Research Design

**From Jumps to Signals:**

**Selecting Countermovement Jump Features for Injury-Risk Classification**

Jordyn Maurer

5 October 2025

Seminar in Data Analytics

1. Research Overview

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

Because this study will rely on secondary, observational data, it is non-experimental in nature. Therefore, we cannot determine whether specific CMJ metrics cause soft-tissue injuries. Instead, our analysis is limited to identifying associations between performance metrics and injury occurrence. We cannot establish direct causation between a specific metric and soft-tissues injury, but our findings can highlight potential risk factors to guide intervention strategies. Thus, our results will be interpreted as evidence of statistical relationships rather than proof that changes in specific CMJ metrics will produce soft-tissue injuries.

Further, to explore the relationship between the CMJ and soft-tissue injuries, we will employ several different methods of analysis. We will primarily explore Extreme Gradient Boosting (XGBoost) with SHAP values for prediction and feature selection (SAMueL project team, 2022). If this approach does not yield meaningful or interpretable results, we will also consider alternative methods. For feature selection, we will explore penalized logistic regression with Elastic Net and, for comparison, Lasso regularization which can reduce the large number of CMJ variables to a subset that best predicts injury risk (James el al., 2021, p. 242). Additionally, we may consider Pearson correlation heat mapping to visually identify relationships among metrics and identify redundant or collinear variables to remove prior to modeling (Khalil et al., 2024). These approaches will provide us with flexibility and ensure other techniques are available to guide variable selection and predictive model construction.

1. Data Preparation

The datasets for this study were provided by Denison University Strength and Conditioning and include 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 will be taken to ensure the data is properly cleaned, merged, and de-identified. First, we have access to the Hawkin Dynamics API and have already extracted CMJ records for the 2024-2025 academic year using the private hawkinR library in RStudio. This data includes raw performance metrics for each jump and contains identifiable athlete names. Prior to analysis, we performed a VLOOKUP in Excel using a separate spreadsheet of unique athlete codes and removed all names for de-identification purposes. Next, we manually compiled injury records into a single dataset. These records span ten months of data collection and will be filtered on the variable “InjuryType” to include only non-contact, soft-tissue injuries. Once both datasets are cleaned, they will be merged by athlete ID, and we will create our binary dependent variable. For injured athletes, jumps occurring two weeks prior to each documented injury will be labeled as ‘1’ while all other jumps will remain labeled as ‘0’. This approach allows the model to focus on identifying performance patterns that precede injury events rather than those that occur after an injury has already taken place.

Because our dataset includes over seventy CMJ metrics per jump, we will not include the entire set of variables in the final model. Rather than applying strict inclusion criteria prior to analysis, the majority of filtering will occur through feature selection methods during the exploratory phase of the analysis. This avoids the occurrence of arbitrarily excluding variables based solely on existing literature. However, the first part of the CMJ, known as the weighing phase, calculates the athlete’s body weight and serves as a reference point for the subsequent jump phases (Hawkin Dynamics, 2021). Because this phase does not involve active movement, it is unlikely to provide context about injury risk and will be excluded from the model. We will rely on other feature selection methods to further refine the variables to reduce redundancy and improve the final model.

1. Analytical Methods
   1. Exploratory Analysis and Feature Selection

Given the large number of potential predictors, feature selection is essential to avoid overfitting and ensuring the final model highlights the most relevant metrics for injury prediction. Firstly, we will explore logistic regression with Lasso regularization. Lasso regularization applies a 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 is particularly useful for our data, where many variables are redundant or highly correlated. To determine the appropriate level of penalty, we will use cross-validation to select the tuning parameter (λ) that minimizes the prediction error (James et al., 2021, p. 250). This analysis will be implemented in *R* using the glmnet package, specifically designed for fitting Lasso and other penalized regression models (R Core Team, 2025; Friedman et al., 2025). One step further, we will compare Elastic Net (EN) regularization, which blends both Ridge and Lasso penalties. Compared to pure Lasso, EN is often more stable when predictors are correlated and tends to select correlated CMJ metrics together in a "grouping effect,” improving interpretability without sacrificing sparsity (Zou & Hastie, 2005). Implementation will mirror our Lasso setup using the glmnet package, tuning both the mixing parameter (α) and λ using cross-validation grouped by athlete ID. This approach is a well-suited starting point for our study because it manages the high dimensionality, reduces redundancy among correlated variables, and provides an interpretable subset of predictors that can guide subsequent modeling. Together, the results from Lasso and EN will inform which CMJ metrics should be retained for further analysis and guide methods used to examine the relationship among variables.

Moreover, we will calculate Variance Inflation Factors (VIFs) to assess multicollinearity among the remaining predictors. The VIF quantifies how much the variance of a coefficient is inflated due to correlation with other predictors, providing a more precise diagnostic than simple pairwise correlations (James et al., 2021, p. 102). Although Lasso and EN penalization mitigate some of this redundancy, computing VIFs offers an additional check to ensure the final subset of CMJ metrics does not include predictors that challenge the interpretability of the model. In addition, we will explore Pearson’s correlation matrix and heatmapping to interpret the relationship and dependence between CMJ metrics. Specifically the squared Pearson correlation coefficient gives 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 transformed into a color-coded heat map, we can interpret the relationship between various features in bulk (Khalil et al., 2024). This method offers an alternative feature selection method by identifying highly correlated and redundant variables for potential removal. In doing so, we can better ensure our final model is better suited to identifying the key CMJ metrics associated with soft-tissue injury risk with easier interpretability.

* 1. Predictive Analysis

Because soft-tissue injuries are rare events, our dataset is highly imbalanced with more healthy jumps than pre-injury jumps. To address this imbalance, we will implement a bagging approach prior to our predictive analysis. Bagging involves generating multiple bootstrapped samples from the original training set, training a separate model on each sample, and then averaging the predictions, reducing variance and increasing test set accuracy (James et al., 2021). To avoid bias towards the non-injured class, we will train each model on balanced subsets of data that contain equal numbers of injured and non-injured events. For a given test, the final prediction will be calculated as the mean of all model predictions, ensuring both classes are equally represented in the final prediction (Lövdal et al., 2021). Additionally, we will determine the predictors in this analysis using the feature selection methods described above. By first reducing redundancy and narrowing the set of metrics most strongly associated with injury risk, we can more confidently assess the interpretability and efficiency of the final predictive model.

For our predictive model, we will follow 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 will also use SHAP values to quantify each feature’s direct contribution to the model’s prediction (SAMueL project team, 2022). We will construct the model using the xgboost package in R and obtain SHAP values directly from xgboost’s built-in TreeSHAP function (Chen et al., 2025; Casas, 2019). To prevent information leakage with repeated measures, we will split the data by athlete to make sure jumps from the same athlete do not appear in both training and test sets (Lövdal et al., 2021). With the training data, we will use *k*-fold cross-validation to tune the model and assess predictive performance. This approach repeatedly partitions the data into *k* equally sized subsets, training on *k*-1 folds while validating on the remaining fold, and then averaging the results to estimate the model performance (James et al., 2021, p. 203). After applying our bagging techniques described above, the final XGBoost model will be trained using the optimal parameters identified through cross-validation and applied to the test set.

We will also get SHAP values for each *k*-fold split which will be reported as log odds shifts in model prediction. We will then find summary statistics of the SHAP values to identify the CMJ metrics that consistently have the largest impact on injury prediction. We will provide a beeswarm plot to display the mean absolute SHAP values, providing a ranked overview of which metrics are most influential (SAMueL project team, 2022). By comparing SHAP rankings with the features selected through Lasso and correlation filtering, we can identify the most important variables and improve interpretability of the model. To assess the model’s performance, we will use a receiving operator characteristic (ROC) curve to describe the fraction of true positives versus the corresponding rate of false positives (Lövdal et al., 2021). The associated area under the curve (AUC) describes the performance of the model with a value closer to 1 representing a better predictive model (Lövdal et al., 2021). In addition, we will report precision and the F1 score to further evaluate classification performance under class imbalance. Precision measures the proportion of predicted injury cases that are correctly identified. Sensitivity, 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 sensitivity, providing a single metric that balances both false positives and false negatives (Powers, 2011). Because our dataset contains far fewer injury events than healthy ones, these measures provide a more meaningful evaluation of accuracy than overall classification rate when assessing performance on the injured class. Through a combination of these statistical methods, we hope to identify the most predictive CMJ metrics for soft-tissue injury risk while remaining conscious of the model’s predictive power, context, and accuracy.

*References*

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.

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](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.

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). [https://cran.r-project.org/web/packages/glmnet/index.html](https://cran.r-project.org/web/packages/glmnet/index.html?utm_source=chatgpt.com)

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

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

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.

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.

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](https://www.r-project.org)

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>

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.