Smart Health (2021)

Contents lists available at ScienceDirect

Smart Health

journal homepage: www.elsevier.com/locate/smhl

A Pain Free Nociceptor: Predicting Football Injuries with Machine Learning

Andrew Lyubovsky[1], Zhenming Liu[1], Amanda Watson[1], Scott Kuehn[1], Erik Korem[1], Gang Zhou[1]

ARTICLE INFO

Keywords:
Football
Injuries
Machine Learning
Sensor Data
Ubiquitous computing

ABSTRACT

Injuries are a significant aspect of every sport, with the ability to impact a player's career and the success of a team in their season. As sensor data is able to pick up on a player's physical state, recently it has been analyzed for its ability to predict player injuries. We inspect the predictive power of player stats, subjective player responses, GPS data, and training load data in forecasting game injuries from an NCAA American football team during the 2019 season. Data processing techniques are used to remove noise and decrease correlated data, and as large portions of the data is missing, multiple methods of data interpolation are tested. Survey data and player stats have the most predictive power for injuries with GPS and training load data performing at near-random levels. Overall, when modeling player stats and survey data together, injury predictions had a precision of .47, recall of .74, and an F1 score of .52 significantly outperforming random guessing.

1. Introduction

Football is a popular American sport where two teams attempt to move a ball across the field to score points. In the process, players need to sprint, pivot, tackle, block, and pass the ball to each other. Within each team, players specialize in the role that they play, with each position having a different job on the field. These include throwing the ball, tackling the opponent with the ball, or preventing opposing team members from moving the ball down the field. These movements are typically carried out in rapid bursts called plays, during which injuries can occur due to the force exerted on joints and muscles [42]. Typically, injuries are classified as contact injuries that result from player contact, non-contact injuries in movements like pivoting or accelerations, and chronic injuries that are a result of joint overuse.

In recent years, football's image has been hit hard due to the risk of injuries associated with the sport, especially the number and severity of concussions. While the National Football League (NFL) attempted to create rule changes to make the game safer and less injury-prone, their success has been limited [35]. Beyond affecting the safety of the sport, injuries also affect the short-term performance of players and the players' teams. Some injuries also have a long-term impact on a player's health, and repeated concussions often result in mental health consequences later in life. From the policy-making perspective, the two major ways of reducing injuries in football are through new technology and rule changes [35]. Athletic trainers and coaches also aim to minimize injuries by managing the physical preparation of athletes through their training regiment. In this study, we will examine how machine learning can further reduce the risks of injury associated with football by investigating its potential to foresee injuries before they occur.

While previous research on injury prediction in football is limited, GPS data has been shown to be successful in forecasting injuries in soccer [33]. That being said, American football and soccer are two very different sports, and the physical exertion on players is also very different, as there is much less contact between players in soccer. In Australian football, subjective player data collected through surveys was able to forecast player injuries one week out [22]. While there is contact in Australian football, the play is continuous in Australian football, and the positions are much less rigid compared to American football. Thus, while some information might be useful for injury detection in one sport, it will not necessarily be as useful in other sports. This raises the question: What contextual and collected data can be used to predict injuries in football?

To address this question, we collected player statistics, GPS data, subjective player responses, training load data about practice intensity, and omegawave data about physical recovery. There are other vital sign tests, including PPG or blood pressure testing. However, while these can demonstrate an individual's health condition they might not be as closely linked with athletic injuries. On the other hand omegawave data and the other chosen metrics are commonly linked with athletic performance and have been observed to be indicative of athletic injuries [28, 33]. The data

http://dx.doi.org/10.1016/j.smhl.xxxx.xx.xxx

was collected on a Division 1 College football team over the course of the 2019 season, which consisted of 11 games spanning four months. We collected data on a total of 101 players, where 31 of the top players that saw more game time collected more data.

As with other real-world studies, a significant portion of data was missing in our study, affecting the effectiveness of the machine learning algorithms. Some of the missing data was a result of not having enough sensors to collect on all of the players, while other missing data resulted from not collecting on off days or from athletes simply forgetting to collect data. Typically, the issue of missing data is addressed by relying on contextual information; however, as there was no best way of filling in missing values for our given context, we evaluated multiple methods of data interpolation on each data stream.

Lastly, using the interpolated data, we analyzed the predictive power that GPS data, training load data, survey data, and player statistics have in forecasting game injuries. This was done by fitting a Logistic Regression on each collected feature and evaluating the statistical significance of the slope of that Logistic Regression. Next, we examined if the features that are statistically significant in the training data retained their predictive power in the test set in order to evaluate the predictive power of the selected features.

This study's main contributions are the following:

- We collected a data set by recording player stats, GPS data, survey responses, training load data, and Omegawave recordings that preceded injuries in a NCAA Division 1 American football team over the course of one season.
- We examined the effectiveness of five methods of data interpolation in order to address the common issue of missing data. We found that Matrix Completion methods had outperformed other data interpolation methods in most cases.
- We analyzed the predictive power that GPS data, survey responses, training load data, and player statistics have in predicting American football injuries. We found that survey data and player stats had some ability to predict injuries, while none of the GPS or training load features had statistically significant predictive power.

The rest of the paper is structured as follows. We start by describing previous work done with sport injury predictions in Section 2. Then, the collected data is described in Section 3. Next, as large portions of the collected data is missing, multiple interpolation methods are tested to achieve the best injury prediction model. This is done in Section 4. In Section 5, we remove highly correlated features and create new features that provide novel information, such as how the features change from one week to the next. After that, in Section 6, we select features that have predictive power for game injuries. In Section 7, we model the data using different machine learning techniques to predict game injuries on the top players. Lastly, we wrap up with our discussion and conclusion.

2. Related Works

Machine learning has been applied to sports science to tackle a variety of issues. These include performance prediction and injury risk analysis, amongst other issues such as team management and match attendance [6]. These studies have been carried at both the collegiate and professional levels [6] across different sports, including track and field [29], soccer [33], and Australian football [33]. While some previous studies focused on overall injuries, many studies focused on specific injuries, such as knee injuries [30] or heart malfunction [1]. In extension of these works, we consider the problem of predicting overall injuries in American football. As the rest of our paper does not focus on one specific injury, this section does not discuss work done with predicting specific injuries.

Neural Networks have been used to assess the physical readiness of athletes by forecasting athletic performance at competitions. Peterson [29] used Recurrent Neural Networks to predict sprinting performance based on heart rate variance data collected on an NCAA track and field athlete over the course of 32 competitions. He was able to predict race times with an above random accuracy when testing on 6 held out competitions [29]. While this does not predict injuries and is limited to one athlete, the omegawave data had predictive power for an individual's performance, reflecting individual physiological preparation that can be linked with injuries.

In addition to athletic performance, Peterson [28] also researched sports injuries. He monitored 21 female student-athletes on three undisclosed teams to predict injuries based on accelerometer data, heart rate variance data, and self-reported data. He used Bayesian Networks to look at key injury predictors that can be detected from the collected data. Peterson conducted three rounds of 5 fold cross-validation and was able to outperform the accuracy of naive classification [28]. Similar to Peterson [28], we collect GPS data, training load data, survey data, and omegawave data; however, we look at injury prediction for games in order to address weekly patterns in the data that precede competitions since competitions result in a higher number of injuries.

McCullagh and Whitfort [22] used artificial neural networks (ANNs) to estimate when players will get injured in Australian football. Each week, they analyzed thirty attributes that included physical preparation tests and subjective player responses to fit an artificial neural network with 15 hidden nodes to predict injuries in the following week. They used 10 fold cross-validation and were able to train the ANN with an accuracy of 82% and an F1 score of .59 [22]. Similar to this study, we use subjective player data to predict injuries on a weekly basis. However, we also compare it with GPS data and more detailed training load data when predicting American football injuries. We evaluate our system leaving one game out for testing as opposed to cross-validation in order to better simulate injury predictions for future games.

Lastly, Rossi et al. [33] predicted injuries in soccer based on GPS data using Decision Trees, Random Forests, and a Logistic Regression. They conducted this study on 26 participants with a total of 23 non-contact injuries over the course of 23 weeks and had a total of 931 player points (data-points for a player on any given day). They used thirty percent of their data to select features and tune parameters. Next, then they split the remaining data into two equal folds, training and evaluating each model on each of the folds. This procedure was repeated 1000 times, changing the way that they split their data. In the end, they were able to predict game and practice injuries with an F1 score of .6 [33]. Compared with Rossi et al., we evaluated survey and training load data in addition to GPS data, and we predict injuries during games that the model was not trained on, making the results more rigorous.

Overall, multiple sources of data have shown to have forecasting power to predict injuries across a variety of sports. Physical preparation tests and player surveys had predictive power in predicting injuries in Australian football [22], and GPS features were shown to be effective at predicting injuries in soccer [33]. Omegawave heart rate variance data was also shown to be effective in forecasting athletic performance [29]. In this paper, we predict football injuries, and, as sports differ in the way they are played, it is not necessarily the case that the predictive power of one data source carries over across sports. Thus, we evaluate GPS data, training load data, and survey data for their predictive power in American football. By testing our results on games that are left out of the training set, the results show predictive power for games that have not been observed.

3. Data Collection

The data was collected on an NCAA Division 1 American football team (101 players) over the course of 113 days in their 2019 season. During the season, there were a total of 173 recorded injuries, where injuries are defined as any complications that occur with players that result in partial or full removal from practices, or an inability to participate in games. From Figure 1, we observe that injuries were more common during games, as over half (89) of them occurred during games that made up only 10% of all days. Injuries can be categorized into contact injuries, non-contact injuries, and overuse injuries, yet for the success of a team, players need to practice together, and the absence of any one player hurts the team's performance. Thus, when modeling injuries we focus on all injuries regardless of their type in an effort to enhance the team's performance. To determine ways of foreseeing these injuries, we recorded five different streams of data that include player stats, training load data, survey, GPS, and omegawave data. Player stats, training load, and surveys are collected on all players as they do not require special sensors. However, due to costs of omegawave [25] sensors and Catapult's Playertek [16] GPS sensors, we were not able to collect GPS data and omegawave data on all of the athletes, so we prioritized data collection on some players over others. For example, starters and the players with the most game time were given GPS monitors because these players are most important to the team's success and typically sustain the largest amounts of injuries. Omegawave sensors were also given to players that were considered to be most critical for the team's success and the ones that would be the most consistent with data collection as these players would have the largest impact on the team's performance and the sensors required larger efforts from the players for data collection. Overall, GPS data was collected on 31 athletes, and omegawave data was collected on five players. The study was IRB approved, and all of the players gave their written consent to participate. Due to privacy concerns related to personal data, all of the GPS data is stored in a password protected location on the IBM Cloud, and the rest of the data is stored in a private database on the Athletic Department's Google Drive that has permissions granted only to select individuals within the organization. In the remainder of this section, we describe how the data was collected and what type of information is valuable from each source of data, how it was collected, and when it is missing. We will start by discussing the weekly practice schedule in order to put the patterns observed in the data in context.

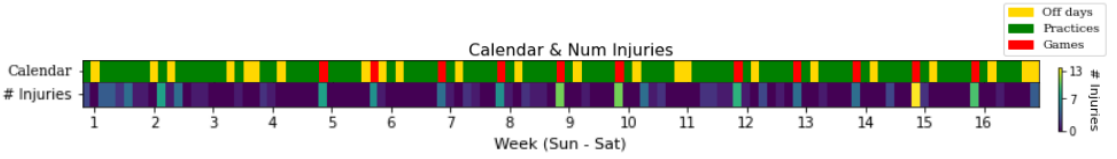

Fig. 1. *Injuries and Schedule*

Schedule. During the season, the football team typically had games on Saturdays. Sundays were lighter practices, and Mondays were off days. Tuesdays and Wednesdays were harder practices with some scrimmages, and Thursdays and Fridays were lighter days before games on Saturday. There was one game that took place on Friday; however, all other games happened on Saturdays. The practice schedule can be seen in Figure 1.

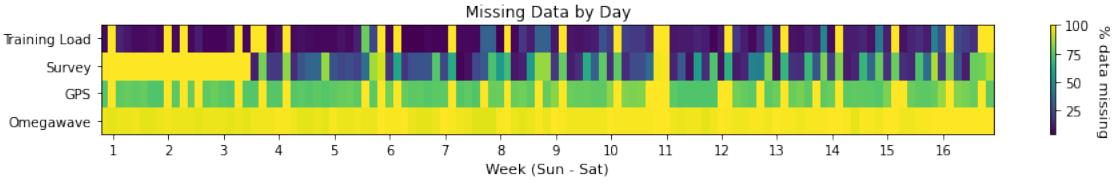

Fig. 2. *This figure presents when data was missing for each source of data. Each data row is broken down into 113 days, and the the percentage of players that collected that sensor is indicated by the color of a box. Dark colors indicate that most players collected a certain sensor on a given day, and light levels indicate that the sensor was not collected on most players on a given day.*

Player Stats. Player stats are measurements that reflect basic information about a player and the type of exercise that they do on a given day. Unlike the other data, while these statistics might suggest how likely a given player will get injured on a given day, they do not reflect information about a player's health. They include the a player's position, class standing, and the number of injuries that they have had during the season. While

Table 1. Collected Player Stats

Feature	Description
Date	Date of data-point
Weekday	Day of the week (Mon-Sun)
Player Id	Id associated with player
Position	The position of the player
Class	The year in school
Starter	If a player is a starter or not
Skill Group	Players were separated into three skill groups
Number of Injuries	Number of Injuries that a player has had
Game	Whether or not it is game day
Game Participation	Whether or not a player played in the game
Practice	Whether or not there is practice
Conditioning	Whether or not there is conditioning

more player stats are included in Table 1, only the ones mentioned here are relevant for game injuries. The other statistics such as Day of Week or Game, Practice, and Conditioning indicate player activity and provide useful information about training sessions in general that are not limited to games. These values are collected by coaches and staff rather than the players so none of the data is missing.

Table 2. Collected Training Load Data

Duration	Duration of training session
RPE	Rate of perceived exercise (1-10)
Global Load	Overall training load measure (RPE $\times$ Duration)

Training Load. Training load data reflects information about practice intensity. It is composed of practice and game duration and the rate of perceived exercise (RPE). The rate of perceived exercise is a subjective measure of exercise intensity on a scale from one to ten provided by each player at the end of each practice. Global load is calculated by multiplying RPE by the practice duration. These metrics are commonly collected by teams to keep track of practice intensity [23], and are displayed in Table 2. As RPE and Duration were recorded by the players at the end of practice, training load data is present for 93% of the players during games and practice days. However, no training load data was collected on off days. Thus it can be observed that days with 100% of the training load data missing in Figure 2 correspond with off days in Figure 1.

Table 3. Collected Survey Data

Sleep Quality	Rate the quality of your sleep (deep, restful, satisfying) [1-10]
Hours of Sleep	How many hours did you sleep? [1-10]
Recovered	How physically strong and recovered does your body feel? [1-10]
Mood	How is your mood? [1-10]
Energy	How is your energy (alertness, focus)? [1-10]
Soreness	Do you have any muscle or joint soreness, stiffness, or tightness? [1-10]
Wellness Quotient	Overall subjective measure of player health [0-100]

Survey Data. Survey responses record subjective data about the quality of a player's recovery and how they feel. These include information about sleep quality on a scale from one to ten, the number of hours that a player sleeps, the feeling of how recovered they are, their mood, energy, and soreness. These are commonly collected values adopted from the Hooper-Mackinnon Wellness questionnaire [13, 40]. All of these are on a scale from one to ten and are presented in Table 3. Lastly, this data is used to compute a quotient reflecting a players wellness on a scale from one to 100. It is calculated by finding the z score of a recorded survey value compared to that value's readings for a player in the two previous weeks. These z scores are then scaled to 1 with the following mapping: {>1.5:1, >.9:.9, >0:.75, <1.5:.6}. They are then averaged amongst all of the values for mood, energy, and the other features from a given day, and then multiplied by 100 to find the Wellness Quotient. While methods to calculate an overall quotient vary, it is common for it to be calculated [40]. To collect survey data, surveys were delivered to the players each day as a google survey. As players started recording their responses one week before the first game, the first few weeks are missing. Since the players were also able to complete the surveys at home, they were collected more evenly throughout the week compared with other data sources; however, they still fluctuated throughout the week. Surveys were most frequently collected Tuesdays through Fridays and on Sundays. Thus, an average of 39% of

player surveys are missing from those days, and an average of 75% of the surveys are missing on Saturdays and Mondays. Only the top players were asked to collect surveys on Saturdays as those were game days, and Mondays were typically days off. This corresponds with Figure 2, where Saturdays and Mondays tend to have 70 % of the data missing.

Table 4. Collected GPS Data

Duration	Duration of training session
Distance	Distance a player ran in a training session
Sprint Distance	Distance a player ran at a speed greater than 5 m/s
Power Plays	# of actions in which power exceeds 20 watt/kg of player weight
Energy burned	kcal burned during training session
# Impacts	Detects impacts greater than 5g
# Accelerations	Number of times a player accelerates
# Decelerations	Number of times a player decelerates
# Sprints	Number of times a player sprints
Top Speed	Top speed of a player
Distance Per Min	Average speed of player
Power Score	Power output per kg of player weight
Work Ratio	Percent of time that player was moving faster than 1.5 m/s
Player Load	Sum of all accelerations across training session
Player Load Per Min	Time averaged player load
Distance & Time in Speed Zone (1-5)	Distance that a player covers when running at different speeds
# of impacts by level of impact (1-5)	Number of impacts that a player has at different levels of impact
# Power plays by duration (1-5)	Number of powerplays separated into groups based on time
Distance, Time & Number in Acceleration Zones (1-5)	Time, distance, and number of accelerations in zones based on acceleration intensity
Distance, Time & Number in Deceleration Zones (1-5)	Time, distance, and number of decelerations in zones based on deceleration intensity
Distance & Time in Power Zones (1-11)	Time and distance in zones grouped by the power that a player exerted

GPS Data. While survey data reflects recovery, GPS data reflects practice intensity by recording player motion during practices. It records information about the duration over which it was recording, the total distance that a player traveled, and the distance that a player sprinted, amongst other factors as shown in Table 4. Player motion was also partitioned into different zones, including speed zones, acceleration zones, deceleration zones, heart rate zones, and power zones, recording the information about the distance and duration of each of those zones [16]. While the other zones were separated based on one standard threshold, the threshold for speed zones are broken down into the percentages of the maximum speed averaged over all of the players that play the same position. This is similar to what was done by O'Connor et al. [26]. The specific values of each zone are presented in Table 5, and all of the values that the GPS recorded are presented in Table 4. GPS data was collected using the PlayerTek System [16] composed of a GPS, a tri-axial accelerometer, and a tri-axial magnetometer that is placed on a player's back during practices and games [16, 34]. As we only had access to 28 GPS units, GPS data was consistently collected on 31 players, and 70 percent of the data is missing on practice and game days. From Figure 2, it can also be observed that there is no data collected on rest days, and there are conditioning days when GPS data was not collected. Lastly, for 6 of the 11 games, GPS data was collected on the Fridays prior to the games.

Omegawave. Omegawave data measures physiological information about the central nervous system and the cardiovascular system. This sensor was used to measure DC potential and heart rate variance data in order to reflect athletic readiness. For each of the recordings, a proprietary algorithm is used to produce a number from 1 to 7 to reflect readiness [12]. These values include direct current potential, a measure associated with consciousness in decision making or alertness [39]. Other measures associated with the ability to cope with stress or reaction time are calculated on a scale from one to seven [39]. Omegawave was collected by players when they woke up by placing an ECG belt on their chest, one electrode on the forehead, and another on the thumb. As we had a limited number of omegawave devices, we collected data on those players that were most critical to the team's success and the ones who were most willing to collect the data. Omegawave data was collected on five people, and at the beginning of the season four players regularly collected it. However, due to the challenges in collecting the data and compliance limitations, as the season progressed, athletes got less consistent with the data collection. By the end of the season, none of the players collected it consistently, and 98% of the data collected by omegawave is missing. Thus, it is not investigated further.

4. Data Interpolation

When collecting the data, parts of it were missing, which is a common problem in machine learning. We had a limited number of sensors so some players did not have GPS data collected on them, and we only collected GPS and training load data during practices, so these data sources were missing on off-days. Lastly, some missing data was caused by players simply forgetting to collect data. As the features for each data stream were collected together, each stream has all of the features either missing or present for any given data-point. Thus, we are able to interpolate each source of data individually. For each source of data, we add it to other data that is previously filled, and interpolate the missing values using

Table 5. Zones recorded by GPS

Name	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
Speed Zone (% MAX)	0 - 20 %	20 - 50 %	50 - 75 %	75 - 90 %	> 90 %
Power Play Duration Zones	2.5 - 5 s	5 - 7.5 s	7.5 - 10 s	> 10 s	-
Accel & Decel Zones	0 - 1 m/s/s	1 - 2 m/s/s	2 - 3 m/s/s	3 - 4 m/s/s	> 4 m/s/s
Power Zones (1-5)	0 - 5 w/kg	5 - 10 w/kg	10 - 15 w/kg	15 - 20 w/kg	20 - 25 w/kg
Zones (6-10)	25 - 30 w/kg	30 - 35 w/kg	35 - 40 w/kg	40 - 45 w/kg	45 - 50 w/kg
Zone 11	> 50 w/kg	-	-	-	-

Table 6. Collected Omegawave Data

Cardiac Readiness	1-7 measure reflecting how awake or relaxed a player is
DC Potential	1-7 measure of .5 Hz brainwave reflecting how ready the brain is for stress
Adaptation Reserve	1-7 measure reflecting how long cardiac system can respond to training
Stress	1-7 measure reflecting the tension in response to load
CNS Readiness	comprehensive measure of central nervous system readiness (1-7)

five different interpolation methods. This way previously filled data sources can provide information for the data interpolation. Since player stats has no missing values, we interpolate survey data, adding it to player stats. Then we include training load data, and lastly we add and interpolate GPS data. After each data interpolation, we evaluate the performance of the different interpolation methods by using the interpolations to predict injuries one day out. In this section, we will first discuss common approaches to handling missing data. Then, we will describe how we interpolate the missing data and discuss our results. The filled in data is then used in Section 5.

4.1. Techniques for Addressing Missing Data

There are multiple methods of addressing the issue of missing data. When missing data is encountered, one method of interpolation is typically chosen based on prior knowledge. Imputed values are then impacted by prior assumptions. However, in some cases, it is not clear which method of data interpolation should be used [32]. For example, when GPS data was missing on Mondays, the average value could be used to fill missing data as that is what we would expect to see when everything is normal. On the other hand, missing values could be imputed with zeros as there are no practices on Mondays, and values of zero would reflect a more realistic expectation of exercise levels [44]. Values of negative one could also be used as a way to indicate that the data is missing. Rashid et al. [32] tested three different types of data interpolation methods to see which performs best, including Matrix Completion based methods, Interpolation based methods, and Regression based methods [32]. Matrix Completion based methods are ones where data is interpolated to fit the larger dimensions in the existing data. Interpolation based methods are those where data is interpolated based on surrounding data-points, and regression based methods form models which then predict missing values [32]. As there was no definite method of interpolation that fits our data, we follow a similar approach and test five models of data interpolation.

For mean interpolation, missing data is filled with the average value for a given parameter. For previous interpolation, missing values are interpolated with the previous day's value. K-Nearest Neighbors interpolation fills values with the average of the three nearest points using the 2 norm (Euclidean distance) to calculate distance [38]. Next, a Matrix Completion is used by applying the Soft Impute algorithm. Soft Impute minimizes the Frobenius norm of the new matrix and the Nuclear Norm of the matrix that is imputed [21]. Lastly, we test multiple imputations by chained equations algorithm (MICE) using ten iterations. This method iteratively models all of the feature's values by creating a regression from the other features, filling the missing values of a modeled feature with the result [3].

4.2. Data Interpolation

As we tested multiple methods of data interpolation, we evaluated those imputations by modeling injuries one day out. That is, we use the interpolated data collected each day to predict who will sustain an injury on the following day. More so, as different sensor types can be interpolated using different methods, for each sensor, the sensor's data is interpolated using each method. Then, the data from the interpolation that performs best is added that to previously interpolated data. We start with player statistics, then added survey, training load, and then GPS data in that order.

The interpolations are tested by fitting the data using five models: a Logistic Regression, Support Vector Machine (SVM), Decision Tree, k-Nearest Neighbors algorithm, and a Gaussian Naive Bayes. Using all of the players and all of the days, we use 5-fold cross-validation to test the results. For the Logistic Regression, no regularization is included, and the injury days are weighted heavier than non-injury days in order to address a disproportionate number of non-injury data-points [7, 19]. A cubic SVM with a radial basis function kernel is used without regularization, and with weighted classes [24]. This is a simple SVM for modeling unbalanced data. For the Decision Tree, the "gini" criteria is used to identify the best split of the data, splitting the data until the data-points are fully separated [31]. Again, these are standard hyper-parameters for a Decision Tree. For the k-Nearest Neighbors algorithm, the nearest two points are used, with each point being weighted by the distance to the missing value [17]. Lastly, we use a standard Gaussian Naive Bayes algorithm with prior probabilities adjusted to the injury distribution [15].

Table 7. Best Performing Imputation Results (F1 scores)

Data & Interpolation	Logistic Regression	SVM	k-Nearest Neighbor	Decision Tree	Gaussian Naive Bayes
Player Stats	.06	.2	.02	.09	.03
Player Stats + Survey (MC*)	.1	.16	.04	.18	.03
Player Stats + TL (MC*)	.16	.20	.02	.10	.03
Player Stats + TL + GPS (MC*)	.17	.23	.09	.17	.03

MC* – Matrix Completion TL – Training Load

Upon creating the models of the data, we use the F1 score calculated on the test set to evaluate each model's ability to predict injuries one day out. The F1 score is used as it penalizes models that identify non-injuries as injuries as well as penalizing models that do not detect injuries at all [10]. To calculate the F1 score, the True Positive (TP), False Positive (FP), and False Negative (FN) rates are first computed. True Positives are the days where the model correctly predicts that a player got injured. False positives are the days where a player did not get injured that the model classifies incorrectly as injured days. Lastly, False Negatives are days where a player got injured, yet the model fails to predict an injury. Next, the precision and recall are calculated. Precision and recall can be expressed as P and R with formulas that are expressed below. Precision reflects the number correctly labeled injuries out of all of the days that are labeled as injuries. Similarly, recall reflects the number of correctly labeled injuries out of the total number of injuries that were observed. Lastly, the F1 score is a combination of the two scores with the following equation [10]:

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN} \qquad F1 = \frac{2(P \times R)}{P + R}$$

In order to create a baseline, we first use player stats from one day to predict injuries on the following day for any recorded day. Next, we interpolate survey, training load, and GPS data in that order. The best results after each data stream is added to the previous data are presented in Table 7. Modeling player stats individually, the largest F1 score is .20 when an SVM is fit. Then, survey data is interpolated using the five different methods when combining it with player stats. Matrix Completion performs best, producing the largest F1 score of .16 when using a SVM model. Since the performance is reduced when survey data is included, survey data is left out when interpolating future sensors. Next, we add training load data to the player stats. Matrix Completion interpolation outperforms all other interpolation methods with an F1 score of .2 for Support Vector Machines and an F1 score of .16 when fitting a Logistic Regression. As including training load data improves overall performance compared to player statistics, it is included in the interpolated dataset. GPS data is then interpolated using multiple methods and added to training load and player stats. For GPS data interpolation, Matrix Completion again outperforms other interpolation methods when testing using all models other than KNN. When fitting using the KNN algorithm, the k-Nearest Neighbors interpolation method outperforms matrix completion interpolation. The Logistic Regression improves from an F1 score of .16 to .17 when the interpolation is added, and the Support Vector Machine improves from an F1 score of .20 to .23 when Matrix Completion interpolation is applied to GPS data. As including GPS data improves the F1 score of our prediction, we keep GPS data that is interpolated using Matrix Completion.

Although there is a large amount of data missing, and there is no one best way to interpolate it, we address this issue by testing multiple methods of data interpolation. We find that using Matrix Completion interpolation and an SVM model produces the best result for all data streams. However, for some models, other methods of data interpolation slightly outperform Matrix Completion. This suggests that while Matrix Completion interpolation does not produce the best results for each type of model, in general Matrix Completion produces the best result for most models. While we filled in missing data for predicting injuries one day out, in the next section, we will further process this data in order to model game injuries with better effectiveness.

5. Data Processing

Once the missing data is filled in, it needs to be processed to extract new information and to decrease the correlation between existing data while keeping its intrinsic meaning. We accomplish this by examining how features changed from the previous week and by creating normalized features for each player. Then, we remove features that are highly correlated with other features to address repetitiveness within data. Lastly, we average features prior to games in order to decrease the amount of noise within the data.

Looking at how data changes from one week to the next (differencing) can offer novel information [2]. Hence, We first extract new information by calculating the weekly change in all of the features. The change from each previous day is not calculated as weekly schedules differ day by day, and calculating the change in features from the previous day would add noise. On the other hand, since practice schedules follow a weekly pattern, GPS and training load data is expected to remain similar for each given day of the week. Similarly, survey data is strongly impacted by weekly class and practice schedule, so the change from the previous week is calculated rather than a change from the previous day.

After calculating the new features, we have 178 GPS features from each day, many of which are highly correlated. Specifically, the GPS features that are collected by zones: within each group of zones, the average Pearson correlation coefficient is .61, and the average correlation coefficient between adjacent zones is .73. To reduce the number of repetitive features, Principle Component Analysis (PCA) is applied to each group of zones individually. For example, when PCA is applied to the five speed zone features, five new features are extracted that are uncorrelated. The features extracted from the PCA that account for a variance greater than .001 are then left. This way, all of the features that are left have intrinsic meaning about a specific zoned data while significantly reducing the number of features from each day [43]. Next, to reduce the dimensionality of the features while maintaining the interpretability, when a pair of features has a Pearson correlation coefficient with an absolute value greater than .75, one of the features is removed [4]. This way, the 178 GPS features that were collected each day from GPS data are reduced to 38 features.

As training load has only three features collected each day, none of the training load features are removed. As none of the survey features have a correlation coefficient with an absolute value greater than .75, all of them were also kept.

Lastly, the exercise intensity can vary from player to player, and what would be a difficult work load for one player might be typical for a different player. This would be reflected in different ranges of GPS features amongst different players. Similarly, as survey data is subjective, the range of values that one player records can differ from the range of values that a different player records, and one extreme recording for one player might be normal for a different player. To address this, new features were included in the dataset where the data was normalized amongst each player's data. That is, for each feature, the mean and standard deviation are calculated amongst a player's data. Then, the mean is subtracted from each data-point, and the difference is divided by the standard deviation. By doing so, the days that are more extreme for a player can be identified more easily [36]. In the end, we are left with the original features, the original features normalized by player, features about how how the data changes from one week to the next, and those features normalized.

Once all of the features are processed, we specify our predictions to game injuries. This increases the predictive power that the features have and reduces possible outside noise caused by factors such as variability in practice schedule and player participation. First, we chose to look specifically at game injuries as 52% of the injuries occur during games. By limiting our predictions to games, we prevent the model from simply predicting injuries on game days and outperforming random guessing. More so, multiple days leading up to games have a consistent schedule, so the noise due to the differences in the schedule is reduced and patterns can be picked up more easily. Next, 70% of the 89 injuries were sustained by the top 30 players who consistently play in the games and whose data collection was most consistent. Thus, in order to prevent the model from classifying the top players as those that have a higher risk of injury compared with other players, only the top players are used when modeling game injuries. Lastly, to minimize the effect of missing data and data interpolation, for each sensor we selected players who had collected a sensor's data four or more times during the preceding week. This way, a large portion of the game data-points are kept, and the data interpolation preceding the observed games is minimal. In the end, GPS data had 233 player game points with 51 injuries, there were 316 training load player game points with 62 injuries, and 225 player game points for survey data, with 49 injuries.

Lastly, features multiple days prior to a game could possibly be used to forecast injuries. That being said, individual days have lots of noise and could affect the predictive power of a feature. Thus, we average across multiple days during the week to remove noise.The average of the three days prior to a game and a weekly average is calculated in order to remove noise. The three-day average before games reflects information that is most relevant to a game as it is most recent, and the data averaged over a week aims to reduce the amount of noise that is picked up. Similarly, the average between 2 and 4 days (inclusive) before games is calculated for each feature to average the data collected on the more strenuous practices where the training load could cause fatigue that would carry over into a game. In the end, in order to predict game injuries, we use the features averaged over the past three days, the features averaged over the whole week, and the features averaged over the three harder practice days, tripling the number of features. After all of the features are calculated, there are 446 GPS features, 50 training load features, and 177 survey features prior to each game.

6. Feature Selection

As we have a large number of features compared to the number of data-points, we would like to select those features that have predictive power for injuries. Having features without predictive power can lead to overfitting [11]. In other words, since we have many features, the noise within some of those features might fit injury occurrences well. We would like to separate those features where injury prediction is caused by noise from those that have predictive power. To do this, we first split the dataset into 11 training and testing sets (train/test sets), leaving out a different game for each test set. This way, it is most similar to predicting injuries in future games that have not been observed yet. For each feature in the data set, we calculate the p-values associated with the slope of a Logistic Regression fit on the training set [37]. Then, we evaluate how the Logistic Regression performs on the test set by calculating the area under the curve for the receiver operating characteristic curve (AUC). The AUC measures the ability of a model to predict injuries. It does so by reflecting the model's ability to distinguish game data-points with injuries from game data-points without injuries when given a data-point's features [14].

From Figure 3, we can observe the predictive power for injuries that the three types of data have. For each type of data, we would like to find a threshold for p-values in the training set where the features below that threshold have AUCs that outperform random guessing by a statistically significant amount. On average, survey data outperforms random guessing, where the average AUC across all of the features and train/test sets is above .5. Similarly, we can see that at low p-values, the features outperform random guessing. On the other hand, when the p-value of a feature is large in the training set, this indicates that a feature does not have much predictive power. When selecting features that have a p-value smaller than a .05 threshold, those features have a statistically significant slope associated with the Logistic Regression for the training set. They are also able to forecast injuries in the test set above levels of random guessing by a statistically significant amount. This suggests that p-values can be used to access predictive power for survey features. However, this .05 threshold does not carry over to GPS and training load data. When low p-values are selected for GPS data, they perform worse than average as the average AUC is below .5 for p-values less than .1. More so, as the average AUC is equal to .5, which is the expected AUC score for randomly selected features, GPS data's predictive power is limited. Lastly, training load features have little predictive power when selecting features with p-values below a .4 threshold. From Figure 3, we can see that selecting any lower threshold coresponds to an AUC that outperforms random guessing, yet when comparing AUC values of the features to AUC values of random guessing, the test AUC's p-value (blue line in Figure 3) is above .2 for any threshold below .4. However, the average AUC value for the features is slightly over .5, and the p-value of all of the AUCs is less than .05 suggesting some predictive power across all of the training load data. Thus, while statistical significance in the training set does not carry over to the test set, Figure 3 suggests that some training load features do have slight predictive power.

While there are 177 features in the survey data, if the features are fully random, there is a 37% chance of observing a p-value of .005 (1/177). However, as there exists correlation within our data from averaging days, the probability of seeing a p-value of 1/177 significantly decreases. More

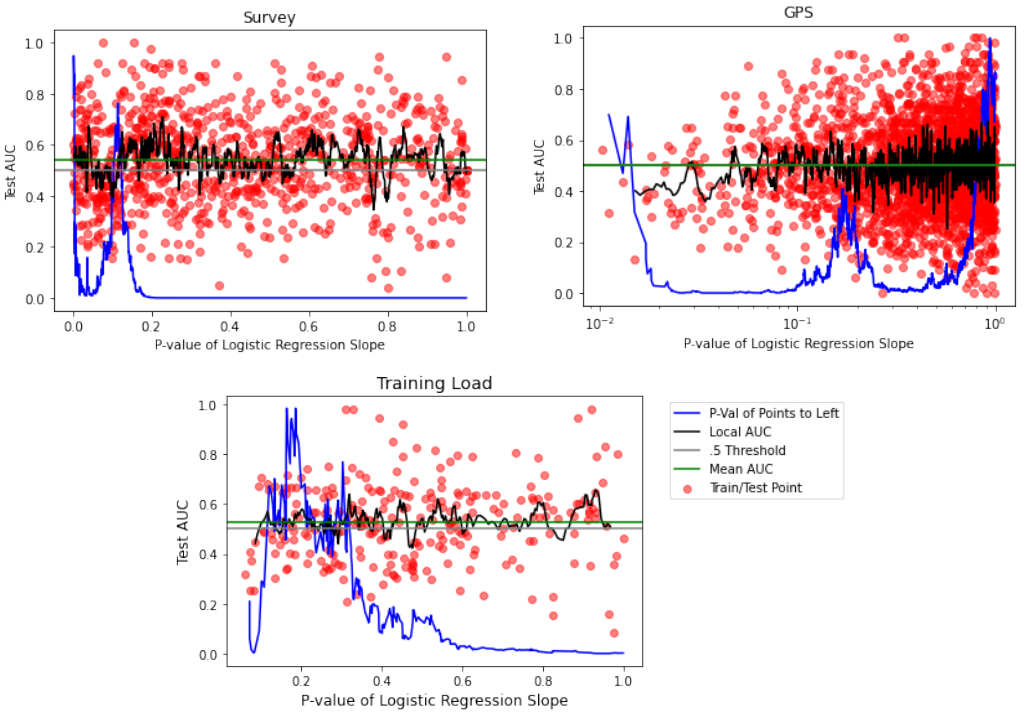

Fig. 3. This figure displays the p statistic that a Logistic Regression's slope has for a training set feature. It also displays the the predictive power that the corresponding feature has in the test set. This relationship will be later used to select useful features. The three different data streams are observed separately. Within each plot, each red point represents a feature in one of the train/test sets. The x value of a train/test point indicates the p-value associated with the slope of the training set Logistic Regression. The y value indicates the test AUC of the Logistic Regression that is fit on a feature in the training set. The black line measures the local test AUC average over the nearest 20 points. The blue line measures the p-value associated with selecting a specific threshold. That is, it reflects the probability that all of the red points to the left of a certain threshold appear at random with a mean of .5. Thus, when the blue line is low, it indicates that the test AUC of the red points to the left of a given x value is not due to chance. Lastly, the horizontal grey line is at .5, which is the expected value of a random test AUC sample, and the green horizontal line is the mean AUC of the test AUCs. That is, the green line is the mean y value of the red points and is where a feature's AUC would be expected to appear if it was selected at random from the sensor's features.

Table 8. Survey p-values

p-value	Feature
0.0036	Player Normalized weekly change in soreness averaged over 3 days prior to game
0.0097	Weekly change in soreness averaged over 3 days prior to game
0.0163	Weekly change in wellness quotient averaged over week
0.0169	Weekly change in wellness quotient averaged over 3 harder practice days
0.0274	Player normalized weekly change in wellness quotient averaged over week
0.0365	Soreness value averaged over last 3 days
0.0380	Player normalized weekly change in wellness quotient averaged over four harder days

so, we observe that the features with p-values below .05 outperform random guessing in the test set. The associated test AUC values with the selected features have less than a .02 percent likelihood of appearing at random, validating the use of a .05 p-value threshold for feature selection.

6.1. Selected Features

We see that survey features with low p-values in the training set have predictive power in the test set. Thus, by examining survey features with p-values smaller than .05, we are able to identify the survey information with the most predictive power. While some GPS features had equally low p-values, when features that had p-values below a .05 threshold were selected, these features performed worse than guessing injuries at random. This suggests that the low p-values in the GPS data are caused by noise. Since GPS data has 446 features, and Survey data has 177 features, p-values below .05 are more likely to be caused by chance in GPS data. More so, while survey features have predictive power for injuries, each individual feature is not examined for a threshold above which an injury is likely. This is because individual features do not have enough predictive power for injuries and would require larger models to be useful. Thus, a threshold for an individual feature would not provide useful information.

After calculating the p-values for survey features over all of the data, the features where the slope of the Logistic Regression have a p-value less than .05 are presented in Table 8. Player soreness in the days leading up to games has predictive power as features that rely on this value

have multiple p-values below .05. While the change in soreness from the previous week that is normalized by each player and averaged over three days has the most predictive power, soreness values averaged over the three days also have statistically significant predictive power. This suggests that normalizing the data and finding the change in features from the previous week improves predictive power in some features. The change in the Wellness Quotient from the previous week also has statistically significant predictive power when averaged over the week and over the three harder practice days. As the Wellness Quotient is a collection of multiple factors, its predictive power suggests that all of the features can provide some useful information, yet there isn't enough data to observe statistical significance in each feature individually. Overall, player soreness and the Wellness Quotient have the most predictive power out of all of the collected data. In the next section, we will evaluate how well all of the features can be combined to model injury prediction.

7. Modeling

In the previous section, we selected the most useful features by using a Logistic Regression to model each feature individually. In this section, multiple, more complex models predict game injuries based on each type of sensor data. Then survey data and player stats model game injuries as both data streams have predictive power for injuries when they are modeled individually.

Using the features selected on each train/test set, a baseline model is created by fitting a Logistic Regression to each feature. These Logistic Regressions vote on each game data-point, and the average of their predictions is used as the final prediction [8]. Next, a variety of models with different hyper-parameters are tested. These include 11 Logistic Regressions, 26 SVMs, 20 KNN models, 21 Decision Tree classifiers, and 8 Random Forest classifiers. Each time, we use 11 train/test sets where a different game is left out for each testing set.

Logistic Regression models are tested using different values of C, an inverse regularization parameter. At small values of C, model complexity is punished, and models are affected less by the noise in the data. On the other hand, when values of C are large, more complex models are formed that can fit more complex patterns related to injuries. We test Logistic Regression models with C values ranging from .01 to 2 and always use the Broyden–Fletcher–Goldfarb–Shanno (LBFGS) algorithm for optimization [18]. The SVM models use different kernels, including linear, polynomial, sigmoid, and radial basis function kernels. Three models with third degree radial basis function kernels are tested with C values of .3, 1, and 3. Five models with linear kernels are tested with C values ranging from .1 to 10. Ten models with third degree polynomial functions are tested with C values ranging from .01 to 100. Lastly, 8 SVM models with sigmoid kernels are tested with C values ranging from .01 to 30. Next, seven KNN models are tested with the number of neighbors ranging from 2 to 90. Ball tree, kd tree, and brute algorithms are tested on KNNs that use both five and eight nearest neighbors. Lastly, seven KNNs using five nearest neighbors evaluate a range of leaf sizes, ranging from five to 150. For Decision Trees, six models are tested where the maximum depth ranges from two to ten. These models use "gini" impurity to evaluate the quality of a split, a measure of well samples in a tree node are split into different classes [31]. Four models are tested by varying the number of samples that are required to split a leaf node, ranging from two to ten. Next, four models test different minimums for the number of samples required in a leaf. These range from three to ten. Next, six models test these ranges where the maximum depth of the tree is limited to four nodes. One model uses entropy as a measure for the split quality. Lastly, the eight Random Forest classifiers are tested with the maximum depth ranging from 2 to 50 [27]. For each model, the training and testing AUC, precision, recall, and F1 score are calculated. We test a range of parameters for each model as one variation of a model might fit the data better than others. In the end, changing parameters did not lead to a significant change in model performance. Thus, it is not discussed in depth.

7.1. Player Stats Model

First, we create a model using player statistics to predict game injuries on top players. This data contains a total of 62 injuries in a total of 316 game data-points. Using a random model, we would expect the AUC to be .5, and if the F1 score is maximised by predicting all game data-points as injured, the F1 score would be .32. However, if (62/316) games are predicted as injured at random, the F1 score would be .19. When fitting a Logistic Regression on each feature, and having each Logistic Regression vote on whether an injury occurs or not, the resulting test AUC and F1 are .63 and .40 respectively. This result suggests that a baseline model outperforms random guessing. When full models are used, on average a Logistic Regression performs best with an average Test AUC of .79, as seen in Figure 4. This suggests that using a full Logistic regression improves the performance over a baseline voting model. Similarly, when observing the best performing Logistic Regression, the test AUC is .8, and the test F1 score is .45. This model has an parameter C of .5. While decreasing the value of C results in a slight improvement in the Test AUC, changing the value of C does not have a strong effect on the performance of the Logistic Regression.

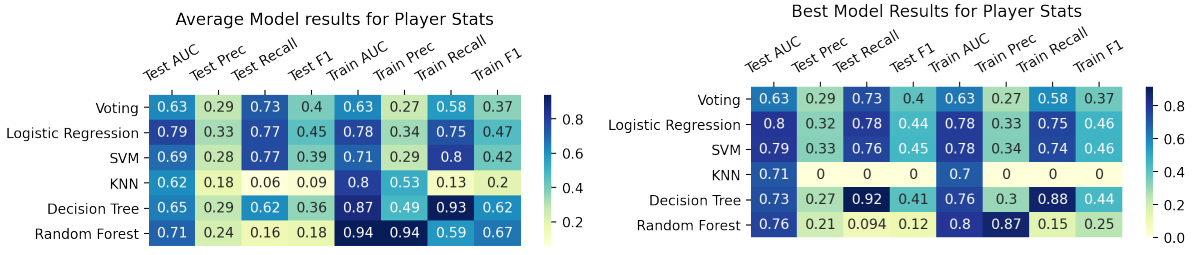

Fig. 4. Modeling using player statistics. The chart on the left presents results that are averaged by each model type. The chart on the right presents results with the largest Test AUC for each model type. Each row presents results from a specific model type, and each column presents a different metric for a given model collected on either the training or testing data.

7.2. Survey Model

Modeling the survey data separately, only the game data-points from the top players are included, and if a player did not complete four or more surveys in the preceding week, the data-point is removed. The resulting data consists of 49 injuries over the course of 225 game data-points. Using a random model, we would expect the AUC to be .5, and if the F1 score is maximized by predicting that all game data-points will have an injury, the F1 score would be .36. However, if (49/225) random game data-points are predicted as resulting in an injury, the F1 score would be .21. When multiple Logistic Regressions vote on whether an injury occurs or not, the resulting test AUC and F1 are .59 and .35 respectively. As the resulting test AUC outperforms a random AUC, the result suggests that Survey data has prediction power and establishes a baseline for other models. Modeling survey data using full models, on average a Logistic Regression performs best, and is able to outperform the baseline voting model as seen in Figure 5. Similarly, the Logistic Regression model achieved the largest Test F1 score, with a regularization parameter C of 2. As the value of C had minimal effect on the Logistic Regression, it is unlikely that the this model was over fitting the data. While the best SVM improved the test AUC, the improvement is minimal and the test F1 score decreased compared with the Logistic Regression. While a Random Forest model had a higher Test AUC, the test F1 score also dropped significantly compared with the Logistic Regression. Reasons why this can occur are discussed in Section 7.7.

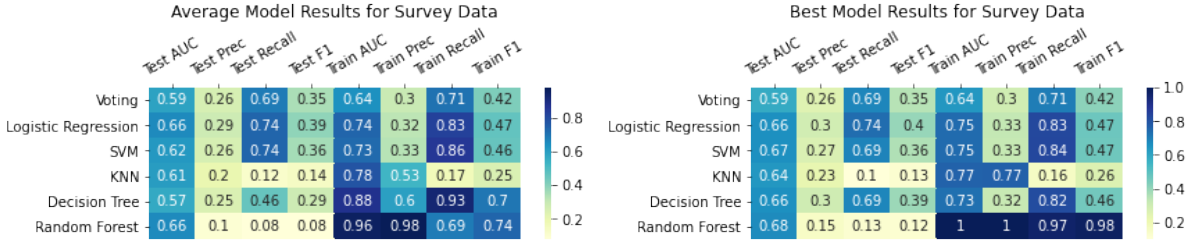

Fig. 5. Modeling using survey data. The chart on the left presents results that are averaged by each model type. The chart on the right presents results with the largest Test AUC for each model type. Each row presents results from a specific model type, and each column presents a different metric for a given model collected on either the training or testing data.

7.3. GPS Model

Next, we model the GPS data separately, and the only data-points that are included are those from the top players where GPS data is collected on four or more days in the week prior. This consists of 44 injuries over the course of 191 game data-points. Using a random model, we would expect the AUC to be .5, and if we maximize the F1 score by predicting all data-points as injured, the F1 score would be 0.37. However, if (44/191) of the players' games are randomly predicted to result in an injury, the F1 score would be 0.23. Fitting a Logistic Regression on each feature, with each of them voting whether or not an injury occurs, the resulting test AUC and F1 are .47 and .28 respectively, performing worse than at random. Similar to the baseline model, the average test AUC across each model type also perform worse than random. This cooresponds with a limited prediction power in GPS data that was seen in Section 6. While there is a a Support Vector Machine model that outperforms random guessing with a Test AUC of .55, it is likely a result of overfitting as other SVMs perform at near random level. Similarly, while on average the test F1 score for SVM models is greater than that of predicting 44/191 as data-points as injured at random, the increase in the F1 score is minimal.

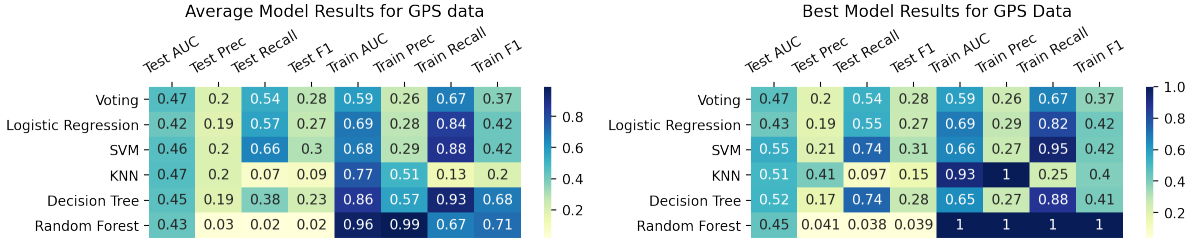

Fig. 6. Modeling using GPS data. The chart on the left presents results that are averaged by each model type. The chart on the right presents results with the largest Test AUC for each model type. Each row presents results from a specific model type, and each column presents a different metric for a given model collected on either the training or testing data.

7.4. Training Load Model

When we select the game data-points on the top players with five or more days of collected training load data, there are 62 injuries in 316 player game data-points. Using a random model, we would again expect the AUC to be .5, and if the F1 score is maximized by predicting all data game points as injured, the F1 score would be 0.32. If we predict (62/316) games as injured at random, the F1 score would be 0.19. As there are only 36 features, nine of the test cases have no features with a p-value less than .05. Thus, we present results on the two games that had such parameters. When using the results to fit a Logistic Regression and vote on game data-points, the model has an AUC of .51 and an F1 score of .18, which is similar to random guessing. The mean AUCs for Logistic Regressions, SVMs, and Decision Trees all outperform random AUCs

slightly as can be seen from Figure 7. Moreso, a test AUC of .68 is achieved when using when using a Decision Tree with a maximum deapth of 10 nodes. That being said, the testing AUC and F1 are larger than the training AUC and F1 for the best performing SVM and KNN models, suggesting over-fitting is occurring for the best performing models. More so, as there are only two train/test splits in the training load data such that a parameter with a p-value less than .5 existed, we observe the predictive power of training load data is very limited in our data-set.

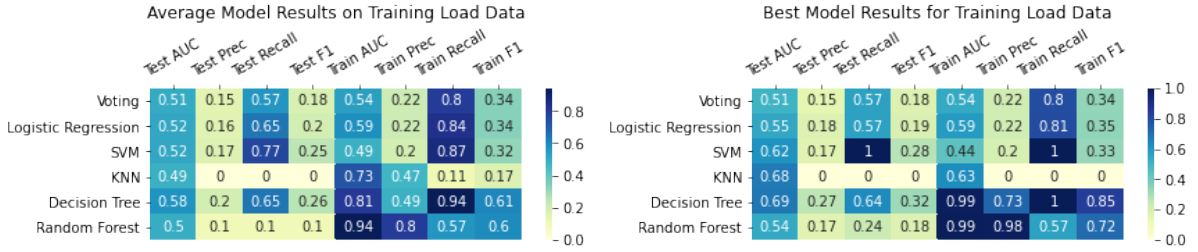

Fig. 7. Modeling using training load. The chart on the left presents results that are averaged by each model type. The chart on the right presents results with the largest Test AUC for each model type. Each row presents results from a specific model type, and each column presents a different metric for a given model collected on either the training or testing data.

7.5. Survey & Player Stats Model

Survey and player stats are both able to model the data separately, performing at above random levels. On the other hand, when training load and GPS data are used, only a few features outperform random guessing. For the dataset, we only use the data-points when survey data is collected four or more times on a given week. This is the same data set as was used for modeling survey data individually. As such, there are 225 game data-points, 49 of which are injuries. A maximized F1 score would be 0.36, and .21 if (40/173) games are predicted to be injuries at random. When multiple models vote on injuries, the test AUC is .67, and a test F1 score is .41. This results in improved baseline results compared to individual models for player stats and survey data. Modeling using a Logistic Regression results in the largest average test AUC with an average AUC of .81. When fitting a Logistic Regression with a C value of 3, the test AUC is .82 and the test F1 is .52, outperforming all other models and predicting injuries with above random accuracy. This is the best model that is achieved on the dataset, and can be seen in Figure 8.

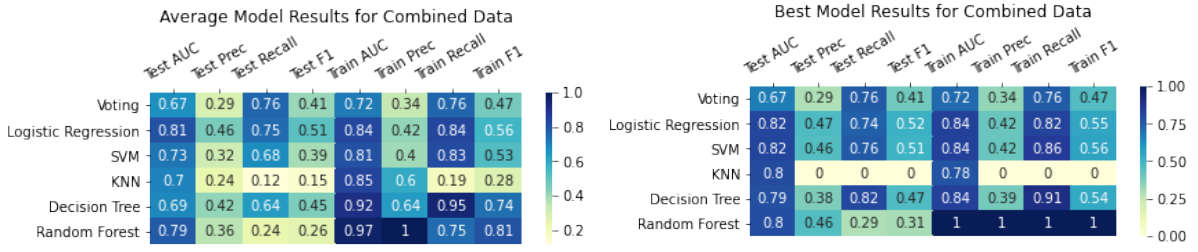

Fig. 8. Modeling using player statistics and survey data. The chart on the left presents results that are averaged by each model type. The chart on the right presents results with the largest Test AUC for each model type. Each row presents results from a specific model type, and each column presents a different metric for a given model collected on either the training or testing data.

7.6. LSTM Model

In the previous sections, we used more basic models to evaluate the performance of Logistic Regressions, Support Vector Machines, K-Nearest Neighbors, Decision Trees, and Random Forests when forecasting game injuries one day out. However, recently more complex models such as Long-Short Term Memory networks (LSTMs) have been successful in modeling temporal data [20]. These are recurrent neural networks that address the problems of exploding gradients and gradient decay that most traditional recurrent neural networks encounter [20]. In this section, we test the ability of LSTMs to model game injuries on the top players as there may be patterns in the data that an LSTM is able to detect while more traditional approaches are not.

In order to model our data using an LSTM and keep it consistent with the other testing results, we limit our dataset to game injuries that occur on the top players. This way, the LSTM cannot outperform random guessing by detecting the players that are most likely to be playing on a certain day or by detecting which days are game days. While this reduces the number of data points that are present in the dataset, it allows us to compare the LSTMs' results with previous results and retains a large portion of injuries that are present in the initial dataset. Secondly, as combining player stats with survey data provides us with the best results in Section 7.5, this data is used again for training and testing the LSTMs. As done previously, survey data is interpolated with Matrix Completion as this yields the best results in Section 4.2. Next, we will describe the LSTM models that we test.

For the LSTM models, we test multiple different hyperparameters to evaluate which combination of hyperparameters produces the best result on 5 fold cross-validation. Again, we give more weight to data points that contain injuries compared to data points that do not in order to prevent

the models from predicting all days as non-injured. As there are a total of 292 data points, we want to limit the number of trainable parameters that a model contains, so we test architectures that include 2, 4, and 8 hidden LSTM nodes. Similarly, a fully connected layer is included after the LSTM layer, where we also test a range of 2, 4, and 8 nodes for the same reason. For both layers, a softplus activation function is used, and for the output layer, a sigmoid activation function is used to produce an output value that ranges between zero and one. An Adam optimizer is used to optimize a binary cross-entropy function as it is a commonly used method for optimizing binary data [9]. Next, a learning rate of .001 is used iterating over 1000 epochs. These values are chosen as they successfully converge in testing. Next, to test a range of regularization techniques, we include a dropout layer after the LSTM layer and after the fully connected layer, testing dropout rates of 0, .1, and .4. And lastly, we add equal l1 and l2 regularization to the optimizer, testing regularization values of 0, .01, .03, and .1 as these help prevent overfitting on the training data [9]. The resulting test AUCs averaged over five folds within the data are presented in Figure 9 for each LSTM configuration.

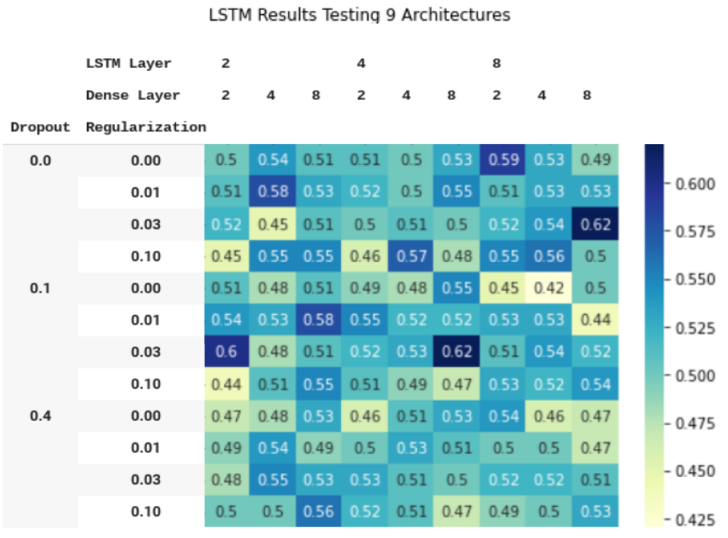

Fig. 9. Modeling multiple LSTM configurations on survey and player stats. This model presents the average AUC result that each LSTM configuration produces when modeling game injuries on the top players. The average AUC is calculated amongst each of the five folds in 5 fold cross-validation. The model architectures are presented along the X-axis, and the regularization parameters that are tested are displayed on the Y-axis. All possible combinations of LSTM configurations are presented in the table.

From Figure 9, we can observe that the LSTM is able to outperform random guessing as the average AUC value for all of the LSTM models is .514. While multiple models have little to no predictive power and result in AUC values smaller than .5, an average AUC value that is greater than .5 suggests that some LSTM models do have predictive power, even if it is limited. There are two models that perform best, resulting in test AUC values of .62. One of these models has 8 fully connected nodes, 4 hidden nodes in the LSTM layer, and used .03 regularization for the l1 and l2 parameters while having a dropout rate of .1. The other model that performs best has 8 nodes in the fully connected layer, 8 hidden nodes in the LSTM layer, and it uses a regularization of .3 for the l1 and l2 parameters. Both of these models outperform other LSTM models and have an AUC of .62 as can be seen in Figure 9. This suggests that while an LSTM is able to pick up on some patterns and predict injuries based on survey data and player stats, more basic models such as a Logistic Regression outperform the LSTM in predicting injuries. This occurs because the data set is too small to fit a more complex model, and thus, the LSTM misses the more basic patterns that a Logistic Regression is able to pick up on. Overall, the LSTMs' results suggest that while these more complex models are able to pick up on some patterns within the data, simpler models outperform LSTMs due to the limited size of the dataset.

7.7. Modeling Analysis

In this section, independent stats, survey data, GPS data, and training load data are used to model game injuries on the top players. In Section 6, survey data had features with statistically significant prediction power, and in this section, multiple models were able to model game injuries when using player stats. Thus, player stats and survey data were combined to model game injuries. We observe the average model results in each model and the best results. The average results of each model type reflect how well the data is able to be modeled by any machine learning algorithm. This allows us to compare the model with random guessing, where if the data fails to be modeled, then the average result of the models will be similar to that of random. On the other hand, the best results from any given model type reflect how well any model is able to fit the data. That being said, it is subject to overfitting as the best model is selected from multiple models. This provides an estimate for the best performance that can be achieved by using a specific data type.

Overall, survey data and and player stats are able to model game injuries on the top players while training load and GPS data are not able to significantly outperform random guessing. First, player statistics are able to predict game injuries with above random accuracy as all of the model types have a test AUC greater than .6. Specifically, the Logistic Regression has the best model performance with a test AUC of .8. However, we can observe that the best performing Logistic Regression does not significantly improve performance over the other logistic regressions. Similar to player statistics, survey data is also able to fit game injuries well. All of the model averages outperform random guessing, and the worst model average AUC is .57. While the test AUC is above average for all of the models, the test F1 score is worse than random guessing for KNNs and

Random Forests. This suggests that while the data is able to be split, the threshold used to classify a data point as injured is too large, leading to a low recall and a low F1 score. Similar patterns are observed in the GPS, training load, and combined models. For survey data, all of the best performing models have a similar test AUC, with an average AUC of .66. This corresponds with survey data having features with predictive power in Section 6. When modeling GPS data, none of the model averages outperform random guessing when observing the mean AUC. This cooresponds with Section 6, where the GPS features had limited predictive power. While there is an SVM that is able to model the game injuries, this is likely due to chance. Lastly, there are only two train/test sets where training load features have low p-values in the training set. Thus, there are only two train/test sets for the training load data. Out of all of the models, only Decision Trees consistently outperform random guessing. Similarly, within each model type there are models that are able to outperform random guessing, yet this is likely a result of the small train/test sample. As a result, neither GPS or training load data is included in the combined model. When combining player stats and survey data, a Logistic Regression has the largest average AUC of .81, which is .02 greater that that when modeling player stats individually. When observing the best performing models, the test AUC again improves by .02 for the best Logistic Regression model. On the other hand, the SVM, KNN, and the Decision Tree models improve slightly more when survey data is included with player stats. This suggests that survey data stores useful information that is not in player stats as it improves the predictive power of a model when it is used along with player stats.

8. Discussion

In this section, we will first summarise our findings related to sport injury prediction and compare them to previous findings. We explain why some sensor data that was useful in other work was not useful for us. As a few GPS features had predictive power while the others did not, we discuss their significance. Next, we discuss our findings with regard to interpolating missing data, comparing it with related works. Lastly, we discuss the limitations that existed within our study and suggest direction for future work.

In this study, we analyzed the predictive power that GPS, survey, and training load data have in forecasting football injuries. Survey data about player health and player statistics have predictive power, where the features that had the most predictive power were the number of previous injuries, how sore a player feels in the days leading up to games, and a general quotient summarizing how recovered a person feels. While we modeled only the top thirty players on the team, the results that we observed passed rigorous testing (i.e., statistically significant) and had shown to significantly outperform random guessing on those thirty players. Specifically, we show that when features have predictive power in the training set, then they also have predictive power outside of the training set, allowing us to extrapolate the conclusions outside of our data set. While some GPS features and training load features have predictive power, the overall data has limited predictive power for injuries. More data or novel approaches are necessary in order to more accurately identify the usefulness of GPS and training load data.

Our survey data results correspond with McCullagh & Whitfort's [22] results where they used subjective player data to quantify injury risks in Australian football one week out [22]. On the other hand, GPS data was useful in predicting soccer injuries, while it has limited predictive power in football. One possible explanation for this is that differences in the physical demands of the two sports cause variation in what data is useful. In soccer, the gameplay is more continuous compared with rapid bursts in American football, which affects the features that are picked up by the GPS [5, 41]. More so, in football, the physical demands vary more between players than in soccer, where a player's positions in football can cause a large difference in the running distance, amongst other player movement characteristics [41]. Lastly, the physiological stress exerted in soccer differs from that in football due to the physical contact in football. This causes a higher prevalence of contact injuries in football. Thus, the different injury types can effect the ability of GPS sensors to pick up on injuries. Since the usefulness of GPS data in forecasting injuries differs amongst sports, this is an area that could be investigated further.

It is the case that using all of the GPS features leads to results without any statistical significance. However as there are many GPS features, there may be a few with limited predictive power that are not selected due to noise. If only features derived from the impacts and the power zone durations are used, some predictive power is observed. When the same procedure for data processing and feature selection is used as described in section 5, a model reflecting votes from multiple Logistic Regressions produces an AUC of .56 and an F1 score of .37. This model's AUC outperforms a random AUC, and the F1 score is consistent with the maximum F1 score that a trivial model would perform at. More so, when this is modeled using a sigmoid kernel, with a C value of .3, the AUC is .61, and the average F1 is .33 for the testing data. While these observations are subject to overfitting, they do suggest predictive power in the GPS data that would require further inquiry. As more data is collected, future studies can inspect for similar effects.

While we investigated what sensors have the most contribution to injury prediction, our data also contained many missing values, so we investigated the effectiveness of various data imputation methods. Similar to Rashid et al. [32], we were not able to use domain knowledge to impute missing values as is commonly done in practice; thus, we tested five different methods [32]. After interpolating our data, we modeled the following day's injuries based on a current day's data using a Logistic Regression, a SVM, a Decision Tree, and a Gaussian Bayes model. We used 5 fold cross-validation and computed the F1 score to evaluate the performance. Matrix Completion based methods outperformed previous interpolation, mean interpolation, KNN interpolation, and MICE imputation methods on each of the sensor imputations that we did. This is consistent with Rashid et al.'s findings, where Matrix Completion methods also outperformed the other imputation methods [32].

One of the major limitations of this study was the amount of data that was collected. While there were 30 different players, they were all on the same team, and the study spanned only one season with a consistent practice schedule. As this study took place over the course of one season, we only analyzed data from 11 games, limiting the ability to test consistency between games. More so, as we only observed 11 games, the features that we examined may have had little predictive power in our data set due to unaccounted for variability, and more data would help us account for that variability. Lastly, as we only collected data over the course of a season, we limited our predictions to game injuries to minimize the amount of noise. As more data is collected, this noise may have less adverse effects, and training injuries could be further analyzed. The quality of the data is also impacted by the sensor, where newer sensors may be able to pick up less noise. Lastly, omegawave data was collected in limited amounts due to challenges associated with the data collection. As Peterson [29] had been able to use it in forecasting game play, and other studies had used it to

forecast injuries, it should be further inspected. Similarly, it's relationship with other sensor data can be analyzed to see if a relationship exists. As new vital sign tests and sensors are developed, this new data can have predictive power for football injuries, leading to better models for injuries. In addition, new data processing techniques may be identified that will have stronger predictive power. Lastly, similar studies can be applied across other sports, across more football teams, and over the course of a longer time frame. These may all provide new insight into methods of predicting injuries in sports.

9. Conclusion

In this study, we collected sensor data on a team of 101 NCAA Division 1 football players over the course of a season. As there were challenges involved in data collection, where large portions of the data was missing, multiple interpolation methods were tested and Matrix Completion had shown to be most effective. Next, we processed features from GPS, survey, and training load data in order to create a set of features prior to games that would be used to predict injuries. Survey data, especially the Wellness Quotient and soreness had the most predictive power. Player stats also had predictive power, especially the number of injuries that a player had previously sustained. While GPS and training load had limited predictive power, our data sample was limited to 11 games, and a more extensive study could prove otherwise. In the end, we were able to predict injuries with a precision of .46, recall of .75, and F1 score of .51.

References

[1] Adetiba, E., Iweanya, V. C., Popoola, S. I., Adetiba, J. N., & Menon, C. (2017). Automated detection of heart defects in athletes based on electrocardiography and artificial neural network. *Cogent Engineering*, *4*, 1411220.

[2] Anderson, C., Burt, P. J., & Van Der Wal, G. (1985). Change detection and tracking using pyramid transform techniques. In *Intelligent Robots and Computer Vision IV* (pp. 72–78). International Society for Optics and Photonics volume 579.

[3] Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, *20*, 40–49.

[4] Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing* (pp. 1–4). Springer.

[5] Bradley, P. S., Sheldon, W., Wooster, B., Olsen, P., Boanas, P., & Krustrup, P. (2009). High-intensity running in english fa premier league soccer matches. *Journal of sports sciences*, *27*, 159–168.

[6] Claudino, J. G., de Oliveira Capanema, D., de Souza, T. V., Serrão, J. C., Pereira, A. C. M., & Nassis, G. P. (2019). Current approaches to the use of artificial intelligence for injury risk assessment and performance prediction in team sports: a systematic review. *Sports medicine-open*, *5*, 1–12.

[7] Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, *35*, 352–359.

[8] Friedman, J., Hastie, T., Tibshirani, R. et al. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Annals of statistics*, *28*, 337–407.

[9] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

[10] Goutte, C., & Gaussier, E. (2005). A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *European conference on information retrieval* (pp. 345–359). Springer.

[11] Grus, J. (2019). *Data science from scratch: first principles with python*. O'Reilly Media.

[12] Heishman, A. D., Curtis, M. A., Saliba, E. N., Hornett, R. J., Malin, S. K., & Weltman, A. L. (2017). Comparing performance during morning vs. afternoon training sessions in intercollegiate basketball players. *Journal of strength and conditioning research*, *31*, 1557.

[13] Hooper, S. L., Mackinnon, L. T., Howard, A., Gordon, R. D., & Bachmann, A. W. (1995). Markers for monitoring overtraining and recovery. *Medicine & Science in Sports & Exercise*, .

[14] Huang, J., & Ling, C. X. (2005). Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, *17*, 299–310.

[15] John, G. H., & Langley, P. (2013). Estimating continuous distributions in bayesian classifiers. *arXiv preprint arXiv:1302.4964*, .

[16] Johnston, R. J., Watsford, M. L., Kelly, S. J., Pine, M. J., & Spurrs, R. W. (2014). Validity and interunit reliability of 10 hz and 15 hz gps units for assessing athlete movement demands. *The Journal of Strength & Conditioning Research*, *28*, 1649–1655.

[17] Keller, J. M., Gray, M. R., & Givens, J. A. (1985). A fuzzy k-nearest neighbor algorithm. *IEEE transactions on systems, man, and cybernetics*, (pp. 580–585).

[18] Kelley, C. T. (1999). *Iterative methods for optimization*. SIAM.

[19] King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political analysis*, *9*, 137–163.

[20] Lipton, Z. C., Kale, D. C., Elkan, C., & Wetzel, R. (2015). Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv:1511.03677*, .

[21] Mazumder, R., Hastie, T., & Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, *11*, 2287–2322.

[22] McCullagh, J., & Whitfort, T. (2013). An investigation into the application of artificial neural networks to the prediction of injuries in sport. *International Journal of Sport and Health Sciences*, *7*, 356–360.

[23] McLaren, S. J., Smith, A., Spears, I. R., & Weston, M. (2017). A detailed quantification of differential ratings of perceived exertion during team-sport training. *Journal of Science and Medicine in Sport*, *20*, 290–295.

[24] Noble, W. S. (2006). What is a support vector machine? *Nature biotechnology*, *24*, 1565–1567.

[25] Omegawave (). Omegawave ble sensor. URL: `https://www.omegawave.com`.

[26] O'Connor, F., Thornton, H. R., Ritchie, D., Anderson, J., Bull, L., Rigby, A., Leonard, Z., Stern, S., & Bartlett, J. D. (2020). Greater association of relative thresholds than absolute thresholds with noncontact lower-body injury in professional australian rules footballers: implications for sprint monitoring. *International journal of sports physiology and performance*, *15*, 204–212.

[27] Pal, M. (2005). Random forest classifier for remote sensing classification. *International journal of remote sensing*, *26*, 217–222.

[28] Peterson, K., & Evans, L. (2019). Decision support system for mitigating athletic injuries. *International Journal of Computer Science in Sport*, *18*, 45–63.

[29] Peterson, K. D. (2018). Recurrent neural network to forecast sprint performance. *Applied Artificial Intelligence*, *32*, 692–706.

[30] Qilin, S., Xiaomei, W., Xiaoling, F., Yuanping, C., & Shaoyong, W. (2016). Study on knee joint injury in college football training based on artificial neural network. *RISTI (Revista Iberica de Sistemas e Tecnologias de Informacao)*, (pp. 197–211).

[31] Raileanu, L. E., & Stoffel, K. (2004). Theoretical comparison between the gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence*, *41*, 77–93.

[32] Rashid, H., Mendu, S., Daniel, K. E., Beltzer, M. L., Teachman, B. A., Boukhechba, M., & Barnes, L. E. (2020). Predicting subjective measures of social anxiety from sparsely collected mobile sensor data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, *4*. URL: `https://doi.org/10.1145/3411823`. doi:`10.1145/3411823`.

[33] Rossi, A., Pappalardo, L., Cintia, P., Iaia, F. M., Fernández, J., & Medina, D. (2018). Effective injury forecasting in soccer with gps training data and machine learning. *PloS one*, *13*, e0201264.

[34] Rossi, A., Perri, E., Trecroci, A., Savino, M., Alberti, G., & Iaia, F. M. (2017). Gps data reflect players' internal load in soccer. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)* (pp. 890–893). IEEE.

[35] Sheth, S. B., Anandayuvaraj, D., & Patel, S. S. (2018). The impact of rule changes on the number and severity of injuries in the nfl. *bioRxiv*, (p. 503227).

[36] Singh, D., & Singh, B. (2020). Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, *97*, 105524.

[37] Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: A review. *Data classification: Algorithms and applications*, (p. 37).

[38] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics*, *17*, 520–525.

[39] de la Vega, R., Jiménez-Castuera, R., & Leyton-Román, M. (2020). Impact of weekly physical activity on stress response: An experimental study. *Frontiers in Psychology*, *11*.

[40] Vescovi, J. D., Klas, A., & Mandic, I. (2019). Investigating the relationships between load and recovery in women's field hockey–female athletes in motion (faim) study. *International Journal of Performance Analysis in Sport*, *19*, 672–682.

[41] Ward, P. A., Ramsden, S., Coutts, A. J., Hulton, A. T., & Drust, B. (2018). Positional differences in running and nonrunning activities during elite american football training. *The Journal of Strength & Conditioning Research*, *32*, 2072–2084.

[42] Wellman, A. D., Coad, S. C., Goulet, G. C., & McLellan, C. P. (2016). Quantification of competitive game demands of ncaa division i college football players using global positioning systems. *The Journal of Strength & Conditioning Research*, *30*, 11–19.

[43] Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, *2*, 37–52.

[44] Zhou, X.-H., Zhou, C., Lui, D., & Ding, X. (2014). *Applied missing data analysis in the health sciences*. John Wiley & Sons.