al. [2018]. “Workload and non-contact injury incidence in elite football players competing in European leagues”. Em: *European Journal of Sport Science* 18.9, pp. 1280–1287.
- Di Credico, A. et al. [2021]. “The prediction of running velocity during the 30–15 intermittent fitness test using accelerometry-derived metrics and physiological parameters: A machine learning approach”. Em: *Intern. J. of Environmental Research and Public Health*.
- Dwyer, D. B. e Gabbett, Tim J. [2012]. “Global positioning system data analysis: velocity ranges and a new definition of sprinting for field sport athletes”. Em: *Journal of Strength and Conditioning Research* 26.3, pp. 818–824.
- Eetvelde, H. V. et al. [2021]. “Machine learning methods in sport injury prediction and prevention: a systematic review”. Em: *Journal of Experimental Orthopaedics* 8.1.
- Gabbett, T. J. [2016]. “The training-injury prevention paradox: should athletes be training smarter and harder?” Em: *British Journal of Sports Medicine* 50, pp. 273–280.
- Jiang, Z. et al. [2022]. “A Systematic Review of the Relationship between Workload and Injury Risk of Professional Male Soccer Players”. Em: *Inter. J. of Environmental Research and Public Health*, p. 13237.
- Malone, J. J. et al. [2015]. “Seasonal training-load quantification in elite English Premier League soccer players”. Em: *Inter. J. of Sports Physiology and Performance*, pp. 489–497.
- Melo, M. et al. [2024]. “Data-Centric AI for predicting non-contact injuries in professional soccer players”. Em: *XXXIX Simpósio Brasileiro de Bancos de Dados*, pp. 167–180.
- Pitka, T. et al. [2023]. “Predicting Injuries in Football Based on Data Collected from GPS-Based Wearable Sensors”. Em: *Sensors* 23.3.
- Rossi, A. et al. [2018]. “Effective injury forecasting in soccer with GPS training data and machine learning”. Em: *PLoS ONE* 13.7.
- Salces, J.N. et al. [2014]. “An examination of injuries in Spanish Professional Soccer League”. Em: *The Journal of sports medicine and physical fitness* 6, pp. 765–771.
- Vallance, E. et al. [2020]. “Combining internal- and external-training-loads to predict non-contact injuries in soccer”. Em: *Applied Sciences* 10.15.