Domain Review

From Jumps to Signals:

Selecting Countermovement Jump Features for Injury-Risk Classification

Jordyn Maurer

21 September 2025

Seminar in Data Analytics

I. Introduction

As previously explored, injury prevention matters for both athlete wellbeing and team availability. If we can reduce injuries, we can keep athletes healthier, stronger, and more competitive. However, effectively preventing injuries requires detailed insights into an athlete's workload and performance. In response, sports scientists have become increasingly invested in performance-tracking devices to capture and analyze this information. With the rise of these technologies, we now have access to sports data previously unavailable and unexplored. Several devices have been invented to track training in athletes, including force plates. Bourdon et al. (2017) highlights that although many of these devices have shown strong validity and reliability, applying them in real-world settings remains challenging. For example, one countermovement jump (CMJ) on the force plate results in approximately seventy distinct metrics, each spanning weighing, unweighting, braking (eccentric), propulsive (concentric), flight, and landing phases of the jump (Hawkin Dynamics Metric Database, 2022). Many of these variables are highly correlated and vary by athlete or sport, making them difficult to directly apply to athlete performance tracking. This creates a specific need for determining which subset of CMJ force plate metrics are most informative for classifying soft-tissue injury risk in athletes. Previous work has shown that machine learning models can predict injury from GPS and heart-rate training-load features, suggesting a similar approach could be applied to selecting CMJ force plate metrics for classification (Lövdal et al., 2021; Tsilimigkras et al., 2024). Further, the following domain review (i) establishes a foundational understanding of the CMJ and its associated metrics, (ii) examines strategies for handling imbalance datasets where injury cases are rare, and (iii) explores relevant design methods for predicting injury risk.

II. The Countermovement Jump

To better understand the study, we first begin examining the CMJ. Schuster et al. (2020) describe the CMJ as the most detailed force plate test with the best athlete compliance, making it well-suited for frequent testing. Additionally, their study suggests that force plates offer versatile, fast, and simple solutions to monitoring athlete injury risk given a full understanding of each output metric. While the study does not define any methods for proactive injury prevention, it applies specific use-cases of the CMJ in NBA player monitoring, rehabilitation, and benchmarking which we can apply to our foundational knowledge of the CMJ metrics. Further, to make these outputs actionable in the context of injury prevention, we will organize the metrics by CMJ phase to pinpoint those most tied to injury risk. For CMJ phases, we reference McMahon et al. (2018) which facilitates the understanding and application of the exhaustive list of metrics and provides our foundation for selecting features that best classify injury risk. The first part of the jump, the weighing phase, calculates the athlete's body weight and acts as the foundation for many of the metrics in later stages of the jump (Hawkin Dynamics, 2021). However, because this phase does not require athlete movement, it will be excluded from the classification of injury risk. Although McMahon et al. do not specify which phases may be most relevant to injury prediction, structuring the data in this way creates a foundation to explore these potential relationships systematically. Because each jump produces a high number of interrelated metrics, determining which variables are most relevant to injury risk becomes a significant challenge. We must employ feature selection techniques to remove irrelevant, noisy predictors that may contribute to overfitting and inaccuracy of our model (Yan & Zhang, 2015). By leveraging these strategies along with foundational knowledge of the CMJ metrics, we can ensure our feature selection process is both biomechanically informed and statistically supported.

III. Handling Imbalanced Data

As Lövdal et al. explain in their injury prediction model in competitive runners, when we split observations into “injured” and “not injured,” the injured class becomes an extreme minority. Because of this phenomena, the overall accuracy of our model becomes misleading. A model can predict “no injury” most of the time and still appear to perform well. In order to avoid bias toward the non-injured class, we must train our machine learning classifier on a balanced dataset (Lövdal et al., 2021). Depending on the desired machine learning model, researchers have utilized several different methods for balancing data. Referring back to Lövdal et al., their model implements a balanced bagging approach in which they create multiple balanced subsets of the training data, fit a model to each, and average the predicted probabilities. Because sports-injury prediction is still an emerging subject, we will also draw on methodology from traffic-crash mortality prediction, a domain that faces similarly imbalanced outcomes. While the following references also utilize bagging approaches, we will also explore two other methods: under-sampling and over-sampling.

While comparing mortality prediction models for road traffic accidents, Boo and Choi (2022) reflect that imbalanced classification becomes increasingly difficult when compounded by factors such as dataset size, label noise, and data distribution: all of which will be present in our sports-injury data. To address these problems, we must consider re-sampling techniques to balance the data. In classifying motor vehicle crash injury severity, Jeong et. al (2018) reference only 0.34% of their records pertaining to fatal accidents and suggest handling with under-sampling or over-sampling. In another road crash severity prediction model, Fiorentini and Losa (2020) point to Synthetic Minority Oversampling Technique (SMOTE) which takes each minority class and creates new instances using k-nearest neighbors and bootstrapping. Undersampling, in contrast, balances datasets by reducing the number of samples of the majority class. While SMOTE can handle both continuous and categorical features, results generally become overfitted with limited generalizability, a feature that may negatively affect our ability to predict sports injuries (Fiorentini & Losa, 2020). Despite the possible negatives, Boo and Choi (2022) as well as Jeong et. al (2018) noted predictions with samples using SMOTE produced the best results. Furthermore, in order to better choose a resampling approach fit for our sports injury data, we must also explore potential machine learning models to employ.

IV. Design Methods

Given the binary nature of the dependent variable (injury vs. no injury), we first examine logistic regression as an initial approach for modeling and predicting injury occurrence. Gabbett (2010) analyzed noncontact, soft-tissue injuries in athletes as well as training load data using a logistic regression model and logit link function, finding 62.3% true positive predictions. However, we assume our CMJ metrics are nonlinear and interdependent, meaning logistic regression may not be the proper fit for our model. Further, as described in *An Introduction to Statistical Learning*, logistic regression struggles when the sample size is small relative to the number of predictors, leading to unstable estimates and overfitting: a likely issue with CMJ and injury data (James el al., 2021, p. 143). Because of these limitations, alternative methods are often explored for injury prediction and similar health-related applications. For example, a 2009 study predicting early mortality after variceal bleeding demonstrated the value of other statistical methods beyond logistic regression which have also been applied to predicting both motor vehicle crash mortality rates and sports injuries (Augustin et al., 2009). Moreover, in a 2019 review of sports analytics research, researchers found that "the main AI technique or method used for injury risk assessment and sporting performance prediction was artificial neural network," while "decision tree classifier and support vector machine…were the next mostly used techniques" (Claudino et al., 2019). These findings suggest that tree-based classification models are gaining traction for injury prediction.

In particular, classification and regression trees (CART) have been widely used for modeling injury data, including Fiorentini et al. (2018) and Augustin et al. (2009). CART builds a decision tree by applying a series of rules to the predictor variables, stopping when no further improvement can be made or when a pre-defined stopping criterion is met (Jeong et. al., 2018). This model creates a simple, interpretable baseline, making it a natural starting point for injury prediction models. However, a single decision tree generally does not have the same predictive accuracy as other classification approaches (James el al., 2021, p. 340). Thus, we can increase the complexity of the decision tree in an attempt to produce more accurate predictions with ensemble methods. Further, a random forest (RF) includes multiple decision trees trained on bootstrapped data, determining the final predictions based on averages or majority voting (Boo & Choi, 2022). These prove advantageous as this method decorrelates the trees (James el al., 2021, p. 343). This approach is especially relevant to our dataset, where CMJ metrics are highly correlated and injuries are rare events. By decorrelating the trees, random forests reduce overfitting and improve the prediction accuracy of our model.

We take this approach one step further with gradient boosting, a method that also builds decision trees, but does so sequentially rather than independently. In this approach, one tree is built at a time with each new iteration set to correct the errors of the previous model (Friedman, 2002). This process enhances the predictive power particularly relevant to our dataset with complexly related predictors that outweigh the events of our independent variable. Recent literature has demonstrated a leading implementation of gradient boosting in Extreme Gradient Boosting, or XGBoost. Developed by Chen and Guestrin (2016), XGBoost promises a “a sparsity-aware algorithm” that gives “state-of-the-art” results on complex problems in a wide range of domains. This method was applied to both crash prediction and sports injury prediction. Particularly, XGBoost “is faster than conventional gradient boosting machines and allows a generalized model to be obtained” due to its system optimizations and regularization features (Boo & Choi, 2022). These strengths make it well-suited for predicting rare outcomes like injuries, where both speed and model generalization across athletes are critical. Most relevant to our research, Lövdal et al. (2021) applied XGBoost to injury prediction in competitive runners, remarking on its ease of application, generalizability, and accuracy. In addition to the actual predictive model, their study applied XGBoost to determine the importance of each input feature with respect to the output. Thus, the model guides feature selection, potentially recognizing the most important CMJ metrics for injury risk assessment. However, while XGBoost provides built-in feature selection methods, this process can be further refined with SHAP (SHapley Additive exPlanations). Shap offers a more interpretable approach that calculates each variable’s contribution to model predictions, both at the individual and global level (SAMueL project team, 2022) . When combined with subsetting the CMJ indicators into their specific jump phase, this method allows us to identify the most impactful CMJ metrics while also providing a clear, visual explanation of how these variables influence injury risk. Using SHAP alongside XGBoost ensures our model is both accurate and interpretable, allowing us to better implement results from our research and improve athlete performance and injury prevention.

V. Conclusion

Through this domain review, we have gained an understanding of how CMJ metrics can be structured to predict soft-tissue injuries. By examining previous research on injury prediction, we identified strengths and limitations of existing approaches and adapted methods to our specific, highly imbalanced injury data. While these findings provide a foundation for the remainder of the research, we still must implement and validate the proposed models, particularly paying attention to the generalizability of the model across different sports and athletes. We will continue to update our research as we continue to work with the data, making sure to rely on measures of accuracy when validating our models. Ultimately, our goal is to create a reliable, interpretable tool that supports proactive injury prevention and enhances athlete performance which this foundational knowledge grants us the first step toward achieving.

References

Augustin, S., Muntaner, L., Altamirano, J. T., González, A., Saperas, E., Dot, J., Abu-Suboh, M., Armengol, J. R., Malagelada, J. R., Esteban, R., Guardia, J., & Genescà, J. (2009). Predicting early mortality after acute variceal hemorrhage based on classification and regression tree analysis. *Clinical Gastroenterology and Hepatology, 7*(12), 1347–1354.

Boo, Y., & Choi, Y. (2022). Comparison of mortality prediction models for road traffic accidents: An ensemble technique for imbalanced data. *BMC Public Health, 22*, 1476.

Bourdon, P. C., Cardinale, M., Murray, A., Gastin, P., Kellmann, M., Varley, M. C., Gabbett, T. J., Coutts, A. J., Burgess, D. J., Gregson, W., & Cable, N. T. (2017). Monitoring athlete training loads: Consensus statement. *International Journal of Sports Physiology and Performance, 12*(s2), S2-161–S2-170.

Claudino, J. G., Capanema, D. d., de Souza, T. V., Serrão, J. C., Machado Pereira, A. C., Nassis, G. P., & Mochizuki, L. (2019). Current approaches to the use of artificial intelligence for injury risk assessment and performance prediction in team sports: A systematic review. *Sports Medicine - Open, 5*, 28.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery.

Fiorentini, N., & Losa, M. (2020). Handling imbalanced data in road crash severity prediction by machine learning algorithms. *Infrastructures, 5*(7), 61.

Gabbett, T. J. (2010). The development and application of an injury prediction model for noncontact, soft-tissue injuries in elite collision sport athletes. *Journal of Strength and Conditioning Research, 24*(10), 2593–2603.

Hawkin Dynamics. (2021, August). *The countermovement jump playbook* (eBook v1). Hawkin Dynamics.

Hawkin Dynamics. (2022, November 1). *Hawkin Dynamics metric database*.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer.

Jeong, H., Jang, Y., Bowman, P. J., & Masoud, N. (2018). Classification of motor vehicle crash injury severity: A hybrid approach for imbalanced data. *Accident Analysis & Prevention, 120*, 250–261

Lövdal, S., den Hartigh, R., & Azzopardi, G. (2021). Injury Prediction in Competitive Runners with Machine Learning. International journal of sports physiology and performance, 16(10), 1522–1531.

McMahon, J. J., Suchomel, T. J., Lake, J. P., & Comfort, P. (2018). Understanding the key phases of the countermovement jump force–time curve. *Strength & Conditioning Journal, 40*(4), 96–106.

SAMueL Project Team. (2022). *Explaining XGBoost model predictions with SHAP values.* GitHub Pages. https://samuel-book.github.io/samuel-2/samuel_shap_paper_1/xgb_with_feature_selection/03_xgb_combined_shap_key_features.html

Schuster, J., Bove, D., & Little, D. (2020). Jumping towards best-practice: Recommendations for effective use of force plate testing in the NBA. *Sports Performance Science Reports, 68*, 1–4.

Tsilimigkras, T., Kakkos, I., Matsopoulos, G. K., & Bogdanis, G. C. (2024). Enhancing sports injury risk assessment in soccer through machine learning and training load analysis. *Journal of Sports Science and Medicine, 23*(5), 537–547.

Yan, K., & Zhang, D. (2015). Feature selection and analysis on correlated gas sensor data with recursive feature elimination. *Sensors and Actuators B: Chemical, 212*, 353–363.