Prédiction des blessures au Foot 5 à l'aide d'une méthode de machine learning.

D. Jacob[1], R. Tievant[2,3], L. Cervoni[1], M Roudesli[4]

[1] Centre de recherche et d'innovation, Talan, Paris
[2] Institut Régional de Médecine du Sport et de la Santé, IRMS², Bois Guillaume
[3] Faculté de Médecine et de Pharmacie, Rouen
[4] Centre de l'appareil locomoteur de l'estuaire, Le Havre

{damien.jacob, laurent.cervoni}@talan.com,
romain.tie@gmail.com, m.roudesli@skeewai.com

Résumé

Le football est un sport d'équipe populaire dans le monde entier, avec plusieurs variantes comme le futsal, le beach soccer et le foot 5 (Fa5). Si le football a fait l'objet d'études approfondies concernant la prédiction des blessures à l'aide de techniques d'apprentissage automatique, le Fa5 a reçu moins d'attention. Cette étude vise à développer un outil d'apprentissage automatique pour prévenir les blessures dans la pratique du Fa5 en comprenant les facteurs contribuant aux blessures. Les données anthropométriques, sportives et relatives aux blessures ont été recueillies auprès de 1827 participants par le biais d'un questionnaire. L'algorithme employé est un random forest, ce qui a permis de déterminer l'importance des variables en amont de la phase d'optimisation. Nous avons ajusté le seuil du modèle pour augmenter la recall aux faux négatifs, qui sont plus problématiques que les faux positifs dans la détection des blessures. Le modèle final a atteint un rappel de 0,70 et une accuracy de 0,65, les attributs physiques et la durée de l'entraînement étant identifiés comme des caractéristiques importantes.

Mots-clés

Foot 5, machine learning, prévention, blessures.

Abstract

Football is a popular team sport worldwide, with several variants, including Futsal, beach soccer, and five-a-side football (Fa5). While football has been extensively studied regarding injury prediction using machine learning techniques, Fa5 has received less attention. This study aims to develop a machine learning tool to prevent injuries in Fa5 practice by understanding the factors contributing to injuries. We collected data from 1827 participants through a questionnaire survey, including physical attributes, practice habits, and injury details. Using a tree-based approach, we reduced the dataset's dimension and selected a decision tree ensemble method, the random forest, to achieve better regularization. After cross-validation, we adjusted the model's threshold cutoff to increase sensitivity to false negatives, which are more problematic than false positives in injury detection. The final model achieved a recall of 0.70 and an accuracy of 0.65, with physical attributes and duration of practice identified as important features.

Keywords

Five-a-side-football, machine learning, prevention, injuries.

1 Introduction

Le football est l'une des activités physiques les plus pratiquées dans le monde. En 2006, la FIFA (Fédération Internationale de Football Association) recensait 265 millions de footballeurs dont 226 millions de non-licenciés [1], tous pays confondus. Il existe cependant de nombreuses variantes qui découlent du football traditionnel : le futsal, le beach soccer, le football à 5, le tennis-ballon, le jorkyball, le foot fauteuil, le cécifoot, etc.
Nous nous intéressons dans cette étude à une pratique qui tend à se développer en France : le Foot 5 aussi appelé couramment « five » ou « urban ». D'après la FFF (Fédération Française de Football), il existe à l'heure actuelle 250 centres de Foot 5 soit 1000 terrains. Il est estimé qu'en France, il y aurait à ce jour entre 2 millions et 3 millions [2] de joueurs réguliers de Foot 5. Comme son nom l'indique, le Foot 5 se joue à 5 contre 5.
L'épidémiologie des blessures dans le Foot 5 est peu décrite. Il paraît donc intéressant de définir des "profils" de sportifs à risque de blessures, afin de pouvoir réaliser de la prévention auprès de ces patients.
L'usage de l'intelligence artificielle et notamment l'approche « Machine Learning » (ou apprentissage automatique) connaît un intérêt croissant dans la littérature pour développer des modèles qui prédisent le risque de blessure [3]. Un modèle de Machine Learning (ML) est un outil informatique qui, à partir d'une liste de variables d'entrées décrivant un participant, va pouvoir fournir une probabilité de risque de blessure pour ce participant [4], et c'est à partir de cette probabilité que la décision sera prise de qualifier si le participant se blessera ou non. Le machine learning peut être utilisé comme un outil d'aide à la prise de décision grâce aux probabilités calculées, modulo une valeur seuil, mais il peut aussi être un outil

capable d'identifier plus efficacement des variables d'entrées importantes dans la prédiction de blessure au football. Plusieurs études ont proposé des approches de ML [5–7], qui ont permis de mettre en évidence des variables d'entrées pertinentes comme la qualité du sommeil, l'historique des blessures, l'âge du pic de croissance, la taille, la longueur des jambes, pourcentage de masse grasse.
Devant l'émergence de cette pratique du football, il est intéressant de connaître les facteurs de risques de blessures. Une fois ceux-ci connus, cela permettrait de diminuer le nombre de blessures en lien avec cette pratique. Nous avons essayé de développer un modèle prédictif de blessure, à l'aide des algorithmes de machine learning sur la base des données recueillies par les travaux de *Tiévant* [8].

2 Matériel et méthode

2.1 Les données

Nous avons réalisé une étude épidémiologique observationnelle, transversale et rétrospective. L'étude a été conduite de manière multicentrique en France. Les patients ont été recrutés via un auto-questionnaire uniquement disponible sur internet. Les patients pouvaient y accéder via le QR code d'une affiche, installée dans les salles privées de Foot 5, ou par un lien disponible sur les réseaux sociaux. Les réponses ont été collectées via le logiciel LimeSurvey®. Les critères d'inclusion étaient de pratiquer le Foot 5 et d'avoir 18 ans ou plus, en raison de pathologies potentiellement spécifiques aux moins de 18 ans. Le questionnaire a été mis à disposition de tous les participants de Foot 5 pendant 6 semaines, du 11 mai au 22 juin 2022.
Ce questionnaire, anonyme, comportait des informations générales sur le sportif :

- Des données générales et anthropométriques : sexe, âge, taille, poids.
- Des données sur l'hygiène de vie : tabagisme, consommation d'alcool, hydratation quotidienne, alimentation.
- Des données sur la pratique sportive : pratique antérieure de football et niveau de pratique, autres sports pratiqués.
- Des données précises sur la pratique du Foot 5 : nombre de séances par semaine, temps de pratique, renforcement musculaire adapté, échauffement et type d'échauffement, le poste joué.

Dans un second temps, en cas de blessure survenue lors des 12 derniers mois, la description de cette blessure a été demandée:

- Le type de blessure, la localisation, le mécanisme, le poste joué au moment de la blessure, l'échauffement pratiqué avant la séance, le moment de la blessure, la consommation d'alcool avant la séance.
- Le diagnostic lié à la blessure : nécessité de consultation aux urgences, si le diagnostic a été posé par un professionnel de santé (et si oui, quel est son métier), avec quels examens complémentaires.
- La prise en charge de cette blessure : médicale, chirurgicale ou sans prise en charge adaptée.
- Les conséquences de cette blessure : la durée d'arrêt de la pratique (permettant de définir la sévérité de la blessure), la durée d'un éventuel arrêt de travail, les séquelles actuelles.
- Le nombre de blessures totales lors des 12 derniers mois (si les sportifs se blessaient plusieurs fois, ils ne pouvaient décrire qu'une seule blessure).

2.2 Définition de la blessure

Nous nous sommes basés sur la définition proposée par Fuller et al. [9] pour étudier les blessures dans le Foot 5. La blessure est définie comme : « toute plainte physique exprimée par un joueur, à la suite d'un entraînement ou d'un match de football, indépendamment du besoin de soins médicaux et de la durée où la pratique du football n'était pas possible. » [9].

2.3 Cible, métriques et traitement des données

Avant d'élaborer un modèle de machine learning il est primordial de définir la cible (ou « target » en anglais) et les métriques qui permettront d'apprécier la qualité des prédictions du modèle. La cible de cette étude est la prédiction du risque de blessure d'un participant avant de pratiquer sa séance de Foot 5. Les métriques utilisées sont l'accuracy, le *recall* et la précision et l'AUC (« Area Under the Curve") qui permettra de comparer différents modèles de classification.
Le jeu de données est composé de 1827 lignes (les participants) et de 78 colonnes. Cependant sur ces 78 colonnes, seules 46 décrivent les participants et ont été utilisée pour l'entrainement du modèle, les 32 autres colonnes ne concernent que la blessure et elles n'ont pas été utilisées dans l'entrainement afin d'éviter le *data leakage*.
L'entraînement est fait selon la méthode *cross validation* avec 5 *fold*. Cette méthode consiste à subdiviser le groupe de données d'entraînement en 5 sous-groupes. Le modèle va être entraîné sur 4 sous-groupes puis va être validé sur le 5ème sous-groupe. Cette opération est répétée 5 fois, de sorte que chaque sous-groupe soit utilisé une fois pour l'évaluation [10]. L'étape finale de l'entraînement calcule la moyenne des métriques pour chacune des 5 validations.
Un échantillon de 80% des participants de l'étude sont choisis aléatoirement pour construire le groupe d'entraînement, les 20% restants constituent le groupe test. Le groupe d'entraînement va servir à optimiser un modèle de machine learning. Le groupe de test ne sera utilisé que pour évaluer les performances du modèle, sur des données qu'il n'a jamais vues, en utilisant le modèle optimiser sur l'échantillon de 80% des patients.

2.4 L'algorithme

Étant donné que nous disposons d'un jeu de données labélisés, nous nous sommes concentrés sur des approches de machine learning supervisé. Les algorithmes de machine learning supervisé les plus courant dans la prédiction des blessures [3] sont les arbres de décision, les régressions logistiques binaire, le support vector machine et les modèles ensemblistes [11] tel que le random forest (RF). Après un premier benchmark des différents algorithmes (Tableau 1), le RF est le modèle qui obtenu le score AUC le plus élevé, c'est pourquoi nous avons retenu ce dernier. L'algorithme de RF est une méthode d'apprentissage automatique supervisé pour la classification et la régression. Il s'agit d'une extension de la méthode d'arbre de décision qui combine plusieurs arbres de décision pour améliorer la performance des prédictions. Le principe de base

de l'algorithme de RF est de construire un grand nombre d'arbres de décision indépendants et de combiner leurs prédictions pour obtenir une prédiction finale. Les étapes principales de l'algorithme de RF sont :

- La sélection aléatoire d'un échantillon avec remplacement (bootstrap) des données d'entraînement.
- La sélection aléatoire d'un sous-ensemble de variables d'entrée (caractéristiques) pour chaque arbre de décision.
- La Construction d'un arbre de décision pour chaque échantillon bootstrap et chaque sous-ensemble de variables d'entrée.
- L'Agrégation des prédictions de tous les arbres de décision pour obtenir une réponse finale. Pour la classification, la réponse finale est obtenue par vote majoritaire.

Tableau 1 : Valeurs moyennes et écarts types des différentes métriques en pourcentage obtenues lors de l'entraînement du modèle selon la méthode cross validation.

Modèle	*Recall*	Précision	AUC	*Accuracy*
RF	$71{,}1_{(+/- 4)}$	$61{,}9_{(+/- 9)}$	70,6	$66{,}1_{(+/- 9)}$
XGBoost	$67{,}6_{(+/- 5)}$	$61{,}2_{(+/- 9)}$	70,5	$65{,}3_{(+/- 8)}$
SVM	$74_{(+/- 3)}$	$60{,}8_{(+/- 8)}$	69,3	$64{,}4_{(+/- 7)}$
Arbre de décision	$74{,}4_{(+/- 4)}$	$58_{(+/- 8)}$	68,2	$59{,}8_{(+/- 7)}$

3 Résultats

3.1 Entrainement du modèle

L'entraînement et l'optimisation du modèle aboutissent à un *accuracy* de 66,1 % avec un *recall* de 71,1 % (Tableau 1). Ces premiers résultats sont encourageants mais l'utilisation de l'outil, à des fins de prévention, doit être plus sensible à la détection de faux négatifs. Dans cette optique, il est possible de modifier, pendant la phase d'entrainement, la valeur seuil qui différencie une prédiction négative, d'une prédiction positive. Ainsi pour le modèle Random Forest le seuil a été modifié manuellement pour diminuer la prédiction de faux négatifs. Cependant cette action a pour conséquence de diminuer la précision ce qui s'accompagne d'une augmentation de la prédiction de faux positifs (Tableau 2).

Tableau 2 : Valeurs des différentes métriques en pourcentage obtenues lors de l'entraînement avec une nouvelle valeur seuil.

Modèle	*Recall*	Précision	AUC	*Accuracy*
Random Forest Valeur seuil de 0,425	85,9	58,2	70,6	64,7

Avec l'ajustement de cette nouvelle valeur seuil égal à 0,425 au lieu de 0,475, le *recall* du Random Forest a été augmentée de 14,8 point pour une diminution de 1,45 point d'*accuracy*, mais comme anticipé une diminution de la précision de 6%.

3.2 Test du modèle

L'objectif est de tester le modèle précédemment optimisé sur des données inconnues et ainsi mesurer la qualité des prédictions. Ainsi, nous avons testé les 2 modèles de Random Forest précédemment présentés sur les 20% de patients non connus par l'algorithme.

Tableau 3 : Valeurs des différentes métriques en pourcentage obtenues lors du test du modèle.

Modèle	*Recall*	Précision	AUC	*Accuracy*
Random Forest Valeur seuil de 0,475	83,3	54,2	69,2	60,0

On constate une légère dégradation de le *recall* entre l'entraînement et le test, mais la précision est le paramètre le plus fortement dégradé. La dégradation combinée de ces deux métriques se traduit par une diminution de l'*accuracy* passant de 64,9% à 60,0% (Tableau 3).

Tableau 4 : Matrice de confusion obtenue lors de la phase de test.

Matrice de confusion		Prédiction	
		0	1
Observation	0	79	118
	1	28	140

La matrice de confusion (Tableau 4) nous montre que le nombre de prédictions correctes s'élève à 219 sur un total de 365. On dénombre seulement 28 faux négatifs, autrement dit seuls 28 blessés n'ont pas été identifiés par le modèle mais cela se fait au détriment de la détection de 118 faux positifs.

4 Discussion

Nous n'avons pas identifié d'études ayant développé un modèle de machine learning pour prédire le risque de blessures au foot 5. Nous avons donc comparé notre étude aux résultats produits dans le football, en lien avec une approche machine learning.

On peut noter que la taille des échantillons des autres études varie de 86 à 952 participants (Tableau 5). Nous proposons une analyse avec une taille d'échantillon plus importante constituée de 1827 participants. Cette différence de taille d'échantillon peut s'explique par le fait que les joueurs de football des différentes études évoluent au niveau professionnel alors que notre approche repose sur des participants pouvant être débutants, amateurs ou professionnels.

Nos métriques sont dans le même ordre de grandeur que celles obtenues par Oliver et al. [12]. Nous retrouvons également un modèle dont le *recall* est supérieur à la précision. Quelques variables d'entrées sont communes à nos deux études comme le poids, l'âge et l'IMC. Cependant Oliver el al. [12] utilisent d'autres variables d'entrées que nous n'avons pas été en mesure de recueillir, comme des mesures neuromusculaires (saut de contre-mouvement unipodal, pic de force verticale de réaction au sol) et des données anthropométriques supplémentaires (comme la longueur des jambes).

Les études de Ayala et al. [7] et Rommers et al. [6] obtiennent des résultats plus précis. Même si notre *recall* est du même ordre de grandeur que les 3 études précédemment, leur précision est nettement supérieure. Nous avons identifié 2 facteurs qui expliqueraient cette différence :

- L'utilisation de certaines variables d'entrées plus pertinentes dans la prédiction des blessures. Pour Rommers et al. [6], la variable d'entrée la plus importante est l'âge du pic de croissance. Ayala et al. [7] mettent en évidence l'impact de la qualité du sommeil sur le risque de blessure, ce facteur aggravant a déjà été cité par Cresswell et Eklund [13]. Et pour finir, une dernière variable semble importante dans l'étude réalisée par Ayala et al. [7], qui est l'historique des blessures du participant, ce qui est cohérent avec la littérature [14,15].
- La blessure à identifier peut, si elle est très spécifique, influencer les performances du modèle. Comme le souligne Ayala et al. [7], "Perhaps, the fact that the current study [7] built on injury-specific predictive model might explain the slightly better predictive performance results obtained in comparison with the non-specific injury risk model". Dans notre étude la blessure n'est pas spécifique à une localisation ou à un type d'atteinte. Cela peut donc expliquer, en partie, nos moins bonnes performances.

Il convient de noter que cette étude sur la réalisation d'un modèle prédictif des blessures au foot 5 a quelques limites. Tout d'abord, la population étudiée n'est probablement pas représentative de l'ensemble des pratiquants de foot 5. En effet, les participants à l'étude sont principalement des étudiants (32,3%), des employés (24,7%) et des cadres et professions intellectuelles supérieures (22,7%). De plus, le modèle développé ne dépasse pas l'existant en termes de métriques (Tableau 5). Néanmoins, l'amélioration de notre modèle doit passer par une augmentation de la précision, cela permettra de diminuer la proportion de faux positifs. Nous suggérons deux axes pour l'amélioration des performances du modèle. En premier lieu, nous suggérons une augmentation de la taille de l'échantillon car les algorithmes de machine learning peuvent gagner en *accuracy* avec une augmentation de la taille de la base de données [16]. De plus, une augmentation de la base de données permettra d'envisager des algorithmes plus précis comme les réseaux de neurones et autre forme de deep learning. Deuxièmement, en s'appuyant sur les méthodes utilisées dans d'autres analyses sur le football, nous pouvons envisager des pistes d'amélioration pour une future étude : nous pourrions questionner les participants sur la qualité de leur sommeil, être plus précis quant à l'historique de leur blessure ou affiner les données anthropométriques.

Cette étude peut servir pour développer une application à destination des sportifs pour les sensibiliser aux risques de blessures et les prévenir en amont avec une consultation médicale dédiée. Elle permet également de sensibiliser les professionnels de santé aux facteurs de risque de blessures et de promouvoir des stratégies de prévention pour réduire l'incidence de ces blessures chez les joueurs de foot 5.

5 Conclusion

Grâce aux données récoltées par l'étude de Tievant [8] sur la pratique du Foot 5, un modèle prédictif a pu être développé et testé. Le modèle repose sur un algorithme de machine learning supervisé : le Random Forest. Ce modèle est constitué de plusieurs centaines d'arbres de décision qui vont participer pour prédire le risque de blessure d'un participant en fonction de certaines informations d'entrées. Dans notre cas, ce sont les informations sur le temps de jeu et quelques caractéristiques anthropométriques qui participent à fournir les meilleures prédictions. Le modèle se révèle être précis à 64,9% avec une précision de 60,1% et un *recall* de 70,3%. Ce qui en fait un modèle assez efficace pour ne pas prédire trop de faux négatifs.

Cette approche prédictive avec un algorithme de machine learning est nouvelle, à notre connaissance, pour le foot 5. Cependant il existe plusieurs approches similaires dans le football. En comparant nos résultats avec la littérature, deux pistes d'améliorations sont envisageables. D'une part l'augmentation de la taille de l'échantillon qui permettrait de gagner en *accuracy* grâce à l'utilisation d'algorithmes comme les réseaux de neurones et d'autre part le questionnement des participants sur la qualité de leur sommeil, l'historique de leur blessure et affiner les données anthropométriques. Ce questionnaire [8], dans une étude de plus grande envergure, va donc pouvoir évoluer afin d'améliorer notre compréhension et la prévention des blessures dans le foot 5.

Tableau 5 : Comparatif des différentes études sur l'application de modèle de machine learning dans la prédiction de la blessure au football.

Nom de l'étude	*Recall*	Précision	AUC	Taille de l'étude	Variables d'entrées importantes
Rossi et al., 2018	80,0	87,0	Non communiqué	952	Historique des blessures, vitesse de course, distance parcourue
Ayala et al., 2019	77,8	83,3	87,3	86	Qualité du sommeil
Oliver et al ., 2020	74,2	55,6	66,3	355	Données anthropométriques (poids, âge, IMC)
Rommers et al., 2020	85	85	Non communiqué	734	L'âge du pic de croissance, la taille, la longueur des jambes, pourcentage de masse grasse.
Notre modèle	83,3	54,2	69,2	1827	Fréquence de jeu, âge, IMC, temps de jeu

6 Références

1. FIFA. FIFA Big Count 2006: 270 million people active in football [Internet]. FIFA Communication divisions; 2007 mai [cité 8 nov 2022]. Disponible sur: https://digitalhub.fifa.com/m/55621f9fdc8ea7b4/original/mzid0qmguixkcmruvema-pdf.pdf

2. La mode du Five, une pratique en pleine expansion [Internet]. L'Équipe. [cité 16 sept 2022]. Disponible sur: https://www.lequipe.fr/France-Football/Actualites/La-mode-du-five-une-pratique-en-pleine-expansion/1331413

3. Van Eetvelde H, Mendonça LD, Ley C, Seil R, Tischer T. Machine learning methods in sport injury prediction and prevention: a systematic review. J Exp Orthop. 14 avr 2021;8(1):27.

4. Hastie T, Friedman J, Tibshirani R. The Elements of Statistical Learning [Internet]. New York, NY: Springer; 2001 [cité 2 avr 2023]. (Springer Series in Statistics). Disponible sur: http://link.springer.com/10.1007/978-0-387-21606-5

5. Rossi A, Pappalardo L, Cintia P, Iaia FM, Fernàndez J, Medina D. Effective injury forecasting in soccer with GPS training data and machine learning. PLOS ONE. 25 juill 2018;13(7):e0201264.

6. Rommers N, Rössler R, Verhagen E, Vandecasteele F, Verstockt S, Vaeyens R, Lenoir M, D'Hondt E, Witvrouw E. A machine learning approach to assess injury risk in elite youth football players. Med Sci SPORTS Exerc. 2020;52(8):1745-51.

7. Ayala F, López-Valenciano A, Martín JAG, Croix MDS, Vera-Garcia FJ, García-Vaquero M del P, Ruiz-Pérez I, Myer GD. A Preventive Model for Hamstring Injuries in Professional Soccer: Learning Algorithms. Int J Sports Med. mai 2019;40(5):344-53.

8. Tievant R. Étude épidémiologique des blessures liées à la pratique du foot 5 : identification des profils sportifs à risque. Rouen: UFR de sante de Rouen Normandie; 2022 p. France.

9. Fuller CW, Ekstrand J, Junge A, Andersen TE, Bahr R, Dvorak J, Hägglund M, McCrory P, Meeuwisse WH. Consensus statement on injury definitions and data collection procedures in studies of football (soccer) injuries. Scand J Med Sci Sports. 2006;16(2):83-92.

10. 3.1. Cross-validation: evaluating estimator performance [Internet]. scikit-learn. [cité 27 sept 2022]. Disponible sur: https://scikit-learn/stable/modules/cross_validation.html

11. Géron A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media, Inc.; 2022. 878 p.

12. Oliver JL, Ayala F, De Ste Croix MBA, Lloyd RS, Myer GD, Read PJ. Using machine learning to improve our understanding of injury risk and prediction in elite male youth football players. J Sci Med Sport. 1 nov 2020;23(11):1044-8.

13. Cresswell SL, Eklund RC. The Nature of Player Burnout in Rugby: Key Characteristics and Attributions. J Appl Sport Psychol. 1 sept 2006;18(3):219-39.

14. Fousekis K, Tsepis E, Poulmedis P, Athanasopoulos S, Vagenas G. Intrinsic risk factors of non-contact quadriceps and hamstring strains in soccer: a prospective study of 100 professional players. Br J Sports Med. 1 juill 2011;45(9):709-14.

15. Hägglund M, Waldén M, Ekstrand J. Injury incidence and distribution in elite football—a prospective study of the Danish and the Swedish top divisions. Scand J Med Sci Sports. 2005;15(1):21-8.

16. Brill E, Lin J, Banko M, Dumais ST, Ng AY. Data-Intensive Question Answering. In: TREC. 2001. p. 90.