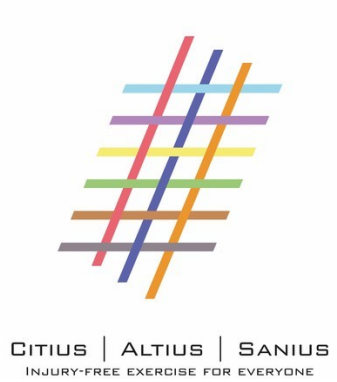

Upper Extremity Injury Prediction in Elite Youth Baseball Pitchers using Classification Methods

by **P. S. Sengalrayan, BSc**

A Thesis

Presented to the Faculty of the Department of BioMechanical Engineering

Delft University of Technology

In Partial Fulfilment of the Requirements for the Degree of

MSc Mechanical Engineering

October 5, 2020

TU Delft Delft University of Technology

KNBSB

(This page has intentionally been left blank)

Patrick S. Sengalrayan
student #: 4206525

Preface

This thesis is the culmination of my student time in the beautiful city of Delft. After a bachelor at the faculty of Aerospace Engineering, I got interested in the field of BioMechanical Engineering, in which I pursued a masters degree. Last year I went on an internship at NeuRA in Australia, where I specifically focused on being the bridge between technology and the human body. Here I designed perturbations on a high-tech treadmill for future studies on people's reaction to falls, which will influence knowledge about falls in elderly or disabled people in particular. During my studies, I have also been working parttime at an R&D firm of a renowned dutch energy venture. That significantly fostered my interest in programming and machine learning, together with their application to real-life problem-solving.

All of these events came together when I started with my thesis topic. In a broad sense, my research was again aimed at bridging the gap between technology and the human body, more particularly between data science and movement sciences. I believe that combining knowledge of various sciences is the right course to pursue the aim of a greater understanding of the world around us. The challenge then lies in the way the results are communicated with the usage of knowledge of all fields involved. This required a steep learning curve, which I experienced as such but nonetheless enjoyed until the last moments.

In this personal message to the reader, a mention of the COVID-19 pandemic can not be missing. In this context, I would like to thank all of the people supporting me during this time of insecurity and social distancing, both in real-life and online. I also want to address special thanks to Bart van Trigt, who is a competent mentor and a positive person. He helped me out by being always available to discuss issues at hand and reacting promptly to the influence of the pandemic on my research. Last, I would like to thank DirkJan Veeger and Jakob Söhl for helping out with expert advice during the entire duration of my thesis project. They represented an essential factor when making sense of the results obtained.

I wish you, the reader, a pleasant reading and a lot of inspiration.

Patrick Sengalrayan

Thesis Committee:

Prof. dr. H.E.J. (DirkJan) Veeger	*Faculty of 3ME, TUDelft*	Responsible Professor
Bart van Trigt, PhD candidate	*Faculty of 3ME, TUDelft*	Daily Supervisor
Dr. Jakob Söhl	*Faculty of EWI, TUDelft*	External Expert

Abstract

In baseball, pitchers have a high rate of throwing arm injury, which could lead to disability and less training time. This paper aims at investigating whether upcoming injuries of youth baseball pitchers can be detected before the athlete experiences injury symptoms. A total of 118 elite youth baseball pitchers from the Dutch national baseball team and six Dutch academies were followed over three years.

Promising variables like Range of Motion, Muscle Force, Ball Speed and Training Time were included for use in a supervised classification problem. Prediction accuracy performance was then measured for different algorithms in the form of F1 and F2 scores. Results showed deficient performance for injury prediction using single-point-in-time measurements for all examined algorithms, with scores of both F1 and F2 reaching maximums of 0.5. The results, however, revealed the importance of measuring variables like hip force and hip range of motion for shoulder injury prediction, and force in the hip and shoulder together with the total rotational motion (TRM) of the shoulder for elbow injury prediction. Ball speed and training time contributed less for the tested models.

Higher frequency data is needed for better injury prediction performance. Future studies are recommended to measure data with a time between measurements of one to two weeks. This high frequency makes it possible to use time-series analysis to detect slight asymptomatic pathology developments progressing over time, to help youth baseball pitchers avoid injuries and keep their performance ready for top-level play.

Keywords: shoulder injury, elbow injury, range of motion, maximum contraction force, ball speed, training time, decision tree, logistic regression

Contents

1 Introduction

Data analysis is an essential tool for the sports world, with the primary goal of improving individual players' and teams' performance. In baseball, the statistical analysis of game data even has a separate name: sabermetrics, which arose in the 1980s from the combination of the acronym SABR (Society for American Baseball Research) and the word metrics [1].

The modern development of an increasing number of easy to use and powerful technologies has steepened the learning curve both for sabermetrics as well as for other sports analytics, resulting in a significant benefit for the athletes and the entire sports community by providing relevant feedback [2]. It is elementary to say that any type of sports activity is related to some risk of injury [3]. In elite sports, it is essential to reduce injury risk to get the most performance out of the individual athlete's potential, as injured hours can best be spent training or merely recovering from training to improve performance (more about the influence on training load later in this text).

In the Netherlands, a country populated by 17 million people [4], every year there are 4.5 million injuries. In 2016 alone, 121.000 people sought emergency care because of sports injuries. Direct medical costs for these injuries account for over 5 million euros every single year in the Netherlands alone, and experts say half of these injuries could potentially be prevented through support and self-management [5]. That makes injuries an exciting topic to investigate.

Since predicting when an injury might occur can be of high value, quantifiable ways to study the pitching motion are considered. Athletes who repetitively perform pitching motions are exerting significant mechanical stresses on the upper extremities. Especially the shoulder and elbow joints are affected by these mechanical, repetitive stresses, which are known to lead to anatomic changes in youth athletes [6, 7]. Some studies indicate that overuse stresses on the skeletal structure even result in asymptomatic pathologies [8]. The term *asymptomatic* (also referred to as subclinical) indicates that symptoms related to an injury are not showing, and are thus often not detected by the affected person. It could mean that the problematic symptoms are yet to show up (and thus become a *symptomatic* pathology), or that it will resolve itself by the adaptation of the athlete's body. It is the case for both the shoulder and the elbow that injuries occur when these adaptive changes and asymptomatic pathology progress, exceeding the compensation abilities of the athlete's body [8]. Due to the occurrence of asymptomatic deterioration, it is theoretically possible to detect an upcoming injury before the athlete presents its symptoms.

Numerous studies are showing how measurable performance metrics can be related to injury. Training load quantified as pitch count was found to correlate with injury. Here, a maximum of 75 pitches per game is advised for youth pitchers to lower risks of pain, one of the symptoms associated with injury progression [9]. More generally, the effects of training volume, frequency and intensity (or load in general) on athletic performance and injury have been researched widely in the past. It is believed that increases in training load correspond to improving performance until a certain threshold, after which injury is more likely to occur [10–13]. That threshold can lie at different magnitudes for each athlete and is related to the athlete's body ability to compensate [8]. Due to these inter-individual differences, no general threshold can be given to work for all athletes. This concept has been illustrated by Verhagen and Gabbett (Figure 1). They propose that an athlete's load and load capacity should always be seen in relation to one's context and environment [14].

Upper arm injuries are one of the primary injuries in baseball, and they are believed to be mostly related to the late cocking phase [8, 15]. As can be observed in Figure 2, cocking is the phase that takes place right before the acceleration phase of the arm forward and is identified as the transition from external to internal shoulder rotation, concentrating all stresses on the shoulder and the elbow joints [16]. In the case of the shoulder, this transition is considered risky because the *Caput Humeri* (the humeral head) moves in an abnormal motion relative to the *Glenoid Cavity*, causing injury in the superior and posterosuperior *Glenoid Labrum* (the glenoid ligament). This injury has

accordingly been named posterosuperior glenoid (or internal) impingement [15, 17].

Some MRI (magnetic resonance imaging) studies conclude that 80 % of elite overhead athletes have shoulder abnormalities compared to healthy subjects [18], and that 40 % of elite overhead athletes have symptoms related to rotator cuff muscle tear in the dominant shoulder compared to the non-dominant one [19]. Even though asymptomatic, the papers refer to these abnormalities going together with muscle degradation, meaning that continuation of training increases the chance of sustaining an injury on the shoulder joint [8]. One could argue that the degradation of the shoulder muscles is reflected by a decrease in measured joint maximum force.

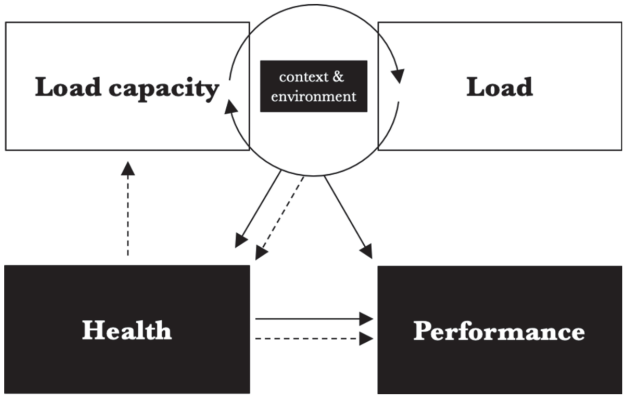

Figure 1: "An integrated view on load, load capacity, performance and health in sports. Dotted lines represent negative relationships and solid lines represent positive relationships." [14]

Elbow injuries are another one of the most common causes of disability among throwing athletes [8, 20, 21]. According to Ciccotti *et al.* elbow injuries contribute to over 7 % of the measured scholastic injury types, and elbow injury is almost twice as likely to occur in pitchers than in position players [21].

Research on elbow injury by Olsen *et al.* in youth athletes reported injured players throwing fastballs at a significantly higher speed compared to healthy subjects, concluding that a relationship exists between increasing ball speed and elbow injuries [22]. A recent study by Kurokawa *et al.* confirmed these results, showing increases in pitch velocity are related to both abnormalities of the medial epicondyle and elbow pain [23]. These findings are in line with the previously mentioned studies on training load and their relation to injury, which makes one assume pitch velocity could be used as a marker for training load.

One paper, by Garrison *et al.*, showed that analysis of elbow injuries could also be performed by investigating the total rotational motion (TRM) of the shoulder joint, which is the amount of combined internal and external rotation of the shoulder at 90° abduction. They specifically focus on players with UCL injuries (Ulnar Collateral Ligament tear), stating they display deficits in the TRM of the throwing shoulder compared to healthy players on the same position [24]. According to Fortenbaugh *et al.*, if there is insufficient shoulder rotation, the throwing arm can get in an incorrect position and lag behind the elbow. That leads to compensation forces in both the shoulder muscles and the elbow, which can cause further injury [25]. Other than for elbow injury, a previous study by Wilk *et al.* also links TRM to shoulder injury, stating that the TRM of the dominant shoulder should be within 5° of the non-throwing shoulder's TRM to decrease the risk of getting injured [26].

Most research performed on baseball injuries until this day has been set up as a cross-sectional study, meaning it focuses on linking measurements to documented injury by investigating a single moment in time. In order to research cause and effect relationships, examining a single point in time does not take into account what might be happening before or after a specific moment. Data involving repeated observations in time of the same sample is called longitudinal data, and homonymously a study involving this type of data is called a longitudinal study.

A longitudinal study can be set up in two ways, namely as a retrospective or as a prospective study. The difference is that the former analyses historical data of a group that may or may not have sustained an injury, while the latter follows a selected sample group for some time. The need for a longitudinal prospective study to link mechanical patterns to the incidence of shoulder and elbow injuries has been proclaimed by Agresta *et al.* in a review paper investigating risk factors for arm injuries among baseball players [27]. Longitudinal study design can be used to study patterns

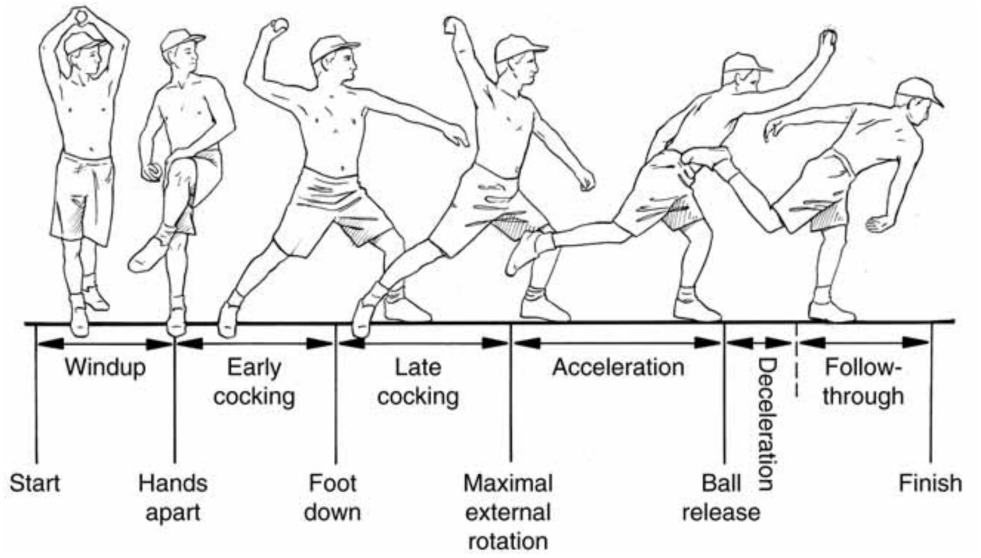

Figure 2: Phases of a pitch from Start to Finish (adapted from [8]).

over time and possibly link measurements of one single mo to injury in the next.

This study aims at investigating whether an indication of upcoming shoulder or elbow injuries can be given to an athlete prior to injury symptoms showing up in youth pitchers. It is hypothesised that injuries can be predicted by the inclusion of measurements of Range of Motion, Muscle Force, Ball Speed and Training Time, which, according to literature, are the most promising.

2 Methods

2.1 Study Population

The KNBSB (the Royal Dutch baseball federation) recruited participants from the national baseball team and the six baseball academies throughout the Netherlands. Inclusion criteria were being male baseball pitchers participating in the elite youth class. No selection was performed on the dominant arm, and no criteria on injury epidemiology were specified for inclusion. A population of 125 athletes satisfied the inclusion criteria throughout the data collection time, of which 118 participated in the study.

The setup of the paper is of exploratory research using existing longitudinal data of observational nature, meaning the conditions of the observed athletes are not altered for the study purposes. The data collection procedures were consistent with the Helsinki Declaration, and the local ethics committee approved them. Participants and their parents signed an informed consent for participation before the start of the measurements.

2.2 Data Collection

The dataset used has been initially captured for a study done by the KNBSB on the development of pitching speed, called the Fastball project. The complete set of data includes several variables; throwing speed, arguably the primary performance variable typically used in baseball, together with passive anatomic measurements of the athletes and self-reported questionnaire data can be found.

Measurements were performed twice a year: one at the end of March/start of April, and the other one at the end of September/start of October. The measurements started in the spring of 2014 over the total span of three years, resulting in six sessions (named trials A-F) each six months apart. Each trial, a variety of parameters were asked/measured. A little overview of each category of variables (a feature set) is made, together with the methodology of data collection.

General Data on the subject's name, age, length, weight and dominant arm was collected.

This feature set also contains information gained by questionnaire data about the level of play, which academy of training and the thrown types of pitch, together with the number of days and amount of hours played every week on average.

Self-reported questionnaires DASH, KJOC and WOSI questionnaires were reported by the athletes for each trial. These questionnaires were only filled in when a player reports being injured, and assess the level of injury. DASH is a disability index, KJOC is a measure of injury and performance, whereas WOSI is an instability index for the arm. These are commonly used questionnaires to assess upper-body injuries.

Injury Self-reported injury data was collected reporting general injury at the moment of data collection and injury in the six months prior to the trial. The latter was also asked more specific, where a player could indicate injury more specifically at the shoulder or the elbow joints.

In this paper, these variables are considered to be target features for the learning algorithm (more detailed in section 2.3.1).

Ball Speed Each player had to throw ten fastballs within one measurement, the speeds of which were measured in mph using a handheld speed gun. Sometimes more than ten measurements were done, in which case they are not considered for analysis.

Range of Motion (RoM) Range of Motion was measured in degrees using a goniometer and consists of measurements at the hip, the shoulder and the trunk. For both the hip and the shoulder, measurements were done on both the left and the right sides, measuring the joint's endo- and exo-rotation. The trunk's RoM is an axial measurement and was measured clockwise and counterclockwise.

Maximum Static Force (Force) Force measurements of the shoulder and the hip were measured in Newton (using a MicroFET® dynamometer, by Hoggan Scientific) doing three measurements of maximum contraction and averaging the three to obtain a single measurement. At the hip, they can be distinguished between hip abduction and hip adduction for both the left and the right side. At the shoulder joints abduction, exo- and endo- contractions were measured, and also measurements of the lower trapezius (LT), middle trapezius (MT) and serratus anterior (SA) were captured.

Laxity Laxity is a measure of looseness of a joint and can be the result of an injury. Measurements were done for the pinky, thumb, elbow and knees, together with a single measurement for the trunk. It is indicated as a binary measurement, indicating positive or negative looseness.

A year after the start of the data capturing (trial C), some changes to the measurement's protocol were made to specify the location of the injury and streamline the data collection protocol. The decision was, therefore, taken to limit the data processing to include only trials C to F.

A list of abbreviations explaining the name buildup of all the features can be found in appendix A.1.

2.3 Data Processing

The data is used to train machine-learning models to predict future injuries before they occur. Data analysis was performed in a classic cross-sectional study design first. The cross-sectional analysis was used to assess the quality of the dataset before using it for longitudinal prediction purposes.

The experimental procedure contained the following phases:

- Data extraction. All data was received in a raw format (.xlsx files) and had to be extracted and named consistently.
- Data pre-processing. The raw data was cleaned from null values by removal of some feature sets and imputing methods. It was then discretised using encoding methods. More about the data pre-processing phase can be found in appendix A.3.
- Feature selection. Features were selected as raw readings, and some were engineered with the help of past literature knowledge.

- Learning phase. In this phase, the model was fit on a learning set (60 % of the dataset).
- Test phase. Here the models were tested on their prediction performance.

All data was processed with Python (`v3.8` [28]). Pre-processing was mostly done using the Pandas library [29], the SciKit-learn package [30] and some custom functions. Because of the unbalance between dominant left (L) and right (R) armed athletes, the data is corrected by simplifying it into Dominant side (`Ds`) and non-Dominant side (`nD`). This way, more data can be used for categorisation.

2.3.1 Feature Selection

In machine learning, a feature is either a single measurable property or a mathematically computed property using one or more available variables. This paper uses supervised learning methods, which means data is labelled, and models can be assessed on their performance of prediction of these labels. The distinction can thus be made between predictor and target features. More specifically, predictors are used as the input of a model that predicts a target feature.

Predictor Features Some information about the engineered features is given. The rest of the features were all taken directly from the raw data.

- RoM The ratio of the endo- over the exo-rotation of the hip and shoulder joints was calculated. The total rotational motion (TRM) of the shoulder joint was found by summing up the two (endo + exo). An extra feature was identified as the difference in shoulder TRM between the dominant side and the non-dominant side.

- Ball Speed For each trial, the minimum, maximum, average and the standard deviation of the throws was calculated.

To investigate which features to select the correlations between the targets and the various features were first explored. Therefore, two different correlation techniques, the Pearson and Spearman correlation coefficients were used. Correlation is limited to measuring strength and direction of the relationship between two variables that is either linear (measured with the Pearson correlation coefficient) or monotonic, i.e. the relation is either entirely increasing or decreasing but not in a constant way (measured with the Spearman correlation coefficient).

Target Features The target variables were all extracted from the self-reported injury data, and are specified in more detail here.

N.B. The reported questions/answers have been translated from the Dutch language.

- Injury Shoulder : Q: "Do you suffer from an injury at the [...] shoulder?", asked for both the left and the right side. Response possibilities are ordinal categories:

- No
- Yes, occasionally
- Yes, regularly
- Yes, long-lasting

- Injury Elbow : Q: "Do you suffer from an injury at the [...] elbow?", asked for both the left and the right side. Response possibilities are the same as for the shoulder injury.

As previously mentioned, during pre-processing, it was chosen to rename left/right indicating which is the Dominant side (Ds) and which side is non-Dominant (nD).

The distribution of the target variables' possible answers for injured vs non-injured items was examined by reporting them on a table. After examination, it was decided to transform the target features into binary ones (moreover can be found in the results, section 3).

2.3.2 Learning Phase

The learning phase is the phase in which a learning model is chosen and is fitted to the training data. With training data the amount of the data used for training is meant; the rest is called test data and, as the name suggests, is used in the next phase of data processing (see section

2.3.3). Splitting the dataset was investigated using the `train_test_split` function from SciKit-learn, and it was concluded a test split higher than 30 % and smaller than 40 % would be ideal. More about this choice of is elaborated in appendix A.3.4.

Various algorithms were selected for comparison of their predictive powers. The choice of these algorithms was made according to the wide adoption of the learning methods they employ. For supervised learning classification problems like this one, the most used models are Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM) and Naive Bayes (NB) [31]. All algorithms were obtained by the open-source SciKit-learn package [30].

All models were fitted for both the cross-sectional as well as the longitudinal datasets (see appendix A.3.3 for information about the preparation of the two datasets). This way, a model could be selected to perform the rest of the analysis.

After model selection, various combinations of features (feature set) were fed to the algorithm. This way, the feature sets were fed separately as well as together to test the change in predictive power with each feature set.

2.3.3 Test Phase

During the test phase, the models were evaluated on the test split of the data. Metrics for evaluation of this type of classification problem were thoroughly analysed by Sokolova *et al.* in a paper on classification metrics [32], and an explanation of the most important ones can be found in appendix A.4. The evaluation of the classifier's performance was done by analysing two scores: the Cross-Validated F1 and F2 Score.

In general, the F-β Score is used to leverage the power of the Prediction (P) and Recall (R) scores to diminish both Type I and Type II errors by taking a harmonic mean of the two scores. For $\beta = 1$ (F1 Score) the P and R are given the same weight, whereas increasing β puts more weight on R. More information about these scores can be found in appendix A.4.

By making use of the F2 Score, this paper prioritises Recall, as False Negative cases (FN: predicted False, but actually injured) are the ones that are the least favourable to have.

Because of the relatively small amount of data, splitting the data could make for results slightly changing each time. For completeness, therefore, evaluation of the F1 and F2 Score was performed with the use of Stratified K-fold Cross-Validation (SCV with $K = 3$), which is cross-validation used for large imbalances of the target variables. Here the data is divided into K folds (chunks of data) which have the same percentage of each class of the target variable. The model is trained on $K-1$ folds and is then evaluated (tested on) the remaining fold. This evaluation is done for K folds and taking the mean score of these folds then gives the average F-β Score of that model. The choice of $K = 3$ comes forth from the investigated test split value as can be found in appendix A.3.4, resulting in a 1/3 test split ($\approx 33\%$) and a 2/3 training split. The mean values of the models' performance were consequently reported, together with their standard deviation.

Next to the model's performance metrics, also the fundamental value of the contribution of the individual features is sketched. For this purpose, the SHAP value is introduced using the `shap` package for Python [33]. SHAP values (or SHapley Additive exPlanations) are based on a value in game theory called Shapley value, which identifies the contribution of each "player" to the "game". What SHAP does is to quantify the contribution of each *feature* to the *model*. It, therefore, does not provide causality, rather the interpretability of a model [34].

It is important to note that the SHAP value is computed at a single observation and is, therefore, a local interpretation of the predictive model [34].

3 Results

Study population The study population consisted of 118 elite youth male pitchers with mean age of 15.4 (SD 1.5, range 11 to 19). Mean body height was 179.3 cm (SD 10.3, range 148.5 to 204.8) and mean body weight was 69.9 kg (SD 13.9, range 35.1 to 111.2). Of the 118 players, 18 were left-handed and none were ambidextrous. The average ballspeed thrown per athlete has a mean of 69.6 mph (SD 6.0, range 53.7 to 81.2), or 112.0 $\mathrm{km\,h^{-1}}$ (SD 9.7, range 86.4 to 130.7).

Included features A first examination of the Person and Spearmann correlations showed weak correlations of the predictor features with the injury measurements (the highest of which were in the range of 0.10 to 0.18). The greatest correlations were achieved for the Range of Motion (RoM) and the Maximum Static Force feature sets, resulting in these being selected for initial model training. Moreover, from literature also Ball Speed and Training Time resulted promising metrics for measuring training load and were thus selected as well.

The distribution of the target variables' possible answers for injured vs non-injured items can be found in Table 1. As can be expected from injury variables, the distribution is quite imbalanced and the injured categories have far too few items to be able to train a machine learning model accurately. Furthermore, the answer possibilities containing "Yes" can be distinguished by the frequency of injury ('Yes, occasionally' and 'Yes, regularly') and the duration of it ('Yes, long-lasting'). The answers can, therefore, be regarded as subjective, which makes them difficult to be interpreted by an algorithm. To correct for imbalance, lack of data and subjectivity, the ordinal categories were quantified into simple binary ones before the learning phase started. That means that 'No' was transformed to a value of 0, and the rest ('Yes') all take on a value of 1.

Table 1: Distribution of injury responses at the dominant side for trials C to F, before transforming them to binary features ('No' and 'Yes' only). Values shown are divided for the shoulder and the elbow joints.

	Shoulder	**Elbow**
No	170	168
Yes, occasionally	57	61
Yes, regularly	24	22
Yes, long-lasting	11	11

Moreover, the average time lost from sport due to injury over six months was examined. It was found that 65 % of injured athletes reported time lost from practices or games to be '1 to 7 days', and 10 % reporting 14 or more days.

Models performance comparison The prediction accuracy of the five chosen models is explored in the form of Stratified K-fold Cross-Validated (SCV) F-β scores (with $\beta = 1$ and $\beta = 2$). Results are shown separately for the Cross-Sectional and the Longitudinal analysis. The difference between the two comes forth from the way the data for these was prepared (see appendix A.3.3). The models used are Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM) and Naive Bayes (NB).

It can be seen how the model's performance stays below scores of 0.5 for all combinations of type of analysis (cross-sectional vs longitudinal), the model used (LR, DT, RF, SVM, NB) and injury classification (shoulder vs elbow). It can also be noted how some standard deviation (SD) values are of 0.1 or higher, indicating the spread of the scores obtained.

In the cross-sectional analysis (Figure 3A), the highest performing models for the shoulder are LR, DT, RF and NB with similar magnitudes, while SVM performs the worse. For the elbow, LR and NB perform similarly, succeeded by DT, RF and SVM. In this graph (Figure 3B), the models generally score higher for the elbow joint compared to the shoulder joint.

In the longitudinal analysis, the highest performing models are DT, LR and NB for the shoulder (Figure 4A), and DT, LR and RF for the elbow joint (Figure 4B). Both for the shoulder and the elbow it can be observed how the RF model performs worse than the DT one, which is especially noticeable for the shoulder. Also, SVM scores worse in both classifications compared to the other models, with scores always under 0.4.

Furthermore, no significant difference in prediction accuracy scores was observed between the cross-sectional (Figure 3) and the longitudinal analysis (Figure 4).

Input feature set performance The performance was then further analysed for longitudinal predictions. The effect of training the model on different feature sets is shown using the best performing models (DT and LR). Six sets were examined:

1. RoM only
2. Force only
3. Ball Speed only

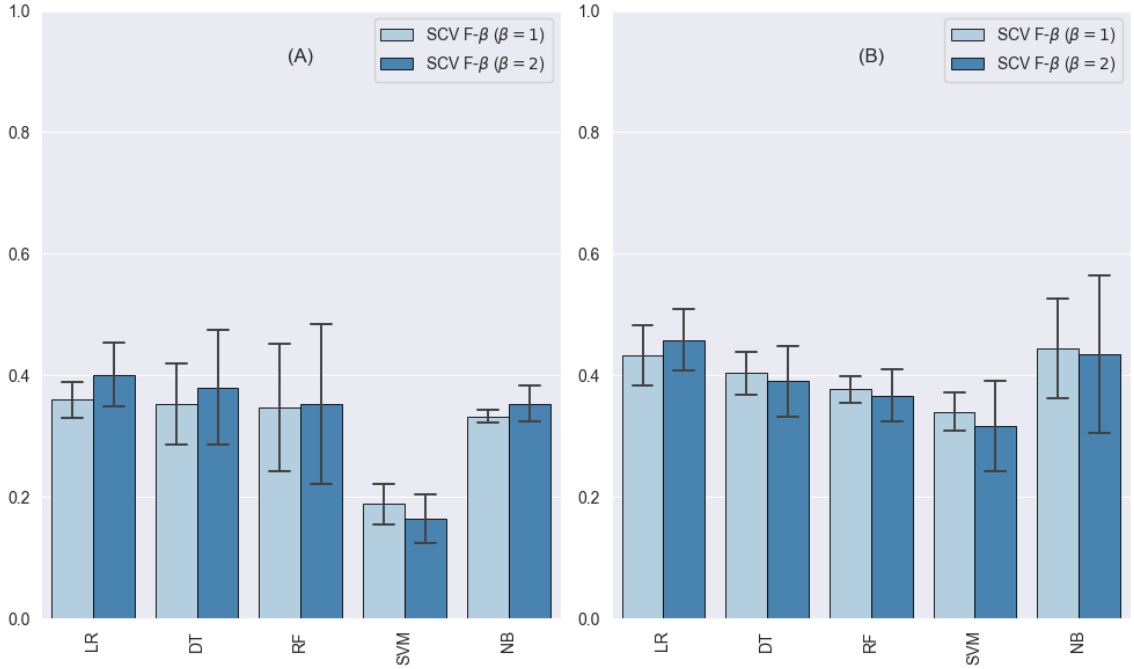

Figure 3: Stratified K-fold Cross-Validated F1 and F2 Score mean value (bar height) and SD (line) for different classification algorithms, plotted in a barplot format for the **cross-sectional analysis**. (A) shows values for injuries in the dominant shoulder, while (B) focuses on injuries for the dominant elbow. Models are Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM) and Naive Bayes (NB).

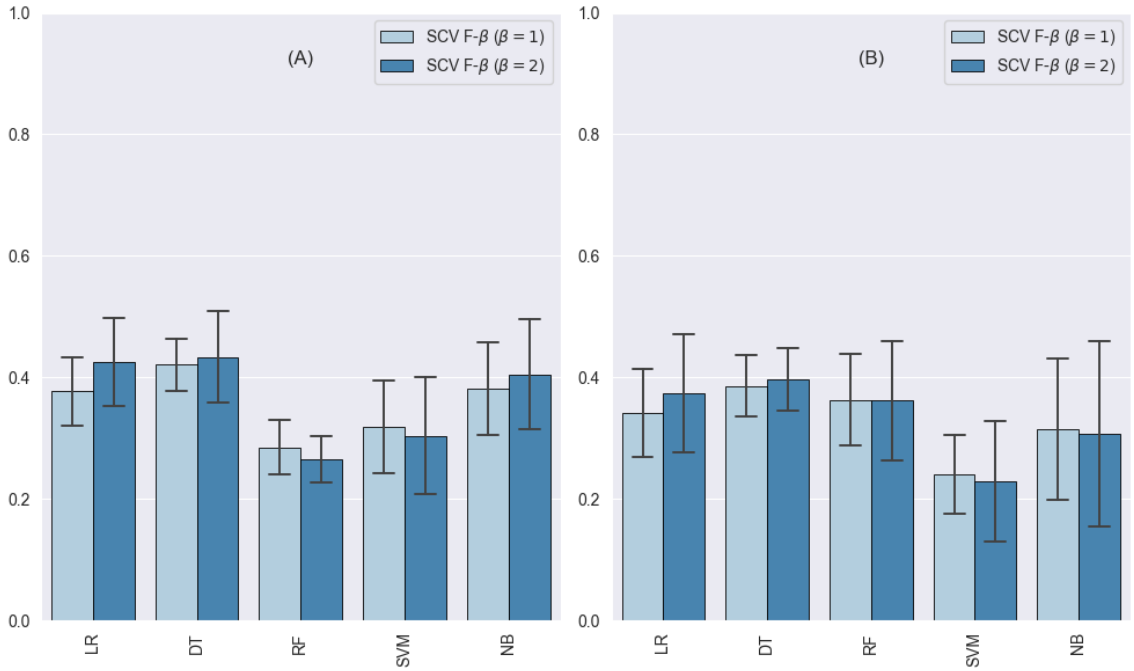

Figure 4: Stratified K-fold Cross-Validated F1 and F2 Score mean value (bar height) and SD (line) for different classification algorithms, plotted in a barplot format for the **longitudinal analysis**. (A) shows values for injuries in the dominant shoulder, while (B) focuses on injuries for the dominant elbow. Models are Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM) and Naive Bayes (NB).

4. Training Time only

5. RoM and Force

6. RoM, Force, Ball Speed and Training Time

The results for the DT model were analysed first. For the shoulder (Figure 5A), the feature sets 1 to 3 performed similarly. Set 4 shows the lowest performance, whereas the combinations shown by 5 and 6 scored the highest. For the elbow predictions (Figure 5B), the best performing feature set is number 3 with a mean score of approximately 0.5, while set 4 scores the lowest at around 0.2.

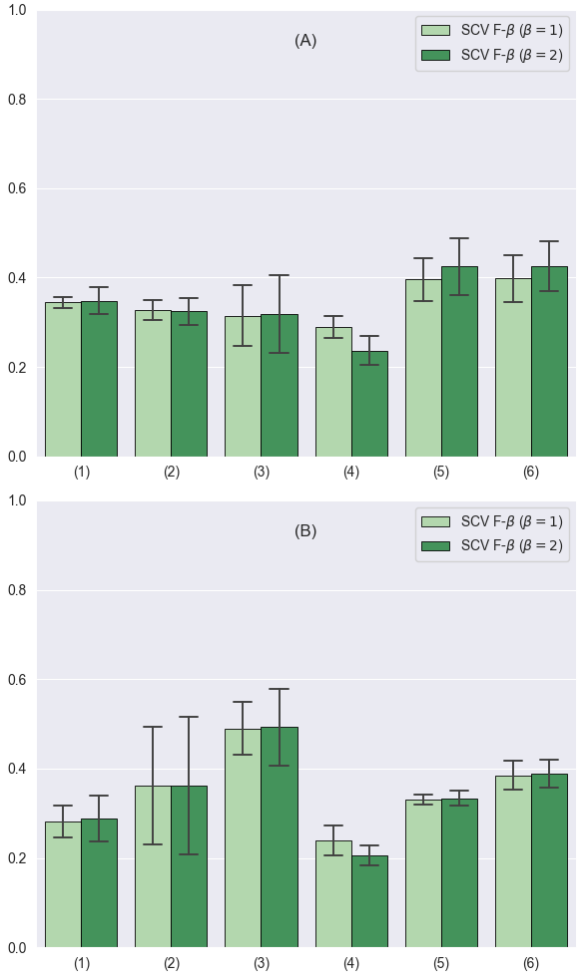

Figure 5: Stratified K-fold Cross-Validated F1 and F2 Score mean value (bar height) and SD (line) for different feature sets inputted in the `DecisionTree` model for longitudinal prediction. Options include: (1) RoM; (2) Force; (3) Ball Speed; (4) Training Time; (5) RoM and Force; (6) RoM, Force, Ball Speed and Training Time. (A) shows values for injuries in the dominant shoulder, while (B) focuses on injuries for the dominant elbow.

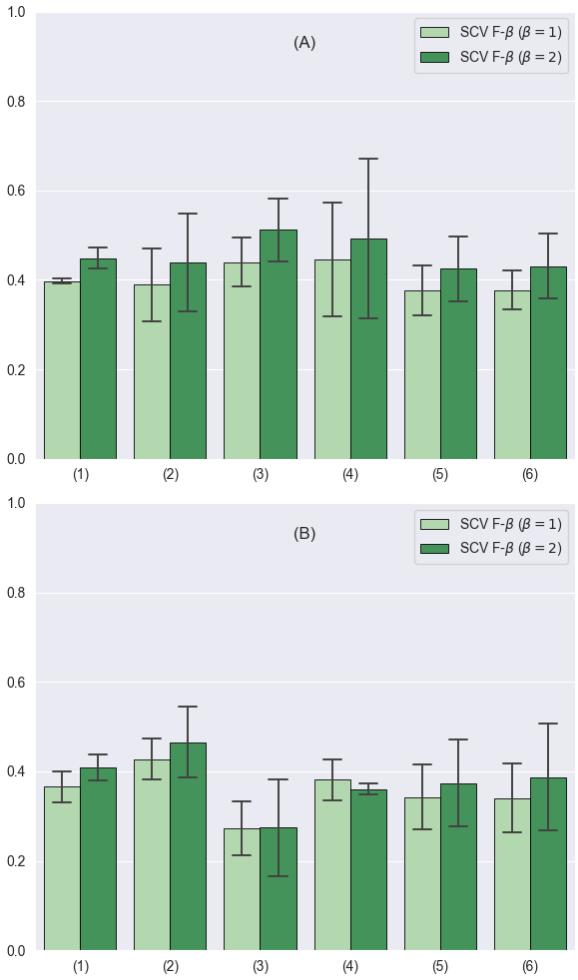

Figure 6: Stratified K-fold Cross-Validated F1 and F2 Score mean value (bar height) and SD (line) for different feature sets inputted in the `LogisticRegression` model for longitudinal prediction. Options include: (1) RoM; (2) Force; (3) Ball Speed; (4) Training Time; (5) RoM and Force; (6) RoM, Force, Ball Speed and Training Time. (A) shows values for injuries in the dominant shoulder, while (B) focuses on injuries for the dominant elbow.

Interestingly, set 2 has an SD of almost 0.2, indicating a big spread in score results for the model training on the Force set only.

The results for the LR model are then shown. For the shoulder (Figure 6A), all feature sets reported mean scores of 0.4 or higher, with set 3 and 4 performing somewhat higher at a mean score of 0.5. Notably, the SD of set 4 is very large compared to the rest of the sets. For the elbow (Figure 6B), it is shown how the sets all report a mean score just lower than 0.4, with an exception for set 3 performing worse (≈ 0.3) and set 2 reporting a higher mean score at around 0.5.

Individual feature contribution In order to identify the contribution of the individual features, the SHAP values for the two best performing models (DT and LR) were analysed. The models were trained using all four of the selected feature sets (RoM, Force, Ball Speed and Training Time), and the SHAP values for only the top twenty contributing individual features were plotted. A list of abbreviations explaining the name buildup of all the features can be found in appendix A.1.

The average impact on the output for the DT model, for the shoulder and elbow, is shown first. For the shoulder (Figure 7A), the hip's Exo RoM at the dominant side (`RoM_hip_exo_Ds`) is shown to be of most notable importance compared to the other features. For the elbow (Figure 7B) the biggest contribution was found to be given by the difference in shoulder's TRM between dominant and non-dominant sides (`RoM_sh_trm_change`), the amount of hours trained per week (`train_hours_week`) and the Exo RoM of the shoulder at the non-dominant side (`RoM_sh_exo_nD`) respectively.

The results for the shoulder and elbow joints using the LR model are then shown. The most important difference between the results in this plot and the regular SHAP plots of Figure 7 is that Figure 8 shows the relationship between the target's magnitude and positive or negative SHAP. Positive values show the model putting a greater weight towards the formation of an injury, and negative values indicate the model learning an injury is not forming. Figure 8A shows the results for the shoulder, with the hip force at both the dominant and non-dominant sides scoring high (`Force_hip_ab_nD`, `Force_hip_ab_Ds` and `Force_hip_ad_Ds`). It is worth noticing how the magnitudes for abduction (`Force_hip_ab_Ds`) and adduction (`Force_hip_ad_Ds`) have an opposite effect on the SHAP value. Figure 8B shows the results for the elbow, and here it is valuable to see how the hip strength (`Force_hip`) features are valued highest, after which numerous features of shoulder strength (`Force_sh`) are ranked high.

Mind that all these SHAP results show the model's weights for a single run. It is noted that, during runtime, the top-performing (highest SHAP values) features for the DT model were not consistently the same, as each run made the tree model learn different relations. The presented values are, therefore, to be interpreted with caution. Overall, for both shoulder and elbow, the Force features with higher SHAP values scored consistently higher compared to the other features when using the LR model.

4 Discussion

This study sought to explore if longitudinal data of elite youth pitchers can be used to predict injury before symptomatic pathology develops. In order to do so, the data was analysed first cross-sectionally and then longitudinally. The cross-sectional analysis was used to assess the quality of the dataset before using it for prediction purposes. The longitudinal analysis then more specifically focuses on the predictive power of the features toward future injury. After identifying which models performed best on longitudinal prediction, these models were trained to identify the contribution of the individual feature sets by selectively using them as predictors, and then by studying the SHAP values of the best performing models.

Cross-sectional analysis First, the cross-sectional results of the five selected models are analysed. The first thing that can be observed is the bad prediction score for all models used. It is concluded that combining the data of all athletes into a single model simplifies the problem too much. This way of structuring the data was necessary due to the scarcity of the data, and it does not take into account the individual differences between athletes. Limpisvasti *et al.* already concluded that injury forms when asymptomatic pathology is seen to exceed the individual's ability to compensate [8]. Verhagen and Gabbett also stressed using an individualised approach, saying performance and load measures have to be seen in relation to health, context and environment [14].

It is, therefore, thought that personalised injury prediction power lies not in a single point in time measurement, but in the development of one's pathology over time. Using an algorithm for time-series analysis of multiple points in time is believed to increase the ability to spot deterioration and thus predict future injury.

It is also worth mentioning that the cross-

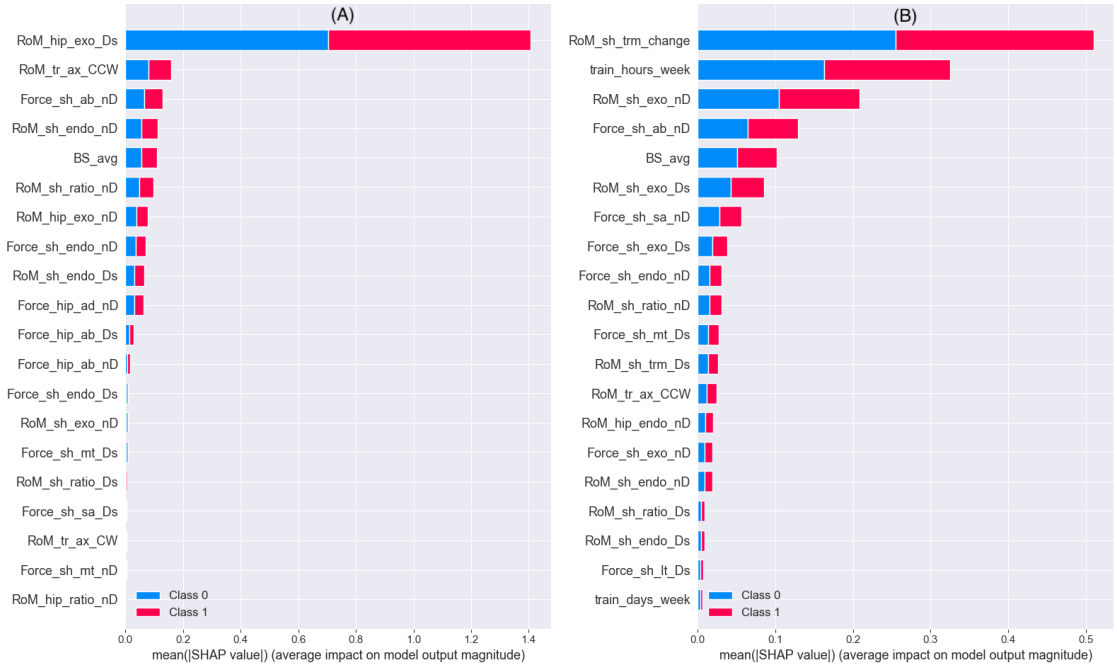

Figure 7: SHAP values reported for the `DecisionTree` model trained with the RoM, Force, Ball Speed and Training Hours feature sets. Greater values show the model putting a greater weight towards a correct prediction. Color differences represent the class it tends to predict more often using this weight (with a binary target these are equally sized). (A) shows values for injuries in the dominant shoulder, while (B) focuses on injuries for the dominant elbow. A list of abbreviations explaining the name buildup of all the features can be found in appendix A.1.

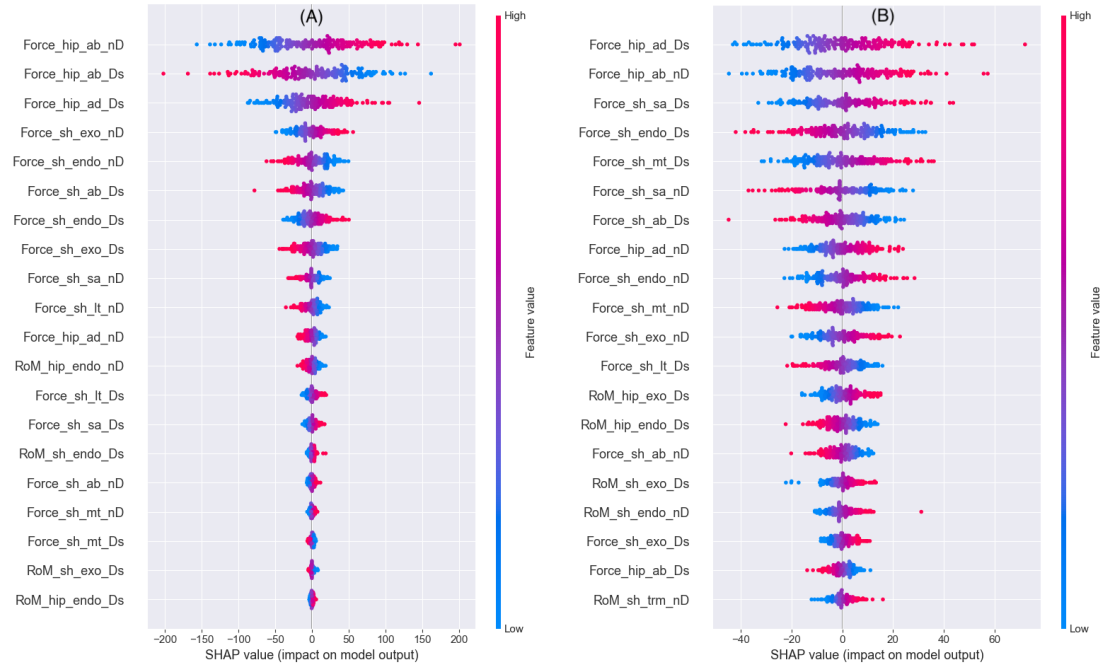

Figure 8: SHAP values reported for the `LogisticRegression` model trained with the RoM, Force, Ball Speed and Training Hours feature sets. Positive values show the model putting a greater weight towards the formation of an injury, and negative values indicate the model learning an injury is not forming. Color differences indicate changes in magnitude of the feature from low (blue) to high (red). (A) shows values for injuries in the dominant shoulder, while (B) focuses on injuries for the dominant elbow. A list of abbreviations explaining the name buildup of all the features can be found in appendix A.1.

sectional analysis did not yield better results compared to the longitudinal analysis, which should be the case given that the feature sets used are considered to be promising by many [8, 9, 22–26].

The cross-sectional results for the elbow joint show slightly better scores than those for the shoulder joint. That is thought to be because of the differences between the elbow joint and the shoulder joint. The elbow joint is a hinge joint, able to perform only movements of flexion and extension, whereas the shoulder's ball-and-socket joint allows for multiaxial movement [35]. It is therefore believed that the shoulder joint's complexity could be a relevant factor when predicting injury and that the less complicated elbow joint injuries are therefore more easily identifiable by these simplistic algorithms.

Model selection The prediction performance of the five selected algorithms trained on the longitudinal data is then analysed. The goal of this analysis was to select the best performing algorithm for further examination of the contribution of the individual feature sets.

It is unusual to see RF perform worse than DT (particularly noticeable in the shoulder joint), as the former should be an improved algorithm of the latter, making predictions by averaging the prediction of individual subtrees. The low prediction scores are related to RF algorithms having difficulties dealing with unbalanced classes, as the classification bias found in the subtrees is emphasised [36]. The results also illustrate how SVM performs poorly on predictions of both joints' injuries. The SVM results in this paper have been obtained using a linear kernel, which is not a flexible method because of its use of linear separation lines. However, using higher-dimensional kernels, the model performed even worse, as higher dimensionality is related to overfitting on noisy data. It is therefore thought to be an optimisation issue.

According to the scores for longitudinal injury prediction, the two best-performing algorithms are DT and LR for both the shoulder and the elbow injuries. The analysis of the importance of the individual feature sets is, therefore, performed with both the DT and LR models for both joints.

Optimisations By analysing the deficient prediction performance of the results, it was determined that prediction of injury in the shoulder or the elbow joints is quite a challenging endeavour when using simplified approaches. Multiple reasons can be thought of to explain the low accuracy of the trained models, the main of which is related to the frequency of data collection, which is a limitation further elaborated upon later in this text. Error in the real-life data is always present, and it can be attributed to the detail of the data, noise in the measurements or the stochastic nature of the algorithms used for modelling [31]. Some improvements were considered during the data processing phase.

An improvement considered has been the careful optimisation of each single machine learning algorithm's parameter. Given the range of results that were outputted in preliminary tests, optimisation was, at last, not regarded as an option as it mostly yielded only marginal improvements over default parameters. Bittencourt *et al.* investigated ways of improving injury predictions and stated that injuries should be investigated as complex systems, by studying the interaction of multiple facets in relation to each other [37]. Based on their findings this paper also tested a complex systems approach with a sequential Artificial Neural Network (ANN), which resulted in a 5% to 8% increase in cross-validated F1 Score for the longitudinal dataset alone. That is due to the nature of ANNs, taking into account complex interactions between variables. However, these results indicate that some other causes should be identified, one of which is the used dataset. Improvements on data collection are highlighted in the future recommendations paragraph.

Important features This paper also sought to identify the most determining features by using different input feature sets and studying the SHAP values for the DT and the LR models trained on longitudinal predictions. The reader is reminded that no intervention was performed during data collection, as this is a strictly observational study.

In order to understand the features in relation to the pitching motion and injury due to pitch-

ing, the concept of the kinetic chain is introduced. The kinetic chain is a coordinated sequence of segments, in which the lower extremities and the trunk act as the force generators, the shoulder then funnels this force to the arm, which acts as a force delivery mechanism [38]. During a pitch, there is an ideal proximal to distal sequence by which greater ball speeds can be achieved, and a reduction in throwing arm injury is observed [25].

For the DT model trained on the shoulder, we observed particular importance given to the external rotation RoM at the hip's dominant side. This relation is in line with a previous paper by Oliver and Weimar, indicating a positive relationship exists between the external hip rotation and the scapular external rotation [39]. That is because dominant hip external rotation is crucial for optimal foot placement to allow for the preparation of the shoulder's maximum external rotation during pitching [40]. The link to injury was explored by Sher *et al.*, who did not find an association with the hip's external rotation directly. However, they did show a positive relationship exists between dominant hip extension and shoulder injury, with injury occurring for higher values of external shoulder rotation [41]. The paper concludes recommending training to improve hip flexibility (non-dominant internal hip rotation and dominant hip extension) to prevent injury [41]. It is therefore concluded that maximising external shoulder rotation is the feature more likely to cause injury, especially when the hip does not allow for optimal foot placement by having a limited external range of motion.

Looking at the most promising variables on shoulder injury from the LR model, we can observe how hip force features score relatively high at both sides (dominant and non-dominant). That confirms previous longitudinal research on the maximum contraction force in the hip by Zeppieri et al., who found a significant decrease in hip strength from preseason to postseason measurements. They argue that the reduction in hip strength (especially abduction) may result in an improper sequencing of force production which may put an increased load on the upper extremities [42]. It is thought that alterations in the normal kinetic chain motion would mostly be felt by the shoulder joint, as it has a low capacity of force-generation and primarily depends on proper positioning and muscle activation to control joint translations [38].

As already mentioned, a decrease in the efficient force sequencing results in injury and in a decreased distal velocity [25, 38], which means decreased pitching speed. Results for the LR model trained on Ball Speed only showed moderate improvements in performance scores compared to other feature sets. The average Ball Speed feature was found to be a slight contributor in the DT model but did not show up in the top twenty contributors for the LR model. Research done on the development of force during a throw (i.e. not maximum contraction force) did find lower energy transfer (i.e. a lower Ball Speed) for time variations between segments and argue that increases in Ball Speed can not be related to reductions in injury [43]. The paper argues that there may be an optimum timing between the forces acting on the shoulder (known to increase injury likelihood [38]) and the elbow to produce sufficiently high Ball Speeds [43], with offsets aiding injury progression over ball speed [25, 38]. Although promising in literature, these results can not be used to draw a conclusion confirming a direct relationship between shoulder injury and ball speed exists, limiting the usefulness of Ball Speed *in regression models*.

The findings on elbow injury were quite inconsistent when considering Ball Speed only, with the DT model showing an increased prediction score when considering it the only input set, whereas the LR model did show a decrease in performance compared to other feature sets. Unlike the findings for shoulder injuries, for the elbow joint it is known that higher pitch velocity is related to elbow injuries [22, 23]. It is thought that the LR model does not show high scoring for predictions of elbow injury using the Ball Speed feature because of the same reason as for the shoulder. That is, Ball Speed measurements do show an optimum timing between shoulder and elbow, making it a feature not appreciated by regression analysis but detected by a tree algorithm like DT.

The results for the DT features (on elbow injury) showed a pronounced contribution given by the reported amount of training hours per week. Training Time was included as the closest indication of the experienced training load available

in the dataset. A previous study by Olsen *et al.* showed elbow injury being significantly related to fatigue and pitch count [22], both being variables arguably related to training load. Although the Training Time feature was found to be an important one for DT prediction of elbow injuries, using it as the only predictor feature did show lower performance compared to other feature sets. It is therefore concluded that the measured Training Time was not a good predictor for injury overall and that more consistent ways of measuring training load should be used for injury prediction.

Results indicated that the difference between the shoulder's TRM at the dominant side and the non-dominant side has a big contribution on prediction performance of elbow injury using the DT model. This is in line with conclusions by Garrison *et al.*, indicating measurements of the shoulder's TRM are related to elbow injury, with a lower TRM indicating a higher chance on UCL injury [24]. It is believed that a low TRM makes for incorrect shoulder rotation during a pitching motion [25], disrupting the kinetic motion and increasing compensation load by the elbow joint. The highest contributing variables for the LR model also confirm this: the hip and shoulder force features (strength). According to Zeppieri *et al.*, a reduction in hip strength increases compensation load further down the sequential chain [42]. The same reasoning can be applied to explain the shoulder strength feature. It is known that, during a throwing motion, shoulder muscles generate a significant load forward which, coupled to adduction of the humerus and rapid elbow extension, generate the high propulsive distal velocities [44]. If shoulder strength results to be lacking, the elbow is thought to compensate, increasing the load on the elbow joint and thus increasing the chance of injury.

Strengths Although low prediction performance was achieved, this paper also comes with some strengths. First of all, pitchers are mostly trained on their pitching motion, which makes that the effects found in this dataset are most likely attributable to pitching in contrast to slidings, falling or colliding with other players. Furthermore, according to Agresta *et al.*, a prospective study investigating injuries was missing in the research community [27]. This paper, although of relatively small size, is a good starting point to fill the need for this type of study.

The most exciting aspect of this study is that it compares classification methods in direct link with pitching mechanics and values used in the baseball world. Using a systematic approach of modelling and establishing results, this paper builds up toward more robust and more informative datasets in the future, intending to make the real-world application of prediction algorithms for injury detection possible in the near future.

Limitations and future recommendations Some limitations exist for this paper. These should be carefully considered for improvements in future studies.

The range of answer possibilities given to the participants in order to determine the seriousness of an injury was not clear enough, with answers alternating frequency and duration of injury. It is hence thought that the answers were somewhat subjective, making it difficult to establish proper cause-effect relations.

The definition of what injury is should be clear, specific and measurable. It is, therefore, recommended that a physician certifies injury for data labelling.

The dataset used in this work is regarded as having a scarce size for a machine learning model to perform adequately on. At first, this data was intended for a study on the performance of youth athletes. When using the same dataset for injury prediction, the frequency of the collection moments is regarded to be too low, resulting both in loss of information (richness) and data scarcity (amount).

For one, the low frequency makes for difficult detection of an injury, as injury length is measured as the time lost from practice or games. The dataset showed most time lost due to injury to be between 1 to 7 days, which is in line with findings by Garrick *et al.*, who reported 27 % of high school baseball players lost at least five days per injury [45]. A study on the epidemiology of collegiate baseball players yielded similar results, reporting that almost 3/4 of injuries resulted in less than 7 days lost from sport and 1/4 of injuries resulting in more than 21 days lost per in-

jury [46]. No special mention about pitchers was found in the studies above, making it difficult to make a one-on-one comparison possible. The average time lost to injury is hence multiple times smaller than the precision that can be measured with a frequency of measurement of once in six months. It could, therefore, be possible that one can recover from an injury and sustain a new one all between one measurement and the next one. The data would not be able to distinguish between two separate injuries.

Also, it is a possibility that, due to a long time between measurements, an athlete could sustain damages resulting in injury some time (e.g. a month) before the next measurement. In that case, the injury could be reported, whereas the pathology which can predict the injury was not measured, making it impossible to find cause-effect relations. A higher data collection frequency would solve this problem.

Furthermore, one more advantage of improving the frequency (and thus the amount) of collected data is for the use of time-series data, i.e. series of data progressing in time. This is in line with the previous reporting in the paragraph on cross-sectional analysis, indicating time-series data can be used to spot an individual's deterioration and thus predict future injury.

The most significant improvement that could thus be made is related to collecting data with a higher frequency. It is recommended to collect data with a frequency of one to two weeks, having the benefit of both more data being available, and of richer details being available to establish the relationship of cause-effect.

One more limitation was not being able to use the self-reported questionnaires for prediction purposes, as they were only asked when the injury was already perceived (i.e. already symptomatic). This makes for the dataset not showing the difference in answers between asymptomatic and symptomatic pathologies. Another effect is that because of this option to skip the questionnaires (if not injured) athletes could be psychologically incentivised not to report an injury at all, further skewing the results. It is recommended to consider not providing a skip option in future data collection.
If new data is collected to build upon this research, next to taking into account this paper's limitations, it could be viable also to explore the following options.

As previously mentioned, using a complex systems approach to data analysis could yield improved performance results. It is therefore recommended to investigate ways to implement this into future studies. This could be in the form of a sequential Artificial Neural Network, but further research on the topic is needed to reveal the best methods for the analysis of such data.

Analysis of the usefulness of the predictor features already showed the importance of the solid execution of the kinetic chain. According to Fleisig *et al.* and Scarborough *et al.* the pattern in kinematic sequence during a pitch can be associated with high torque production in the shoulder and the elbow [15, 47]. Papers by Burkhart *et al.* and Fortenbaugh *et al.* noted that compensation occurs when errors are made in the sequence, resulting in injuries [25, 38]. It could, thus, be especially exciting to look at the incidence of injuries in relation to kinematic data of a throw, to locate potential missteps in the throwing motion and longitudinally link these to future injury.

5 Conclusion

This study explored if longitudinal data of statically measured features can be used for injury prediction before symptomatic pathology develops. The prediction accuracy for the models on both the elbow and the shoulder injuries did not give significant results. Future recommendations propose three main improvements: higher frequency of measurements, the inclusion of kinematic data and the use of complex models for analysis.

It is concluded that scarce and low-frequency data yields bad results when predicting injury of the shoulder and the elbow. It was also found that measurements reflecting a single point-in-time are not capable of giving the data richness needed for analysis of the development of someone's pathology, which can become an injury with time.

Findings also backed previous literature, showing relations indicating that proper development of kinetic chain motion is crucial to avoid injury. Measurements of hip range of motion and hip

strength are found to be good predictors of shoulder injury. Total rotational motion (TRM) of the shoulder, as well as muscle strength in the hip and the shoulder, are thought to be of importance for the detection of elbow injury. Also, ball speed measurements could aid in the prediction of elbow injuries when using models not employing regression learning. Last, training time measurements used as an indication of training load did not provide the expected prediction results.

To summarise, more frequent data collection and the addition of kinetic chain analysis data could help state-of-the-art algorithms with upper extremity injury prediction. Examining the data as time-series developments would advance analysis of cause-effect relationships in the future, helping elite youth baseball pitchers avoid injuries and keep their performance ready for top-level play.

References

[1] M. Lewis, *Moneyball: The art of winning an unfair game*. WW Norton & Company, 2004.

[2] D. G. Liebermann, L. Katz, M. D. Hughes, R. M. Bartlett, J. McClements, and I. M. Franks, "Advances in the application of information technology to sport performance," *Journal of Sports Sciences*, vol. 20, no. 10, pp. 755–769, 2002.

[3] W. van Mechelen, H. Hlobil, and H. C. Kemper, "Incidence, severity, aetiology and prevention of sports injuries," *Sports Medicine*, vol. 14, no. 2, pp. 82–99, 1992.

[4] CBS, "Bevolkingsteller CBS." https://www.cbs.nl/nl-nl/visualisaties/bevolkingsteller, accessed on June 2020.

[5] TUDelft, "Six million euros in research funding to develop technology that prevents sports injuries." https://www.tudelft.nl/io/onderzoek/research-labs/emerging-materials- lab/citius-altius-sanius, accessed on May 2020.

[6] S. D. Mair, T. L. Uhl, R. G. Robbe, and K. A. Brindle, "Physeal changes and range-of-motion differences in the dominant shoulders of skeletally immature baseball players," *Journal of Shoulder and Elbow Surgery*, vol. 13, no. 5, pp. 487–491, 2004.

[7] K. Meister, T. Day, M. Horodyski, T. W. Kaminski, M. P. Wasik, and S. Tillman, "Rotational motion changes in the glenohumeral joint of the adolescent/Little League baseball player," *American Journal of Sports Medicine*, vol. 33, no. 5, pp. 693–698, 2005.

[8] O. Limpisvasti, N. S. ElAttrache, and F. W. Jobe, "Understanding shoulder and elbow injuries in baseball," *Journal of the American Academy of Orthopaedic Surgeons*, vol. 15, no. 3, pp. 139–147, 2007.

[9] S. Lyman, G. S. Fleisig, J. W. Waterbor, E. M. Funkhouser, L. Pulley, J. R. Andrews, E. D. Osinski, and J. M. Roseman, "Longitudinal study of elbow and shoulder pain in youth baseball pitchers," *Medicine and Science in Sports and Exercise*, vol. 33, pp. 1803–1810, 11 2001.

[10] P. S. Krebs, S. Zinkgraf, and S. J. Virgilio, "Predicting competitive bicycling performance with training and physiological variables.," *The Journal of sports medicine and physical fitness*, vol. 26, pp. 323–330, 12 1986.

[11] A. G. Scrimgeour, T. D. Noakes, B. Adams, and K. Myburgh, "The influence of weekly training distance on fractional utilization of maximum aerobic capacity in marathon and ultramarathon runners," *European Journal of Applied Physiology and Occupational Physiology*, vol. 55, pp. 202–209, 3 1986.

[12] I. Mujika, J. C. Chatard, T. Busso, A. Geyssant, F. Barale, and L. Lacoste, "Effects of training on performance in competitive swimming.," *Canadian journal of applied physiology = Revue canadienne de physiologie appliquée*, vol. 20, no. 4, pp. 395–406, 1995.

[13] C. Foster, E. Daines, L. Hector, A. C. Snyder, and R. Welsh, "Athletic performance in relation to training load," *Wisconsin Medical Journal*, vol. 95, no. 6, pp. 370–374, 1996.

[14] E. Verhagen and T. Gabbett, "Load, capacity and health: critical pieces of the holistic performance puzzle," *British Journal of Sports Medicine*, vol. 53, pp. 5–6, 1 2019.

[15] S. Fleisig, Glenn, J. R. Andrews, C. J. Dillman, and R. F. Escamilla, "Kinetics of Baseball Pitching with Implications About Injury Mechanisms," *The American Journal of Sports Medicine*, vol. 23, no. 2, pp. 233–239, 1995.

[16] D. Lintner, T. J. Noonan, and W. B. Kibler, "Injury Patterns and Biomechanics of the Athlete's Shoulder," *Clinics in Sports Medicine*, vol. 27, no. 4, pp. 527–551, 2008.

[17] K. E. Wilk, K. Meister, and J. R. Andrews, "Current concepts in the rehabilitation of the overhead throwing athlete," *American Journal of Sports Medicine*, vol. 30, no. 1, pp. 136–151, 2002.

[18] A. Miniaci, A. T. Mascia, D. C. Salonen, and E. J. Becker, "Magnetic resonance imaging of the shoulder in asymptomatic professional baseball pitchers," *American Journal of Sports Medicine*, vol. 30, no. 1, pp. 66–73, 2002.

[19] P. M. Connor, D. M. Banks, A. B. Tyson, J. S. Coumas, and D. F. D'Alessandro, "Magnetic resonance imaging of the asymptomatic shoulder of overhead athletes: A 5-year follow-up study," *American Journal of Sports Medicine*, vol. 31, no. 5, pp. 724–727, 2003.

[20] S. Conte, R. K. Requa, and J. G. Garrick, "Disability days in Major League Baseball," *American Journal of Sports Medicine*, vol. 29, no. 4, pp. 431–436, 2001.

[21] M. C. G. Ciccotti, K. M. Pollack, M. C. G. Ciccotti, J. D'Angelo, C. S. Ahmad, D. Altchek, J. Andrews, and F. C. Curriero, "Elbow Injuries in Professional Baseball: Epidemiological Findings from the Major League Baseball Injury Surveillance System," *American Journal of Sports Medicine*, vol. 45, no. 10, pp. 2319–2328, 2017.

[22] S. J. Olsen, G. S. Fleisig, S. Dun, J. Loftice, and J. R. Andrews, "Risk factors for shoulder and elbow injuries in adolescent baseball pitchers," *American Journal of Sports Medicine*, vol. 34, no. 6, pp. 905–912, 2006.

[23] D. Kurokawa, T. Muraki, H. Ishikawa, K. Shinagawa, H. Nagamoto, H. Takahashi, N. Yamamoto, M. Tanaka, and E. Itoi, "The Influence of Pitch Velocity on Medial Elbow Pain and Medial Epicondyle Abnormality Among Youth Baseball Players," *The American Journal of Sports Medicine*, p. 036354652091491, 2020.

[24] J. C. Garrison, M. A. Cole, J. E. Conway, M. J. MacKo, C. Thigpen, and E. Shanley, "Shoulder range of motion deficits in baseball players with an ulnar collateral ligament tear," *American Journal of Sports Medicine*, vol. 40, no. 11, pp. 2597–2603, 2012.

[25] D. Fortenbaugh, G. S. Fleisig, and J. R. Andrews, "Baseball pitching biomechanics in relation to injury risk and performance," *Sports Health*, vol. 1, no. 4, pp. 314–320, 2009.

[26] K. E. Wilk, L. C. MacRina, G. S. Fleisig, R. Porterfield, C. D. Simpson, P. Harker, N. Paparesta, and J. R. Andrews, "Correlation of glenohumeral internal rotation deficit and total rotational motion to shoulder injuries in professional baseball pitchers," *American Journal of Sports Medicine*, vol. 39, no. 2, pp. 329–335, 2011.

[27] C. E. Agresta, K. Krieg, and M. T. Freehill, "Risk Factors for Baseball-Related Arm Injuries: A Systematic Review," *Orthopaedic Journal of Sports Medicine*, vol. 7, no. 2, pp. 1–13, 2019.

[28] "Python Software Foundation." Python Language Reference, version 3.8. Available at http://www.python.org.

[29] "pandas-dev/pandas." The Pandas Development Team. Available at https://doi.org/10.5281/zenodo.3509134, 2 2020.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[31] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. Springer Texts, 2006.

[32] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing and Management*, 2009.

[33] "slundberg/shap." Available at https://github.com/slundberg/shap#citations, 2020.

[34] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, pp. 4765–4774, 2017.

[35] E. N. Marieb and K. Hoehn, *Human Anatomy & Physiology*. Pearson Education Limited, 10th ed., 2016.

[36] C. Chen, A. Liaw, and L. Breiman, "Using Random Forest to Learn Imbalanced Data," *Discovery*, no. 1999, pp. 1–12, 2004.

[37] N. F. Bittencourt, W. H. Meeuwisse, L. D. Mendonça, A. Nettel-Aguirre, J. M. Ocarino, and S. T. Fonseca, "Complex systems approach for sports injuries: Moving from risk factor identification to injury pattern recognition - Narrative review and new concept," *British Journal of Sports Medicine*, vol. 50, no. 21, pp. 1309–1314, 2016.

[38] S. S. Burkhart, C. D. Morgan, and W. B. Kibler, "The disabled throwing shoulder: Spectrum of pathology part III: The SICK scapula, scapular dyskinesis, the kinetic chain, and rehabilitation," *Arthroscopy - Journal of Arthroscopic and Related Surgery*, vol. 19, no. 6, pp. 641–661, 2003.

[39] G. D. Oliver and W. Weimar, "Hip range of motion and scapula position in youth baseball pitching pre and post simulated game," *Journal of Sports Sciences*, vol. 33, no. 14, pp. 1447–1453, 2015.

[40] S. R. Tippett, "Lower extremity strength and active range of motion in college baseball pitchers: A comparison between stance leg and kick leg," *Journal of Orthopaedic and Sports Physical Therapy*, vol. 8, no. 1, pp. 10–14, 1986.

[41] S. Scher, K. Anderson, N. Weber, J. Bajorek, K. Rand, and M. J. Bey, "Associations among hip and shoulder range of motion and shoulder injury in professional baseball players," *Journal of Athletic Training*, vol. 45, no. 2, pp. 191–197, 2010.

[42] G. Zeppieri, T. A. Lentz, M. W. Moser, and K. W. Farmer, "Changes in Hip Range of Motion and Strength in Collegiate Baseball Pitchers Over the Course of a Competitive Season: a Pilot Study.," *International journal of sports physical therapy*, vol. 10, no. 4, pp. 505–13, 2015.

[43] M. A. Urbin, G. S. Fleisig, A. Abebe, and J. R. Andrews, "Associations between timing in the baseball pitch and shoulder kinetics, elbow kinetics, and ball speed," *American Journal of Sports Medicine*, vol. 41, no. 2, pp. 336–342, 2013.

[44] F. S. Chen, V. A. Diaz, M. Loebenberg, and J. E. Rosen, "Shoulder and Elbow Injuries in the Skeletally Immature Athlete," *Journal of the American Academy of Orthopaedic Surgeons*, vol. 13, pp. 172–185, 5 2005.

[45] J. G. Garrick and R. K. Requa, "Injuries in High School Sports," *PEDIATRICS*, vol. 61, pp. 465–469, 3 1978.

[46] E. G. McFarland and M. Wasik, "Epidemiology of Collegiate Baseball Injuries," *Clinical Journal of Sport Medicine*, vol. 8, pp. 10–13, 1 1998.

[47] D. M. Scarborough, S. E. Linderman, J. E. Sanchez, and E. M. Berkson, "Baseball Pitchers' Kinematic Sequences and Their Relationship to Elbow and Shoulder Torque Production," *Orthopaedic Journal of Sports Medicine*, vol. 7, no. 7_suppl5, p. 2325967119S0042, 2019.

[48] P. Sengalrayan, "Injury, Performance and Machine Learning Aspects of the Analysis of the Tennis Serve," tech. rep., TUDelft, 2020.

A Appendix

A.1 List of Abbreviations

Here the abbreviated names for the presented variables are found. Combinations of the different parts can be found in a feature, where the parts are separated by an underscore (_). e.g. `Force_hip_ad_Ds` indicates it is a Maximum Static Force (`Force`) measurement on the hip joint (`hip`) during an adduction movement (`ad`) at the dominant side of the athlete (`Ds`).

Table 2: List of Abbreviations

Code	Variable	Computation
`Ds`	Dominant side of body	
`nD`	non Dominant side of body	
`RoM`	Range of Motion	
`Force`	Maximum Static Force	
`el`	elbow joint	
`sh`	shoulder joint	
`tr`	trunk	
`hip`	hip joint	
`CW`	clockwise rotation	
`CCW`	counter-clockwise rotation	
`exo`	external rotation	
`endo`	internal rotation	
`ratio`	ratio of exo over endo RoM	`exo` / `endo`
`trm`	total rotational motion of a joint	`exo` + `endo`
`trm_change`	difference of total rotational motion	`trm_Ds` - `trm_nD`
`ax`	axial	
`ab`	abduction	
`ad`	adduction	
`lt`	lower trapezius	
`mt`	middle trapezius	
`sa`	serratus anterior	
`BS`	Ball Speed	
`avg`	average	`average(BS1 to BS10)`
`min`	minimum	`minimum(BS1 to BS10)`
`max`	maximum	`maximum(BS1 to BS10)`
`train_hours_week`	amount of hours trained per week	
`train_days_week`	amount of days of training per week	
`shap`	SHapley Additive exPlanations	

A.2 Data Exploration

Exploration of the dataset was done to analyse the distribution of data.

The data used in this study is longitudinal with equally spaced observations. It contains an unequal amount of measurements for each participant, as not all participants performed all 6 tests. A distribution of the amount of trials attended by each athlete can be seen in Figure 9.

The figure clearly shows how the measurements were not attended regularly, with just 16 athletes having attended all six trials. It is also noticeable that the attended trials were not always of consecutive measurements in time (e.g. an athlete having attended two trials could have attenden trials A and D, thus having missed B and C). Players dropped out for various reasons, like outgrowing the dataset's target age or simply because of voluntary drop-out from the measurements or from the sport.

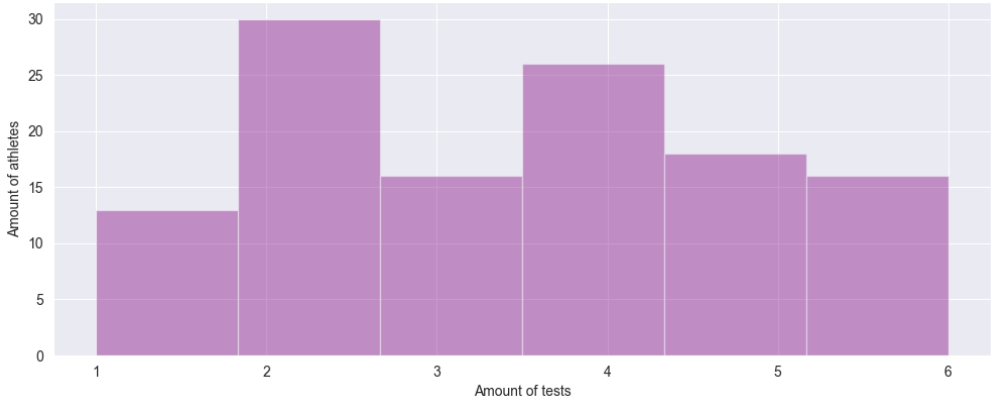

Figure 9: Distribution of amount of trials registered per athlete. The graph shows how many athletes fall in each category (1 to 6 measurements attended).

A.3 Data Pre-Processing

A.3.1 Removing Nulls

A first step taken in order to clean the dataframe was to remove trials without any content, Null values.

Also, outliers in ball speeds, length and weight were removed for values outside of the range of 3 standard deviations from the mean.

A first plot of the missing data (Figure 10) shows all the extracted data on the horizontal axis and all the data rows on the vertical axis. The available points are shown in purple, whereas with yellow the missing data is indicated.

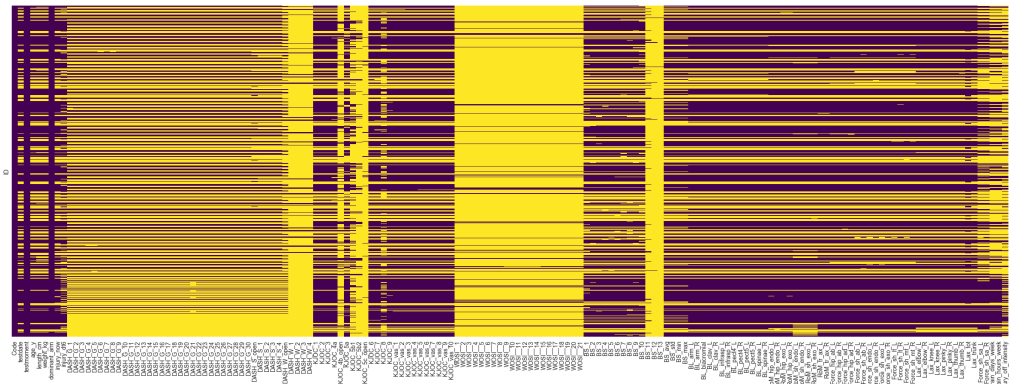

Figure 10: Nulls in the raw extracted dataset. The rows (items) and columns respectively are the Y and X axis. Purple blocks show valid data, whereas yellow blocks show missing data (i.e. Null values).

When further analysing the data one could see how most null rows are simply missed measurement trials. The next bigger contribution to null values could be found in the self-reported questionnaire data feature set, as the DASH questionnaire data was removed from the measurement protocol from trial C onward, and the other questionnaires were only proposed when injury was already present.

The rest of the missing values in the data can now be filled with imputing methods, further covered in the next section.

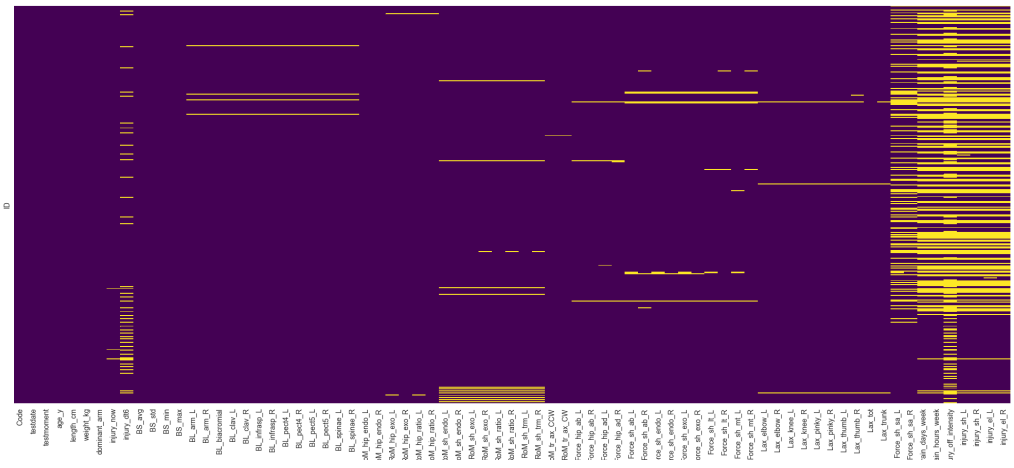

Figure 11: Nulls in dataset, after removal of empty rows and the questionnaire data. The rows (names) and columns respectively are the Y and X axis. Purple blocks show valid data, whereas yellow blocks show missing data (i.e. Null values).

A.3.2 Imputing and Encoding

Imputing is the process of filling up missing data. Imputing was done using a linear fashion for the length and weight parameters wherever they were missing, as these are seen to have an approximately linear progression over time in youth athletes. Other variables were imputed with the mean value of each variable using the `SimpleImputer`, to make sure not to change the variables' distribution.

Encoding is the process of assigning a unique number to each category in categorical variables in order for the algorithm to make sense out of the various categories. The impute and preprocessing modules of the SciKit-learn package [30] were used in this step. This is a ready-to-use package for both data pre-processing (encoding, imputing and scaling) as well as predicting. The encoders used are the `LabelEncoder` and the `OneHotEncoder`. The difference between the two is that the former is used for labelling multi-category within the same feature, whereas the latter encodes a separate binary feature (0 or 1) for each single category.

A.3.3 Making the Cross-sectional and Longitudinal Datasets

The data used for the analyses consists of two datasets, depending on whether it is used for cross-sectional or longitudinal analysis. The cross-sectional dataset simply consists of all data with the exception for the first trials (trials A and B), because they do not contain measurements of the target features.

The longitudinal dataset is made such that the model is fitted with predictor features of the present trial n and the target features of trial $n + 1$. As the target features are present from trial C on, predictors of trial B were therefore included. The predictors of trial F however, are excluded for the same reason. It is essential to state that this process had to be performed on each athlete's data individually and that this data was later combined again into a single dataframe for computation.

A.3.4 Train/Test Dataset Split

Splitting the dataset was done using the `train_test_split` function from SciKit-learn [30]. This function takes on the entire dataset and the testing size (0 to 1: 0% to 100% of data used for testing), and then splits the data completely randomly. When choosing the portion of data allocated for testing,

one has to take into account two main factors. It is favorable to have a testing size that is as small as possible, such that training is done on more data, which improves the accuracy of the model. One has also to make sure that the testing set is representative of the dataset in such a way that the categories that are to be predicted can be subjected to accuracy measurements. As a practical example, if no cases of injury are reported in the test set the model can not be valuated in accuracy.

Having this in mind, Figure 12 shows how the Logistic Regression algorithm performed by showing the F1 Score on the Y axis and splits from 0.1 to 0.5 on the X axis. Because of the small amount of data the way the data was split could make for results slightly changing each time, therefore 100 repetitions were done and consequently plotted. The goal was to identify a testing size in which no scores of 0 were achieved, and to use the same testing size on both the shoulder and the elbow joints. The final decision was thus made to use a test size higher than 0.3 and smaller than 0.4 (i.e. 30 % to 40 % of the total).

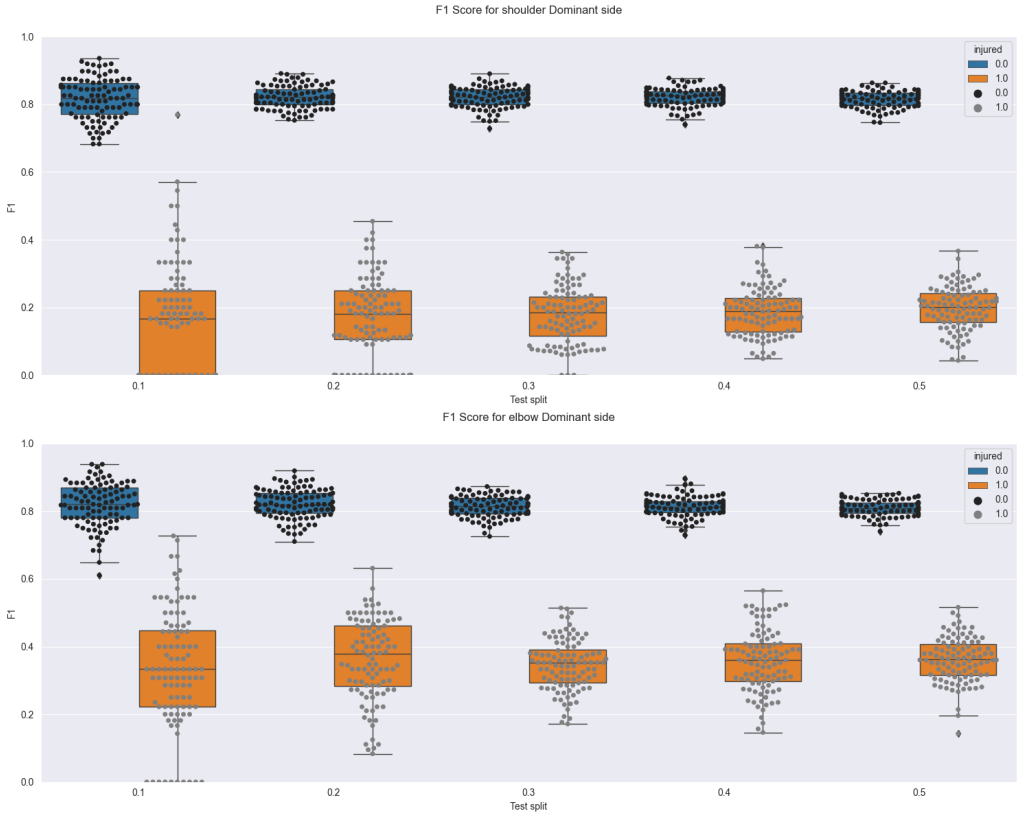

Figure 12: F1 Score for various amounts of testing size and 100 repetitions for each.

A.3.5 Scaler Selection

Scaling is a process done to the data to make sure the model does take inputs which are normally distributed. This is especially important when using algorithms which calculate the euclidean distance between parameters, as bigger parameters will then have a bigger weight. Scaling can be done with a variety of scalers. A short description of the scalers taken into account from the SciKit-learn package [30] follows:

- `MinMaxScaler`: for small standard deviations and non-gaussian distributions, MinMax is rather sensitive to outliers in the dataset. It scales the feature to a value in the range of 0 to 1.

- `StandardScaler`: as the name suggests it is a widely used scaler to distribute the data with mean of zero and SD of one.
- `MaxAbsScaler`: scales each feature individually to a maximum absolute of one. i.e. in a range between -1 and 1.
- `RobustScaler`: the name already indicates it, this scaler is very robust to outliers and keeps linear relationships when scaling.
- `QuantileTransformer`: scales values according to their quantile distribution. Although very robust, often linear relationships are lost with this scaler.
- `PowerTransformer`: used for modelling issues related to varying variability of features.

Even though there is quite some information about the way these scalers work, it is still very much recommended to make the scaling a Trial and Error process as real-world data is not always perfect. The F1 Score of a Logistic Regression algorithm for longitudinal predictions of injury with the different scalers can be found in Figure 13. Here one can see how the `MinMaxScaler`, the `MaxAbsScaler` and the `QuantileTransformer` perform worst. From the rest of the scalers eventually the choice for usage was made by looking at the best mean F1 Score, making the `StandardScaler` the scaler used for all other results in this paper, for both the shoulder and the elbow.

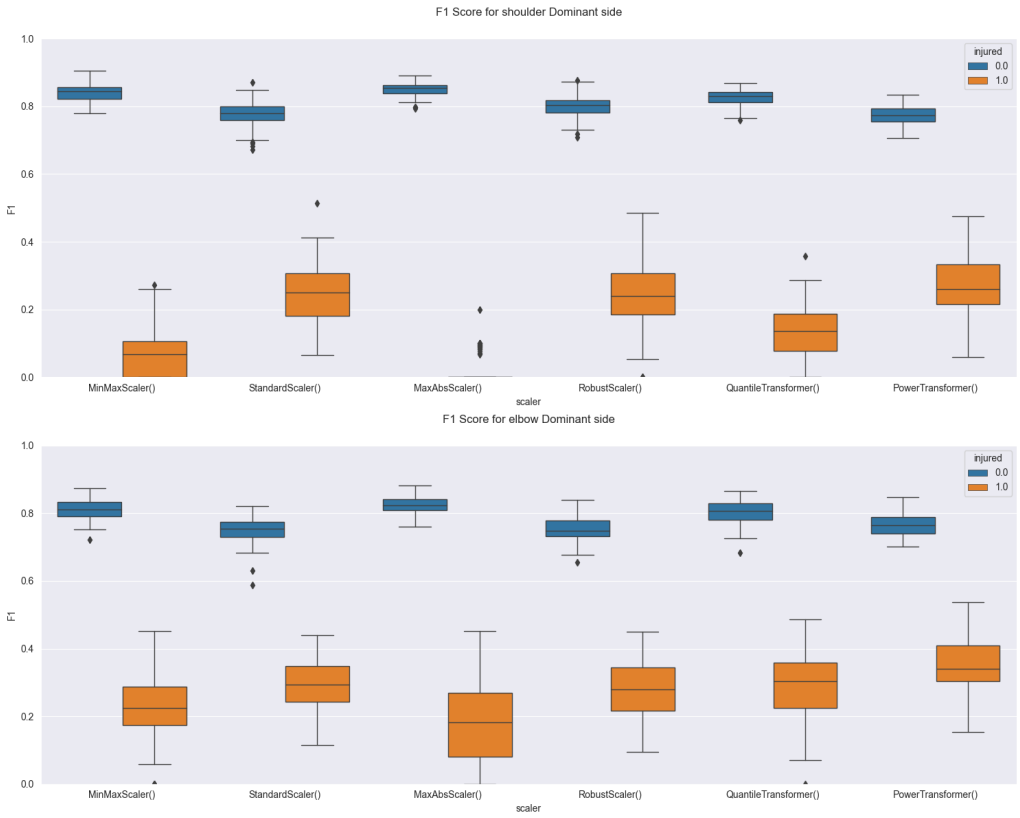

Figure 13: Raw F1 Score for various scalers applied on the data before being used to predict longitudinal injuries with a Logistic Regression algorithm.

A.4 Classification Performance Metrics

In classification problems, the concept of a Confusion Matrix (also known as an Error Matrix) is used to describe the performance of a model. Figure 14 shows such a matrix. The number of correct and incorrect predictions made by the model is summarised into it, and broken down into classes which show how *confused* the model is when making predictions. As shown in the example of Figure 14 the classes are a mix of the actual values and the prediction values both able to take a 1 ("is positive" observation) or a 0 ("is negative" or "is not positive" observation) value. Classes correctly predicted are called either True Positives (TP, prediction=1, actual=1) or True Negatives (TN, prediction=0, actual=0), whereas falsely predicted values are given different error classes. This way, we can distinguish between False Positives (FP, prediction=1, actual=0), also called Type I error, and False Negatives (FN, prediction=0, actual=1), also called Type II error [48].

Using a combination of these 4 classes (TP, TN, FP, FN), various metrics can be used to assess model performance. Because of the imbalance between positive and negative injury counts, the decision was made to use the F-β Score as a score of an algorithm's prediction accuracy. Below first the Precision and Recall metrics are explained as introduction to the F-β Score.

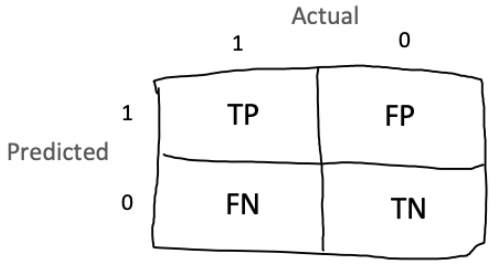

Figure 14: A Confusion (or Error) Matrix with classes of True Positives, True Negatives, False Positives (Type I error) and False Negatives (Type I error). Figure taken from [48].

Precision (P) and Recall (R): Precision is the value of how many of the items predicted positive are actually positive; Recall is the value of how many positive items are predicted correctly, and is also referred to as Sensitivity. So, if we want to minimise FNs recall has to get as close to 100% as possible, whereas to minimise FPs precision should go to 100% [48]. Precision and Recall are calculated with the formulas of equations 1 and 2 respectively. Taking a look at Figure 14 when examining the formulas helps the reader with understanding.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall/Sensitivity} = \frac{TP}{TP + FN} \tag{2}$$

F-β Score: is an accuracy measure that takes into account the balance between P and R, and calculates the harmonic mean of the two. It, thus, takes into account imbalance between classes and is therefore used when data is heavily skewed towards one class. This score attains a value between 0 and 1, where 1 stands for a perfect prediction score. The three most common values of β are 0.5, 1 and 2. When β is equal to 1 the score (called an F1 Score) balances the weight of the P and the R and is used to harmonically diminish FP (Type I error) and FN (Type II error) simultaneously. For a β value of 0.5 the score puts more weight on P and less on R. The opposite is true for a $\beta = 2$, which means R is favored. Using this score one has to consider that it does not take into account the TNs. [48].

$$\text{F-}\beta\text{ Score} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \tag{3}$$