From Jumps to Signals: Selecting Countermovement Jump Metrics for Injury-Risk Classification

Jordyn Maurer
Denison University

December 11, 2025

Abstract

Athlete health and performance depend on detecting early signs of injury risk. Countermovement Jump (CMJ) force plate testing is a reliable, repeatable way to track athlete performance, but correlated metrics, limited interpretability, and few predictive studies creates a challenge when turning these tests into actionable insights. Using Hawkin force plate data from the 2024-2025 academic year, this study integrated Lasso feature selection, XGBoost modeling, and SHAP values to identify CMJ metrics most associated with injury risk among Denison University athletes. After reducing the feature set with Lasso, XGBoost achieved AUC = 0.836, precision = 0.296, recall = 0.450 at the optimal F1 threshold (0.20). Six of the top ten most influential features came from the propulsive phase of the jump. Although injury prediction remains difficult, meaningful signal patterns exist in CMJ data, bridging the gap between athlete monitoring and data-driven performance modeling.

Contents

1 Introduction

Injury prevention matters for both athlete well being and performance. Reducing injuries keeps 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. Among these devices, the force plate is an instrument that allows coaches to detect and measure the force produced during physical exercises (Hawkin Dynamics, 2021). The device is often paired with the countermovement jump (CMJ), which is a vertical jump that involves a downward movement (countermovement) before an upward leap (Walker, 2025). Bourdon et al. (2017) highlight that although many of these devices have shown strong validity and reliability, applying them in real-world settings like injury prediction remains challenging and relatively unexplored. For example, one CMJ 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 injury risk in athletes. Therefore, the goal of this study was to determine whether CMJ force plate data can be used to accurately identify athletes at elevated risk of injury. Using data collected from Division III athletes at Denison University, we applied regularization and gradient-boosting models to classify injured and non-injured athletes and identify the most informative force-time variables. The results of this study aimed to support coaches, strength staff, and sports medicine professionals in making data-informed decisions to reduce injury risk. At its core, this research was driven by a commitment to keeping athletes healthy and able to compete.

2 Ethical Considerations

Because this project used real athlete performance and injury data, there were important ethical considerations to address to ensure the privacy and well-being of the athletes involved, as well as the responsible interpretation of results. Data was used with institutional approval from Denison University's Sports Performance and Sports Medicine staff. Because all data was originally collected as part of routine performance monitoring, no additional physical testing was imposed on the athletes. To protect the privacy of the athletes, all data was void of personally identifiable information such as names, student IDs, or jersey numbers. Any analysis was focused on aggregated trends, rather than individual athletes. Any predictive outputs from this research are intended to support, not replace, the professional judgment of athletic trainers and medical staff. Injury predictions should not be used to restrict participation or make assumptions about an athlete's commitment or durability. We also recognize the risk of model bias, as force plate metrics can vary across sexes, sports, positions, training backgrounds, and genetic predispositions. Care was taken to interpret results within context rather than generalizing findings to populations beyond the scope of this dataset.

At this stage, we avoided labeling individual athletes as "injury-prone" without sufficient evidence. Ultimately, the health and safety of the athletes remained our top priority, and this research was designed to support and protect it.

3 Domain Review

3.1 The Countermovement Jump

To build foundational knowledge for this study, we first examined the countermovement jump (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 athlete monitoring. Their work also highlights that force plates offer versatile, fast, and simple solutions to monitoring athlete injury risk when the output metrics are properly interpreted. While the study does not propose any methods for proactive injury prevention, it applies specific use-cases of the CMJ in NBA player monitoring, rehabilitation, and benchmarking which helped inform our understanding of how these metrics can be used in practice.

To dive deeper into the individual metrics in our study, we referenced McMahon et al. (2018), which facilitated the understanding and application of the exhaustive list of metrics and provided our foundation for understanding 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 involve active movement, it did not contribute to the classification of injury risk. While McMahon et al. do not specify which phases are most relevant to injury risk, this phase-based structure allowed us to systematically explore how different parts of the jump may relate to pre-injury trends.

Because each jump produces a high number of highly correlated variables, determining which metrics are most relevant to injury risk presents a significant analytical challenge. Therefore, feature selection is critical to reduce noise, mitigate multicollinearity, and prevent overfitting (Yan & Zhang, 2015). By combining biomechanical insight from existing CMJ literature with data-driven feature selection techniques, we aimed to ensure that the predictors in our model were both meaningful and statistically supported.

3.2 Handling Imbalanced Data

In injury prediction research, one of the most common challenges is class imbalance: far fewer observations are labeled as "injured" compared to "not injured." 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 may predict "no injury" for nearly every instance and still appear highly accurate, despite failing to detect true injury cases. In order to address this issue, many studies balance the dataset before training a model. 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 to improve sensitivity to the injured class.

Because sports injury prediction is still an emerging subject, researchers have also turned to methods in other domains with similarly imbalanced outcomes, such as road traffic crash mortality. 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: factors also relevant in sports data. In classifying motor vehicle crash injury severity, Jeong et. al (2018), working with a dataset in which only 0.34% of their outcomes were fatal, recommend the use of under-sampling or over-sampling to address imbalance. Other researches, such as Fiorentini and Losa (2020), employ 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 overfit with limited generalizability, a feature that was explored in our study and negatively impacted significance of our results (Fiorentini & Losa, 2020).

Moreover, although these methods are widely used in injury and mortality prediction studies, our approach differed. In this study, we did not modify the dataset through resampling. Instead, we preserved the original class distribution and addressed imbalance during model training by applying class weighting. The rationale for this choice, and its impact on results, is discussed further in the Methods and Discussions sections.

3.3 Model Selection

Given the binary nature of the dependent variable (injury vs. no injury), many studies first examine logistic regression as an initial approach for prediction modeling. Gabbett (2010) analyzed non-contact, 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, CMJ metrics are nonlinear and interdependent, meaning logistic regression may not be the proper fit for this type of data. Further, as described in *An Introduction to Statistical Learning*, logistic regression struggles when sample sizes are 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 health-related applications (Augustin et al., 2009). Moreover, in a 2019 review of sports analytics research, Claudino et al. (2019) found "the main AI technique or method used for injury risk assessment and sporting performance prediction was artificial neural network," while, "the decision tree classifier and support vector machine... were the next mostly used techniques." 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). To improve this, ensemble methods can be used instead. A random forest (RF), for example, includes multiple decision trees trained on bootstrapped data, determining the final predictions based on averages or majority voting (Boo & Choi, 2022). This approach decorrelates trees, which is especially relevant to CMJ data where metrics are highly correlated and injury cases are rare (James el al., 2021, p. 343). By decorrelating the trees, random forests help reduce overfitting and improve the prediction accuracy.

Gradient boosting takes this approach one step further by building trees sequentially rather than independently. In this method, one tree is built at a time with each new iteration set to correct the errors of the previous model (Friedman, 2002). This process improves predictive power in datasets like ours where predictors are numerous, related, and nonlinear. Recent literature has shown strong performance from Extreme Gradient Boosting, or XGBoost. Developed by Chen and Guestrin (2016), XGBoost promises "a sparsity-aware algorithm" that gives "state-of-the-art" results on complex problems in a wide range of domains. It has been 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. In particular, Lövdal et al. (2021) applied XGBoost to injury prediction in competitive runners, noting its generalizability, ease of use, and accuracy.

Additionally, while XGBoost offers built-in feature selection methods, this process can be further refined using SHAP (SHapley Additive exPlanations). SHAP provides more interpretable insight by calculating each variable's contribution to a prediction at both global and individual levels (SAMueL project team, 2022). Thus, this method helps identify the most impactful CMJ metrics, while also providing a clear, visual explanation of how these variables influence injury risk. Using SHAP alongside XGBoost improves both accuracy and interpretability, allowing for better implementation of results.

3.4 Conclusion

Through this domain review, we developed an understanding of how CMJ metrics can be structured and interpreted to support injury prediction. By examining previous research, we identified both the strengths and limitations of existing approaches and considered how these methods could be adapted to a highly imbalanced dataset such as ours. This review provided the foundation for the remainder of the research, particularly in guiding decisions about model selection, feature reduction, and evaluation metrics. While prior literature offers valuable direction, it also highlights gaps, such as limited generalizability across sports, small injury sample sizes, and inconsistent use of force plate metrics, that our study seeks to address. The next stage of this research, therefore, involved implementing and evaluating predictive models using our dataset, with specific attention to interpretability, class imbalance, and practical usability for coaches and sports performance staff.

4 Methods

4.1 Research Overview

The purpose of this study was to identify key countermovement jump (CMJ) force plate metrics that best classify injury risk in collegiate athletes. The research used secondary data from Denison University Strength and Conditioning. The study combined two datasets representing repeated measurements of student athletes from the 2024-2025 academic year. One dataset included information about individual athlete injury status while the other dataset provided individual performance metrics from the CMJ on the Hawkin Dynamics force plate. The analysis involved two main components: (1) exploratory analysis to narrow down potential predictors into a smaller subset, and (2) predictive analysis to classify athletes at greater risk of soft-tissue injuries and interpret the key features involved. The dependent variable was binary (0 = not injured, 1 = injured). Potential independent variables and covariates included over seventy numerical CMJ metrics per jump. Barring any missing data, the final sample size of injured and non-injured athletes included 23,395 jumps.

Because this study relied on secondary, observational data, it was non-experimental in nature. Therefore, we could not determine whether specific CMJ metrics caused athlete injuries. Instead, our analysis was limited to identifying associations between performance metrics and injury occurrence. We could not establish direct causation between a specific metric and injury, but our findings may highlight potential risk factors to guide intervention strategies. Thus, our results were interpreted as evidence of statistical relationships rather than proof that changes in specific CMJ metrics will produce injuries.

Further, to explore the relationship between the CMJ and injuries, we employed several different methods of analysis. We primarily explored Extreme Gradient Boosting (XGBoost) with SHAP values for prediction and feature interpretation (SAMueL project team, 2022). For feature selection, we implemented Lasso regularization to reduce the large number of CMJ variables to a subset that best predicts injury risk (James el al., 2021, p. 242). These methods allowed us to balance predictive accuracy with interpretability, ensuring that the most influential CMJ variables were retained for further analysis. Together, they provided a rigorous framework for identifying which force plate metrics are most strongly associated with injury risk in Denison athletes.

4.2 Data Preparation

The datasets for this study were provided by Denison University Strength and Conditioning and included private, sensitive records of athlete performance and injury status. The department approved the use of the data and required us to sign a confidentiality agreement before use. Prior to analysis, several steps were taken to ensure the data was properly cleaned, merged, and de-identified. First, we gained access to the Hawkin Dynamics API and extracted CMJ records for the 2024-2025 academic year using the private `hawkinR` library in RStudio. This data included raw performance metrics for each jump and contained identifiable athlete names. Prior to analysis, we performed a `VLOOKUP` in Microsoft Excel using a separate spreadsheet of unique athlete codes and removed all names for de-identification purposes. Next, we compiled eight injury datasets into a single injury dataset, separate from the CMJ data. These records spanned ten months of data collection and were filtered on the variable "InjuryType" to exclude illnesses like influenza or allergic reactions. Once the CMJ and injury datasets were cleaned, they were merged by athlete ID, and the binary dependent variable was created. For injured athletes, all jumps occurring within four weeks prior to each documented injury were labeled as '1'. All other jumps were labeled as '0'. This approach allowed the model to focus on identifying performance patterns that preceded injury events rather than those that occurred after an injury had already taken place. However, the selection of the four-week window was rather arbitrary, discussed further in the limitations section. Because our dataset included over seventy CMJ metrics per jump, we did not include the entire set of variables in the final model. Rather than applying strict inclusion criteria prior to analysis, the filtering occurred through algorithmic feature selection. This avoided the occurrence of arbitrarily excluding variables based solely on existing literature, allowing us to rely on feature selection methods to reduce redundancy and improve the final model.

4.3 Analytical Methods

4.3.1 Exploratory Analysis and Feature Selection

Given the large number of potential predictors, feature selection was essential to avoid overfitting and ensure the final model highlighted the most relevant metrics for injury prediction. Firstly, we explored Pearson's correlation matrix to interpret the relationship and dependence between CMJ metrics. Specifically, the squared Pearson correlation coefficient gave an indication on the strength of the linear relationship between two variables (Benesty et al., 2009). When expanded into a matrix of several variables and displayed as a distribution in a histogram, we interpreted the general relationship between various features in bulk (Khalil et al., 2024). This method offered a general descriptive method to support our prior hypothesis of the CMJ metrics' multicollinearity.

Because injuries are rare events, our dataset was highly imbalanced with substantially more healthy jumps than pre-injury jumps. To address this imbalance, we implemented class weighting in our penalized regression models. Specifically, observation weights were applied during Lasso regularization to penalize misclassification of injury events more heavily and prevent the model from overfitting the healthy class (Friedman et al., 2025). Further, for feature selection, we explored Lasso regularization as a variable filtering tool. This analysis was implemented in R using the glmnet package, specifically designed for fitting penalized regression models (R Core Team, 2025). This approach was a well-suited starting point for our study because we wanted to manage high dimensionality, reduce redundancy among correlated variables, and provide an interpretable subset of predictors that could guide subsequent modeling (Friedman et al., 2025). Moreover, Lasso regularization applies an L1 penalty to the model's coefficients, shrinking some to exactly zero and leaving a smaller, more interpretable set of predictors (James et al., 2021, p. 245). This method proved particularly useful for our data, where many variables are redundant or highly correlated. We employed stratified random sampling by athleteID in order to ensure that the same athlete did not appear in both training and test sets (Sanderson, 2024). To determine the appropriate level of penalty, we used 10-fold cross-validation to select the tuning parameter (λ) that minimized the prediction error (James et al., 2021, p. 250). We then calculated the variation inflation factors (VIFs) to assess the multicollinearity among the remaining predictors. The VIF quantified how much the variance of a coefficient was inflated due to correlation with other predictors, providing a more precise diagnostic than simple pairwise correlations (James et al., 2021, p. 102). Although Lasso penalization mitigated some of this redundancy, computing VIFs offered an additional check to ensure the final subset of CMJ metrics did not include predictors that challenge the interpretability of the model. However, several predictors still exhibited elevated VIF values, indicating persistent multicollinearity. This limitation is addressed further in the Discussion section 6.1 and motivated our decision to use these variables only as inputs for more flexible, tree-based models rather than as final regression estimates. Moreover, we implemented each of these methods with the goal of minimizing the amount of CMJ metrics fed into our predictive model to improve functionality and interpretability.

4.3.2 Predictive Analysis

For our predictive model, we followed Lövdal et al. (2021) by employing Extreme Gradient Boosting (XGBoost), a tree-based machine learning algorithm acknowledged for its speed, scalability, and ability to capture complex, non-linear relationships (Chen & Guestrin, 2016). To increase interpretability, we also used SHAP values to quantify the direct contribution of each feature to the prediction of the model (SAMueL project team, 2022). However, prior to modeling, we needed to address class imbalance in our dataset. To preserve the natural class distribution while still addressing imbalance, we did not apply resampling techniques such as bagging or manual oversampling. Instead, we maintained the full dataset and incorporated class weighting directly into the predictive algorithm. Data were first divided into training and testing sets using proportional stratified sampling to ensure that both subsets maintained the original injury to non-injury ratio (Sanderson, 2024). To handle imbalance during model training, we implemented the XGBoost `scale_pos_weight` parameter during all stages of model development, defined as the ratio of negative to positive cases in the training set (XGBoost Developers, 2025). This weighting increased the penalty for misclassifying injury events, improving sensitivity to the minority class without artificially replicating the data. The set of characteristics used in this stage consisted of the reduced subset of CMJ metrics selected during Lasso regularization.

We trained the model using the `xgboost` R package and obtained SHAP values directly from the built-in `TreeSHAP` function of `xgboost` (Chen et al., 2025; Casas, 2019). With the training data, we used 5-fold cross-validation to tune the model and assess predictive performance. This approach repeatedly partitioned the data into 5 equally sized subsets, training on 4 folds while validating on the remaining fold, and then averaging the results to estimate the model performance (James et al., 2021, p. 203). Throughout cross-validation and final training, the model was fit using the same class weighting. The final XGBoost model was trained using the optimal parameters identified through cross-validation and applied to the test set.

We also identified SHAP values for each 5-fold split, reporting log-odds contributions to the model prediction and summarizing feature influence by ranking mean absolute SHAP values (SAMueL project team, 2022). To assess the model's performance, we used the receiver operator characteristic (ROC) curve and its area under the curve (AUC) to describe the fraction of true positives versus the corresponding rate of false positives (Lövdal et al., 2021). In addition, we reported the precision, sensitivity, and F1 score to further evaluate classification performance under class imbalance. Precision measures the proportion of predicted injury cases that are correctly identified. Sensitivity, or recall, on the other hand, measures the proportion of actual injury cases that are correctly detected. The F1 score represents the harmonic mean of precision and recall, providing a single metric that balances both false positives and false negatives (Powers, 2011). Because our dataset contained far fewer injury events than healthy ones, these measures provided a more meaningful evaluation of accuracy when assessing performance on the injured class. Through a combination of these statistical methods, we hoped to identify the most predictive CMJ metrics for injury risk while remaining conscious of the model's predictive power, context, and accuracy.

5 Results

5.1 Exploratory Analysis

A total of 23,954 jumps were recorded across all athletes. Of these, 1,385 jumps (approximately 5.3%) occurred within 28 days prior to a documented injury event. This indicates a substantial class imbalance, with non-injured jumps accounting for approximately 94.7% of all observations. We next examined the structure of the dataset to understand the multicollinearity of the CMJ metrics. We created a table of pairwise Pearson correlation coefficients, finding 70 unique variables with $|r| > 0.8$ (Benesty et al., 2009). This indicates strong multicollinearity, which was expected given the biomechanical nature of the CMJ. Further, this widespread multicollinearity highlighted the need for penalized regression to extract the most important subset of features for injury classification.

5.2 Feature Selection (Weighted Lasso Regularization)

Given the substantial class imbalance, we fit a weighted classifier so both classes contributed equally during training. We applied 10-fold cross-validation to find the optimal λ value for Lasso regularization ($\alpha = 1$). Cross-validation resulted in $\lambda_{min} = 0.00022$, which yielded the highest AUC during training. Slightly higher, $\lambda_{1se} = 0.00047$ represented a simpler model within one standard error of the minimum. The cross-validation curve (Figure 1) shows that model performance steadily increased as λ decreased, peaking near an AUC of 0.64. The gap between λ_{min} and λ_{1se} is relatively small, indicating that the simpler λ_{1se} model sacrificed minimal predictive performance while retaining fewer predictors.

Figure 1: Weighted Lasso Cross-Validation (α = 1)

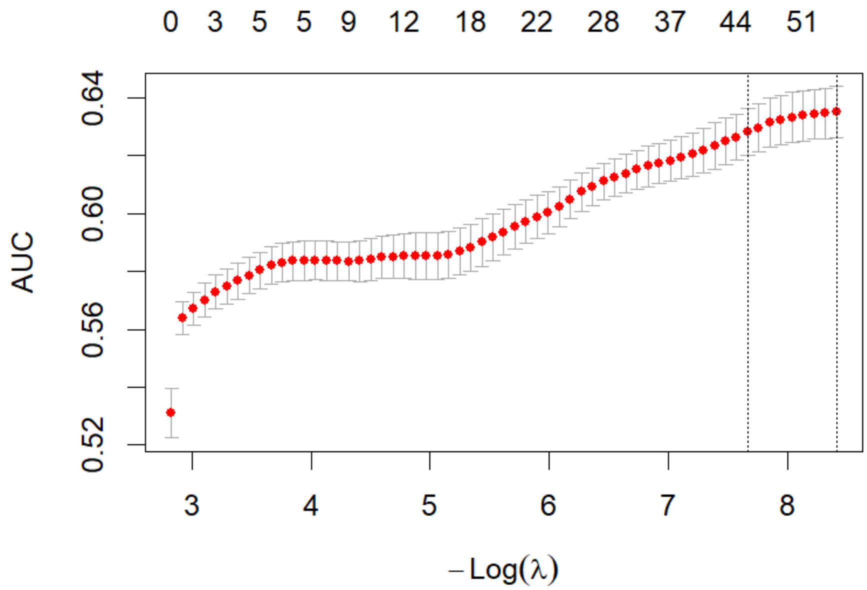

Figure 1: Weighted Lasso cross-validation results showing that predictive performance peaks at λ_{min} (right) while λ_{1se} (left) offers a simpler model with comparable AUC. The number of predictors each λ yielded can be seen in the topmost portion of the graph, below the title.

Specifically, weighted Lasso yielded 43 (λ_{1se}) and 51 (λ_{min}) nonzero predictors, a meaningful reduction from the original set. We explored coefficient trace plots to inspect the model selection process. However, given the large number of CMJ metrics and the fact that our focus was not on using Lasso for prediction, the trace plots were not included. We then tested the performance of both (λ_{1se}) and (λ_{min}) on the test set. To preserve the real-world distribution of injury events, class weights were not applied during evaluation on the test set. This ensured that performance metrics reflected how the model would perform on naturally imbalanced data, improving both interpretability and practicality. Figure 2 shows the discriminatory ability of both (λ_{1se}) and (λ_{min}), indicating how well they separated injured from non-injured jumps. The ROC curve plots the sensitivity vs. specificity with AUC reflecting the model's discriminatory power.

Figure 2: Test ROC for Weighted Lasso (AUC = 0.52 for λ_{1se} and λ_{min})

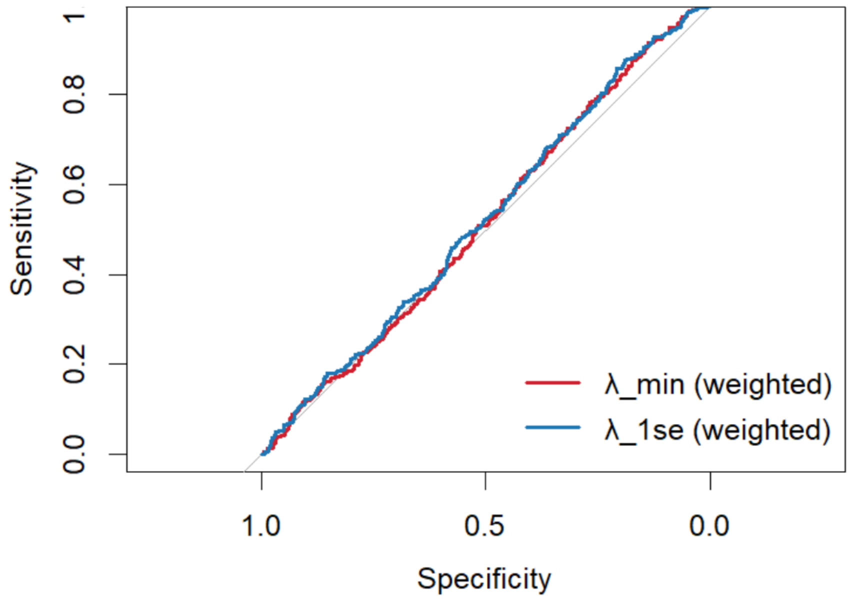

Figure 2: With an AUC only slightly above 0.5, weighted Lasso regularization performed only marginally better than random guessing at discriminating between injured and non-injured jumps on the test set. Despite its limited predictive power, λ_{1se} selected features were retained and used for future modeling.

Because both (λ_{1se}) and (λ_{min}) produced AUC values slightly above 0.5, the model performed only marginally better than random guessing when separating injured vs. non-injured jumps on the test set. Further, the red and blue curves of Figure 2 hug the diagonal, confirming limited predictive power of linear models on this dataset. Altogether, these results imply that injury risk is not well explained by a purely linear combination of the Lasso-selected CMJ metrics.

Despite its low predictive power, we proceeded to calculate the Variance Inflation Factors (VIF) on the (λ_{1se}) selected variables. We received a median VIF of 90.5 with a maximum VIF of 1,474.3. Therefore, high multicollinearity persisted among Lasso-selected CMJ metrics. Yet, the model still performed meaningful feature selection under penalization. Lasso did not eliminate all correlation, but it successfully removed many noisy, redundant, or weakly associated predictors from the original CMJ metrics. The 43 retained variables represent those with the strongest linear association with injury risk.

Importantly, this subset still holds value for modeling. While Lasso assumes linearity and yielded limited predictive accuracy on its own, the reduced feature set can be used as input for non-linear models. Tree-based models do not require linear relationships and are capable of handling noisy input metrics (Chen & Guestrin, 2016). By feeding only the 43 most relevant variables into XGBoost, we reduce noise, improve interpretability, and limit the risk of overfitting while retaining the metrics most linked to injury. Therefore, we created a new dataset containing only these 43 Lasso-selected predictors, serving as the input for our XGBoost model.

5.3 Predictive Modeling (XGBoost)

After feature selection, we trained an Extreme Gradient Boosting (XGBoost) model using only the 43 variables retained from Lasso regularization. We first split the data into a training (80%) and testing (20%) set using proportional stratified sampling, maintaining approximately 5% injured jumps in both sets. Further, we computed the imbalance ratio of the training set by dividing the total non-injured cases by the total injured cases to receive the `scale_pos_weight` value as shown in Table 1.

Table 1: Class Imbalance Summary: Training Set Only

Class / Parameter	Count
Non-Injured (0)	19,661
Injured (1)	1,103
scale_pos_weight	17.83

In the training data, only 1,103 out of all jumps were injury events. So, each injured jump was treated as roughly 18 times more important than non-injured jumps during model training. This balanced the gradient so the model paid attention to both classes equally.

Next, we performed a grid search with 5-fold cross-validation to tune hyperparameters, including `eta`, `max_depth`, `min_child_weight`,`subsample`, and `colsample_bytree`. Early stopping was used to select the optimal number of boosting rounds. The best performing configuration produced an AUC of approximately 0.836 and selected 700 trees as the optimal stopping point. This high number of trees was expected given our low learning rate (eta = 0.05) which slows learning and reduces overfitting (xgboost developers, 2025).

Figure 3: ROC Curve for XGBoost: Test Set

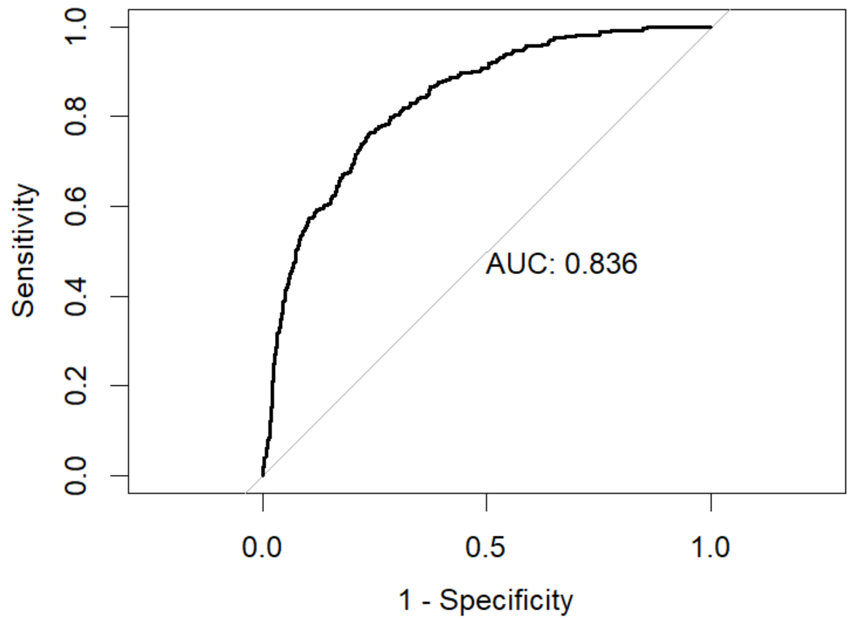

Figure 3: The curve lies well above the diagonal baseline, with AUC = 0.836. This indicates that in roughly 84% of randomly paired injured vs. non-injured jumps, the injured jump receives a higher predicted probability.

We used the fitted model to predict probabilities of injury on the test set. Figure 3 shows the Receiver Operating Characteristic (ROC) curve of the model on the testing data. The curve lies well above the diagonal line of random guessing, and the AUC was 0.836. This means that, in roughly 84% of randomly paired injured and non-injured jumps, the injured jump was assigned a higher probability by the model. This represents strong discriminatory performance despite the imbalanced dataset.

We then converted probabilities into binary predictions using a threshold of 0.5. While this threshold produced high overall accuracy (93.6%), it resulted in low sensitivity (recall) (23.4%), meaning the model missed many actual injury cases. This was expected from the imbalanced dataset where the model learned to favor the majority class.

Because injury prediction prioritizes identifying at-risk athletes rather than overall accuracy, we computed classification metrics across all possible thresholds. We then selected the probability cutoff that maximized the F1-score, which balances precision and recall (Powers, 2011). The best threshold was 0.20 with an F1 score of 0.357.

Table 2: Classification Metrics at Optimal Threshold (0.20) from XGBoost

Threshold	Accuracy	Precision	Recall	F1
0.2	0.912	0.296	0.45	0.357

Table 2 displays the classification metrics at this optimal threshold. Compared to the default 0.5 threshold, sensitivity increased substantially (0.238 to 0.45).

Table 3: Confusion Matrix at Optimal Threshold (0.20) from XGBoost

	Predicted 0	Predicted 1	Total
Actual 0	94%	6%	4,908
Actual 1	55%	45%	282

The corresponding confusion matrix (Table 3) shows the percentage of predicted to actual injured (1) and non-injured (0) cases, showing 6% (302) false positives and 55% (155) false negatives. These false negative cases represent missed injury cases and are of greater concern in injury prevention contexts than false positives. Therefore, prioritizing recall aligns with the practical goal of identifying athletes at elevated risk of injury.

Figure 4: Precision-Recall Curve For XGBoost

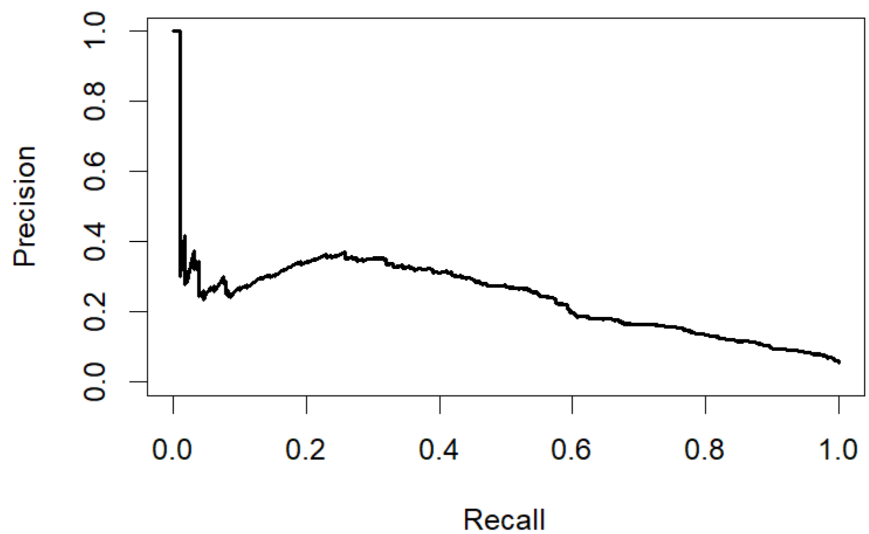

Figure 4: The curve illustrates how precision declines as recall increases, reflecting the trade-off between identifying more true injuries and generating additional false positives. The model achieved its highest F1 score at a probability threshold of 0.20, where precision was 0.296, recall was 0.45, and AUPRC was 0.243.

Because only 6% of the dataset contained injured jumps, the Precision-Recall (PR) curve provides a more accurate representation of the model accuracy. The PR curve illustrates the trade-off between the model's ability to correctly identify injuries (recall) and its correctness when it predicts an injury (precision). At the optimal threshold of 0.20, precision was approximately 0.296 and recall was 0.45, representing the best balance of the two metrics obtainable from this model. At this point, the area under the precision-recall curve (AUPRC) was 0.243. Given that the baseline AUPRC for random guessing equals the prevalance of injuries (6%), our model performed roughly four times better than random chance (Saito and Rehmsmeier, 2015). As recall increased beyond the optimal point, precision drastically decreased, indicating that the model cannot detect all injuries without generating a large number of false positives.

5.4 Feature Importance & Interpretation

We then computed SHAP (SHapley Additive exPlanations) values on the test set to find the most influential CMJ metrics on injury risk (SAMueL project team, 2022). By taking the mean absolute SHAP value, we ranked features by their influence overall rather than considering the direction of their influence.

Figure 5: Top 10 Most Influential Features in XGBoost by Mean Absolute SHAP

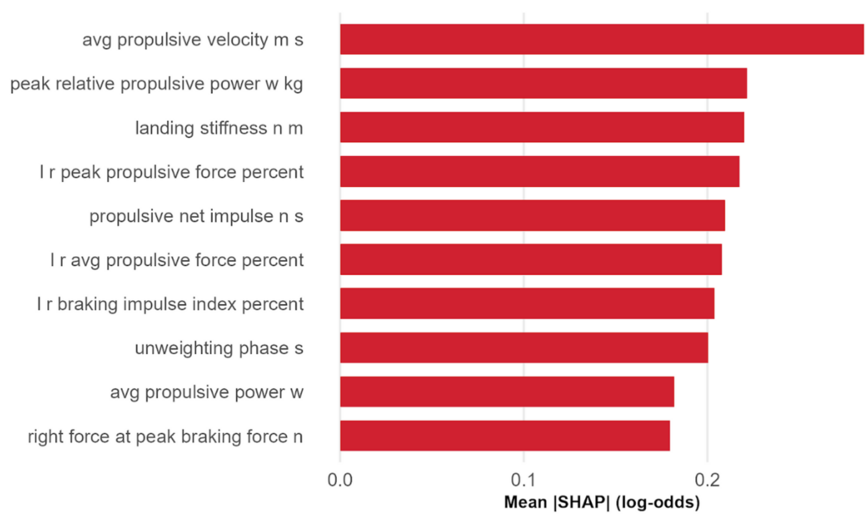

Figure 5: Bars represent each variable's average contribution to the model's prediction of injury risk (on a log-odds scale), regardless of whether the effect increased or decreased risk. Notably, 6 of the top 10 most influential CMJ metrics come from the propulsive phase of the jump.

Figure 5 displays the top 10 most influential CMJ variables. The highest-ranked metric was average propulsive velocity, followed by peak relative propulsive power and landing stiffness. Further, six of the top ten features were from the propulsive phase, while braking and landing-related variables appeared less frequently. There were no variables from the weighing or flight phases of the jump.

The concentration of influential metrics within the propulsive phase suggests that model-predicted injury risk is more sensitive to the force applied during the upward movement of the jump rather than to the landing or eccentric phases. However, SHAP values alone do not indicate causation or practical implementation. In the following discussion section, these results are explored further for the purposes of real-world application.

6 Discussion

6.1 Summary of Findings

This study first explored class-weighted Lasso regularization, prioritizing feature selection rather than predictive accuracy. Lasso yielded a sparser set of 43 CMJ variables that worked only marginally better than random guessing at discriminating between injured and non-injured jumps in the test set. Therefore, Lasso results were used strictly for feature selection, not prediction.

Using the 43 Lasso-selected variables, we trained a balanced XGBoost model with early stopping to select the optimal number of boosting rounds. The best performing configuration resulted in an AUC of 0.836 and 700 optimal trees. To create binary predictions on the test set, we chose a probability cutoff that maximized the F1 score. At this optimal cutoff, the model achieved a precision of 29.6% and a recall of 45%. For interpretation, we computed the mean absolute SHAP values on the test set to find the most influential CMJ metrics for injury classification. Of the top ten most influential features, six were from the propulsive phase, two from braking, one unweighting, and one landing.

6.2 Interpretation of Model Performance

After training, we plotted the ROC curve to measure how well the model ranked injured vs. non-injured jumps (Figure 3). This was done independently of any threshold, meaning it purely showed the model's discrimination ability. This measure was not affected by class imbalance as much as accuracy. The model produced an AUC of 0.836. Thus, in approximately 84% of random injured vs. non-injured jump pairs, the injured jump got a higher predicted probability. Additionally, the ROC curve bowed well above the diagonal baseline of random guessing, suggesting strong separability. Further, the model was able to discern who tends to be injured more often than not. However, this does not mean the model gives perfect injury predictions. In other words, the ROC curve indicates strong underlying pattern recognition, but not necessarily strong classification at a specific threshold.

These results can be misleading given that injuries only made up 6% of the data. The model can rank well, but rarely predict the injured class at certain thresholds. A better estimate of model performance given the class imbalance is the Precision-Recall Curve (Figure 4), which compares the trustworthiness of positive predictions (precision) to the correctly predicted injury cases (recall) with the F1 score representing the balance of both. These measures give a realistic sense of usefulness given the rarity of injury cases in the data.

At the original probability threshold of 0.5, the model promoted high accuracy (93.6%) but less than ideal recall (24%). Because of the imbalanced nature of the data, the model defaulted to predicting "no injury" most of the time with high accuracy but zero value in practice. Therefore, accuracy is not an appropriate metric of evaluation of the model's performance. For the purpose of injury prediction, we needed a model that maximized the F1 score, balancing false negatives and false positives. The maximum F1 score (0.357) occurred at a lower probability threshold of 0.20, meaning that any jump receiving a predicted injury risk of 20% or higher was treated as attention-worthy. This reflects a deliberately cautious approach where the model is willing to generate more false alarms in exchange for identifying more potential injuries. In practice, this could result in more athletes being monitored but fewer unexpected injuries. At this threshold, precision decreased to 0.296, but recall significantly increased to 0.45. Thus, the model correctly flagged 45% of injuries, but approximately 70% of its flags would be false alarms. However, it significantly reduced missed injury cases, aligning the model with the practical goal of early risk detection rather than overall accuracy. The confusion matrix (Table 3) confirms the previous findings, exploring the true positives (127), false negatives (155), false positives (302), and true negatives (4,606). In injury prevention, false negatives (recall) represent missed injury cases and are the most important to minimize. False positives are not dangerous but can cause misleading recommendations to prevent injury, wasting time and athlete training. Thus, prioritizing recall over precision aligns with injury prevention goals. Still, given the model results, it does not support direct injury prediction. Rather, it may support an overly cautious, early warning system for injury.

Moreover, the model recognized athletes whose CMJ signals were changing in concerning ways, detecting injury risk better than random chance. Despite the noise, imbalance, and multi-sport variability, the model still produced meaningful insights into how the CMJ can be used as an early warning system of injury risk. Further, the model still missed 55% of actual injury cases. In addition, there was an extremely high false-positive rate which would raise unnecessary concerns and require more intense human intervention. The model is not a perfect injury predictor but it can be useful for monitoring and screening at cautious probability threshold.

6.3 Feature Importance and CMJ Interpretation

Moreover, rather than predicting injuries, the goal of the research was to determine which features were most important for flagging injury risk. Black-box models like XGBoost offer high predictive power but limited transparency. Therefore, when classifying injury risk, XGBoost is less valuable when explaining why a model flags an athlete as high-risk (Rudin, 2019). Thus, we used SHAP values to interpret which CMJ metrics most strongly contributed to model-predicted injury risk. Although SHAP can provide both local (per-athlete or per-jump predictions) and global explanations, this study focused only on global SHAP analysis. Specifically, mean absolute SHAP values were calculated across all test-set observations to rank features by their overall influence on injury prediction, regardless of whether their effect increased or decreased risk (Lundberg & Lee, 2017). This global approach was ideal for selecting a generalizable list of metrics rather than jump or athlete-specific measures beyond the scope of this study.

Furthermore, average propulsive velocity had the highest mean SHAP value of 0.285, meaning this variable contributed the largest average change in predicted injury risk across all athletes and jumps. Peak relative propulsive power (0.221) and landing stiffness (0.221) were the next highest, showing strong influence on the model's injury prediction. The remaining top-ranked variables all produced mean SHAP values between 0.18 and 0.22, indicating that no single feature drives predictions alone. Rather, several propulsive phase variables contributed comparably to injury risk with 6 of the top 10 originating from this jump phase. Therefore, how athletes generate force to leave the ground proves most predictive of injury over variables from the braking or landing phases (Hawkin, 2021). This could mean that changes in propulsive velocity and/or power may signal fatigue leading up to an injury.

However, while SHAP supports early detection of injury risk, it does not imply causal relationships. Average propulsive velocity resulted in a high SHAP value, suggesting strong influence on injury probability in the model. In other words, in this dataset, it did consistently appear in jumps that preceded injury events. However, it does not show that decreases in average propulsive velocity cause injury. Moreover, these metrics could serve as potential signals coaches can track over time.

6.4 Practical Applications

The final model was not accurate enough to predict injuries or label athletes as injury-prone. However, it did identify concerning CMJ patterns earlier than random chance, making it useful as a monitoring or flagging tool. In practical use, it could act as an alert system rather than a definitive prediction of injury. Importantly, the model informs but does not replace medical or professional judgment. Technically, if a jump exceeds the 0.20 threshold, an athlete could be flagged for review. However, it is up to coaches and medical professionals to monitor the athlete and their performance to determine whether the model's interpretation was applicable. As shown previously, the model tended to predict significant false positives. Therefore, a predicted injury should spark communication between coaches, athletic trainers, and athletes rather than an immediate reduction in training loads or assumption that the model produced a false positive result. When paired with the SHAP-selected variables, the model can serve as a signal to initiate conversations with athletes about potential injury risk.

Moreover, the SHAP-identified metrics could be applied to daily training. Instead of coaches arbitrarily tracking changes in all 70+ metrics from the CMJ, they can focus on the most informative ones. From the analysis, the top variables to monitor include propulsive performance, landing stiffness, unweighting, and left to right braking metrics (Figure 5). In practical use, coaches may track individual baselines for each athlete, tracking declines over time instead of potential outliers from single jumps. When having fatigue-related meetings with athletes, these may inform conversations and provide a statistically supported layer of tracking. Still, these outputs should never be used to label an athlete as "injury-prone", make assumptions about effort or intent, or penalize the athlete in any way. Instead, they should be used to promote athlete well-being.

7 Limitations

While these applications demonstrate the practical value of CMJ-based injury monitoring, the model also carries meaningful limitations. To start, we only analyzed one academic year, which limited the sample size of both injured and non-injured athletes. Moreover, athletes did not have an equal or consistent number of jumps in the chosen 28-day window leading up to an injury. We picked this four-week window rather arbitrarily to capture the most amount of injured-labeled jumps. At Denison, when athletes are in-season, they jump much more frequently than athletes that are out of season. In addition, if an injury occurred within the first couple months of the academic year, there may not have been a sufficient amount of jumps prior to an injury date. More than this, the window of fatigue or decline leading up to an injury may vary depending on the injury type or athlete. Because our outcome variable was binary, we lose the nuances of understanding if a variable changes gradually before injury or drops suddenly and ignore injury severity and recovery time. This lack of of consistency more than likely influences modeling capabilities. Further, the dataset includes all sports, genders, and ages, causing added variation in the model. Each jump could vary by sport, gender, and individual athlete, making the patterns even harder to model. The dataset also does not consider previous injuries or genetic predispositions, which would have a direct impact on the probability of a new injury. Similarly, the study did not exclude upper body injuries. The jump most likely reflects lower-body fatigue more directly and, therefore, can pattern jumps leading to lower body injuries easier. Most importantly, CMJ test quality may vary from jump to jump even without an indication of fatigue or injury-risk. This could depend on an unlimited amount of unobserved factors such as lack of sleep, caffeine, time of day, and even the athlete's warm-up. All of these details increase the variation in the model, making it even more difficult to pattern the CMJ leading up to injury.

In regards to model performance, class imbalance still remains a challenge. Even with class weighting, the minority class is still extremely small, limiting generalizability and affecting precision. Because of this limited sample size, we were unable to explore temporal validation, such as training on Fall 2024 data and testing on Spring 2025, which may further reduce the generalizability of the model across time (Li et al., 2023). Another reduction in generalizability comes from the threshold-based decision making. Using a fixed threshold of 0.20 may not generalize across teams, sports, genders, or seasons. Lastly, the model assumes independence of jumps. Realistically, jumps from the same athlete are all related to each other, limiting the model's accuracy. However, these limitations do not negate the value of our findings. Instead, they emphasize the need for cautious interpretation and further validation across seasons, teams, and individual athletes.

8 Future Work

While informative, this work captures only a fraction of what could be uncovered through continue analysis of the CMJ data. For example, the current study used global SHAP (average importance across all athletes) to show CMJ metric influence. Future work could explore local

SHAP values which could explain why a specific jump from a specific athlete at a specific time was flagged as risky. Athlete 1 may show declining propulsive velocity while Athlete 2 shows increasing landing stiffness to be the red flag. Exploring this route would move the analysis from general to athlete-specific. However, the global analysis could still be used to create a real-time alert system to change jumps to signals for coaches. When an athlete jumps, results could be sent directly to a dashboard displaying individual injury risk and abnormal variables. Staff could then ask the athlete further questions to determine if the jump was simply abnormal or a true signal.

As athletes continue to jump more frequently and more data is collected, the study may be expanded across multiple seasons and multiple teams. This would allow for more time-series modeling (such as temporal XGBoost) that would forecast an athlete's risk trajectory rather than just individual jumps (Kang et al., 2021). This would allow for athlete-specific baselines rather than just one, generalizable model for all athletes. Every athlete is different and, thus, have different susceptibilities to injury. Instead of a single 28 day pre-injury window, future work could test different time windows depending on injury type, history, severity, or location. Each of these factors vary from athlete to athlete and may never be truly measurable but with added aggregation, we may grow closer to uncovering the true relationship between the CMJ and injury risk.

However, CMJ analysis does not have to end at injury classification. Future CMJ models could target changes in specific variables to track recovery, training readiness, or week-to-week performance improvement. Injuries are rare, noisy, and hard to label accurately. However, performance-related outcomes occur every session, offering more balanced data for prediction. The CMJ could be explored in relation to in-game performance indicators, such as hitting percentage, sprint speed, swim split times, accuracy metrics, or even playing time. Using combined performance and force plate data, a similar method could be applied to predict performance peaks or slumps. Overall, with this study as the foundation, the possibilities for future work remain limitless.

9 Conclusion

While there remains significant room for future exploration, this study provides an essential first step toward developing interpretable, data-driven systems for athlete monitoring using CMJ metrics. By identifying which CMJ variables consistently signal changes in athlete condition, this work moves force plate analysis beyond isolated jump metrics and toward continuous athlete monitoring. The combination of Lasso feature selection, XGBoost modeling, and SHAP interpretation demonstrated that CMJ data can reveal subtle patterns in performance even when injury prediction itself remains uncertain. In this way, the study offers an early framework for transforming routine CMJ testing into a real-time signal system that supports data-informed coaching decisions, proactive injury prevention, and improved athlete wellbeing.

10 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.

Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. In Noise reduction in speech processing (pp. 1-4). Berlin, Heidelberg: Springer Berlin Heidelberg.

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.

Brownlee, J. (2024). XGBoosting: Making You Awesome At XGBoost. XGBoosting.com.

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.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., Yuan, J., & XGBoost contributors. (2025). xgboost: Extreme Gradient Boosting (R package version 1.7.11.1) [Computer software]. CRAN. https://CRAN.R-project.org/package=xgboost.

Denison University Strength and Conditioning. (2024). Force plate and injury dataset: 2024–2025 academic year [Unpublished dataset]. Denison University.

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

Friedman, J., Hastie, T., Tibshirani, R., Narasimhan, B., Tay, K., Simon, N., & Qian, J. (2025). glmnet: Lasso and elastic-net regularized generalized linear models (Version 4.1-10) [R package]. Comprehensive R Archive Network (CRAN).

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.

Kang, J., Guo, X., Fang, L., Wang, X., & Fan, Z. (2021). Integration of Internet search data to predict tourism trends using spatial-temporal XGBoost composite model. International Journal of Geographical Information Science, 36(2), 236–252.

Khalil, N. Z., Kong, N., & Fricke, H. (2024). The influence of GNP on the mechanical and thermomechanical properties of epoxy adhesive: Pearson correlation matrix and heatmap application in data interpretation. Polymer Composites, 45(10), 8997-9018.

Li, S., Dou, R., Song, X., Lui, K. Y., Xu, J., Guo, Z., Hu, X., Guan, X., & Cai, C. (2023). Developing an interpretable machine learning model to predict in-hospital mortality in sepsis patients: A retrospective temporal validation study. Journal of Clinical Medicine, 12(3), 915.

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.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

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.

OpenAI. (2023). ChatGPT (GPT-4) [Large language model].

Powers, D. M. W. (2020). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv.

R Core Team. (2025). R: A language and environment for statistical computing (Version 4.4.1) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1, 206–215 (2019).

SAMueL Project Team. (2022). Explaining XGBoost model predictions with SHAP values. GitHub Pages.

Sanderson II, S. P. (2024, July 29). Stratified sampling in R: A practical guide with base R and dplyr. Steve's Data Tips and Tricks.

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.

Walker, O. (2025, March 17). Countermovement jump (CMJ). Science for Sport.

XGBoost Developers. (2025). XGBoost parameters. XGBoost Documentation.

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.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301–320.