A comparative study on deep & machine learning techniques used for football injury prediction & prevention

MSc Research Project
Data Analytics

Michael Dunne
Student ID: 15420892

School of Computing
National College of Ireland

Supervisor: Noel Cosgrave

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Michael Dunne		
Student ID:	15420892		
Programme:	Data Analytics	**Year:**	2021
Module:	Research Project		
Supervisor:	Noel Cosgrave		
Submission Due Date:	16/12/2021		
Project Title:	A comparative study on deep & machine learning techniques used for football player injury prediction & prevention.		
Word Count:	7348	**Page Count:**	18

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Michael Dunne

Date: 16/12/2021

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A comparative study on deep & machine learning techniques used for football injury prediction & prevention

Michael Dunne
15420892

Abstract

Football as a sport is rapidly growing and has reached levels that no other sport was come close to in terms of popularity of watching, playing and the financial enterprise that surrounds this sport. A big part of this professional game is to keep players running at optimal performance and fitness levels to achieve all of the success within the competition and the financial rewards that come with this. To keep football players at this level of performance, the fitness screening of players health is common practice in today's world to negate players getting injured. The purpose of this research is to predict and prevent injuries occurring to footballers so they can preform at optimal fitness levels. This research aims to use deep learning and machine learning techniques to predict an injury occurring to a player and classifying that player as a high or low injury risk. The use of techniques will be compared and contrasted to establish which approach should be used for future research, techniques that have been used in research carried out in this field have been implemented with novel approaches to determine the best fit for the prediction of football injuries. The research looks at predicting and evaluating the injury proneness of a player using both regression and classification methods. The lowest MAE is achieved by DNN model when using LR selected data. This is closely followed by CNN model. However, the difference is not statistically significant. Support Vector Regression performs the worst in all experiments in terms of MAE.

1 Introduction

The sport of football is the most played sport around the world and one of the oldest dating back to the 1800s in England. Nowadays, the sport has transitioned into a yearly multibillion dollar business with some the world's wealthiest investing into teams. Football is a sport played between two teams made up of 11 players per team. Most professional teams have a squad of 25 players as this is maximum number of players allowed per team in a professional season, some leagues in different countries may alter this number but for the majority of major world football leagues this is the criteria to meet. The higher a team finishes in the league determines the reward size they are gifted via the leagues football association. This is all achievable to a football team that performs to the highest standard regularly within a league season. This is why the ability to predict injures occurring to players in a team so that precautions like squad rotation, training alternations and player restfulness can be utilised for maximum productivity within a league season for player performances.

Millions of players playing around the world professionally are at a very high risk of injuries occurring to them (Azzam et al, 2015), this is due to participating in games or training the week of these games in a league format season. The average professionally player in Europe which is the highest level of elite football in any continent is said to miss up to 37 days of gametime and training per league season(Woods, 2002). This means that because of the lack of availability in player selection, a team's performance will decline rapidly due to injuries(McCall, 2016). This in turn will add to financial loss(Ekstrand, 2016) because of commercial income due to sponsorships ending and attendance at games declining but also medical insurance costs of players with rehabilitation costing heavily if the players are seriously injured over a long period of time.

To negate all of these issues, the inclusion of effective injury prevention strategies by means of predicting injuries occurring will significantly help football teams. The main technique that this can be achieved is by developing risk injury factors for professional players, when a player injury risk profile is classified as high risk, the necessary precautions can begin to take effect. The athlete injury risk profile can be monitored and aid the clubs' sports scientists and medical teams in monitoring player health and fitness for optimal performances. Injury screening is widely used in professional football, and it mainly occurs during pre-season when players return just before the league season commences, injury screening allows targets to be set by the medical staff and rehabilitation progression for athletes. Unfortunately, some screening of players can be overlooked, or a misdiagnosis could occur which can lead to serious problems for the athlete and also the club in terms of liability.

The ability to assist clinicians in preventing player harm from occurring like muscular injuries to more serious risks like cardiac related is vital in the area of football sports science and machine learning could provide essential support in this field (Meeuwisse, 2007). Establishing that specific factors have greater influence on the occurrence of an injury occurring in a football game although merits great praise, the lack of current evidence in the game of football that is supported by an excellent standard of research is non-existent.(Woods, 2002).

If the correct approach is taken on data that is compatible with the methodology, then accurate injury predictions can be made with football players who are classed as potential high risk can be assessed further by the medical staff and can set injury prevention techniques thus preventing medical expenses and team performance declining, the predictive profile model can reinforce any diagnosis made by the medical staff throughout a league season.

The research done in injury risk prediction in football does not determine a clear methodology or technique that is widely regarded as the clearest, leaving the question of which approach is the best suited for injury prediction and prevention(Meeuwisse, 2007). Research in this field tends to depend on the data used for the study, many can be hard to acquire for most teams that is not at the elite level of European clubs due the lack of funding/technology available at their disposal like GPS tracking during training exercises or heatmaps throughout the game. Extensive research has been done to predict football match scores using machine learning techniques. However, football injury risk prediction when it comes to machine learning techniques is still novice in its approach.

This is further strengthened when looking at data containing ranked attributes of players consisting of their physical and mental abilities related to football, that any team in world football regardless of financial status can adopt. This research aims to establish that machine learning techniques can aid in the prevention of injuries occurring to football players without the need of exclusive data.

The objectives of this research are listed below.

1. To extensively research all literature regarding techniques that can be used for injury prediction as well as methodologies and evaluation metrics.

2. To implement an injury prediction model that can predict injury occurring to a specific player and classifying them at a low-high risk of injury.

3. To compare, contrast, and evaluate all of the machine learning techniques implemented and provide a conclusive answer to which technique should be used for future prediction models.

4. To establish if deep learning (neural nets) approaches are better suited to injury prediction by comparing results.

The state of the art in section 2 looks at the research already submitted in this area. The review will look at different methods and techniques used to predict football injuries occurring discussing the advantages and disadvantages of past approaches. Section 3 will look at how the data was gathered and cleaned in the methodology section. The research framework and architecture will be outlined in section 4. Section 5 will then build on this and talk about the implementation of said framework approaches. The paper will then conclude at section 6 when the evaluation of the techniques performances and conclude what can built upon this research in future work.

2 Related Work

2.1 Literature review of injury prediction in football

Eetvelde (2021) analyses machine learning approaches in sports injury prediction and prevention, the study suggests that methodological quality of research done in this field is acceptable but more than be built upon, especially when it comes into the interpretation of the models. Eetvelde links the scarce lack of quality data that is publicly available for injury prediction to poor performing machine learning models and suggests that only high quality data can be deemed acceptable for accurate results. Eetvelde mentioned Oliver (2020) who will be looked at in this review as being one of the best approaches to injury prediction. Several other research that has been done has been classed by Eetvelde as either biased (27%) of papers or lacked accurate methodological quality (63%). This research carried out supports the evaluation for the need of improvement in this area, whilst keeping an unbiased approach to the models/data.

Haji (2021) created an injury prediction model using CNN's on over 700 images displaying different parts of the body that had been scraped from googles advanced search. Haji used the sigmoid activation function for prediction of probability of each class and to improve the performance a loss function n-binary cross entropy loss is used. To optimize the model, the Adam Gradient Descent algorithm was used, and this seemed to work very effectively with the loss function, this is something that could be incorporated for the optimization stage. 668 images were used for training whilst 77 images were used for testing, the training accuracy rose to 97% when epochs=13 on the 7 layer CNN model. This model was also compared to a linear SVM detector in which it outperformed.

Huang (2021) proposes artificial neural networks for a sports injury prediction model. 21 football players were used for this experiment, a comparison of the ANN model was contrasted with PCA and decision trees. Although the ANN model performs well, this experiment leans toward being biased due to a number of players have missing data and an average value was entered. The ANN model produced an accuracy 95% compared to PCA(86%) and decision tree(89%). A field programme gate array (FPGA) tool was used for implementation algorithm.

He (2021) approaches sports injury prediction in juvenile players via text classification machine learning technology. The research used real time records of player fitness data between half a season on 48 college football players. Injuries were classified as mild, moderate, or severe, mild injuries accounted for over 40% of the total injury cases. He concluded that this experiment has the capacity to reduce rehabilitation costs drastically with improved training levels for the players based off this prediction model. A vector space model was used for text classification.

Oliver (2020) used six premier league and championship teams for his study in which he used tree based models, logistic regression and SVM. The purpose of this paper was to improve injury risk and prediction in elite male youth football players. From this paper, it can be seen that the main outcome is tree based models, they seem to be one of the most popular approaches to injury risk and prediction in football, this gives an advantage over other ML methods as these models are easy to read and visualize, with the extensive of boosting available also. This paper stats that the logistic regression model performed poorly in contrast to the decision trees. Oliver suggests that body size plays a major role in contributing to injuries, this can be challenged by using player weight and physicality features for the models.

Hughes (2020) uses a multivariable prognostic model to predict the risk of injury on over 150 professional players. The model created was a logistic regression model made up of 12 parameters, this was after they used multiple imputation to remove missing values and replace with substituted values on the data consisting of pre-season performance and injuring history data spanning five years. The authors also created a parsimonious model that used backward selection to remove any factors that surpassed the threshold. The model itself was evaluated with calibration which looks at the agreement stated as the observed outcome with the predictions made and decision curve analysis which is stated as a novel approach method for evaluating prediction models according to Vickers (2008).

Nikki (2020) also looks at injury risk in elite youth players based on physical performance features using extreme Gradient Boosting algorithms. A similar approach used by Sarlis(2021) which results in a near perfect accurate prediction model (99.9%). Nikki uses data from Belgian under 10s to under 15s youth academies taken from a season. Features in the data included similar features to the data such as strength, flexibility, speed, agility, and fitness. Predicted injuries that occurred to players were classified as overuse or acute injuries. The model predicting injury by extreme gradient boosting had precision of 84%, with a recall of 83% and an accuracy of 85% via f1 score. The classification model which depicted an injury as acute, or overuse had a slightly less accuracy and precision with 78%.

In (2019) Ayala worked on a model that will prevent hamstring injuries occurring to professional soccer players throughout the Spanish national soccer divisions. The authors used a pre-season evaluation of the 96 players which contained psychological measures like sleep quality of athlete and athlete burnout and muscular measures like different testing manoeuvres for joints and hamstring strength. The predictive model for hamstring injuries consisted of three decision tree algorithms (J48, AD Tree and SimpleCart)used for classifiers in each method. This paper concludes that a broad range of variables come into play when identifying whether a player is a high or low risk of injury due to real world settings

Wiik (2019) uses Long Short-Term Memory to identify peek readiness for football players to train, this prediction model uses data from two professional teams in Norway. Data was collected using a sport athlete monitoring system called PMSys, this system is stored on data storage unit on the amazon AWS cloud service. The LSTM model operated on a daily basis using the readiness values to predict the next day's values. The variables used to predict readiness to train included stress, mood, sleep quality, fatigue, and soreness. These were classified as either labelled 1-5 very bad to very good. The model had a precision ad recall above 90%.

Rossi (2018) where the use of GPS is incorporated for injury forecasting with machine learning. A multidimensional model was used on the data taken from 26 Italian professional players during a season which lasted 23 weeks gathering just over 930 training sessions to use for the model. 55 features in total are used for injury forecaster which is something to look to reach with this prediction research. A Decision tree is used to classify an injury on said dataset. Rossi (2019) uses recursive feature elimination with cross validation via python package scikit learn, this feature selection consists of each feature with maximum score on the validation data to be best suited, this in turn will reduce dimensionality. Results show that the decision tree can predict 80% of injuries and it stated that 50% of these injuries were labelled as injuries occurred in training sessions. The paper suggests that the model can be used for altering training schedules for players and improving fitness. The author suggests extending the research to include performance features from the games to be used.

Carey (2018) also investigates injury prediction modelling with Australian football. This has been done using GPS devices such as what was seen with Rossi (2018). Data was collected over 3 seasons and contains workload ratio, weighted moving average etc. Predictive models included SVM, logistic regression and random forest, logistic regression produced the best model for hamstring injuries with an AUC equalling 0.76. Underfitting was present with these models and the research stated that increasing the data was needed for better performing models.

Ruddy (2018) uses multiple machine learning techniques to predict hamstring strains in Australian football players. Injury history data of 186 footballers were gathered along with hamstring strength and demographic. AUC was used to compare the prediction of the hamstring strain occurring with the injury outcome class and the median value for the models were 0.26, 0.91 and 0.58. Logistic regression, neural network, random forest, SVM and Naïve bayes were all used for building predictive models.

Liu (2018) proposes a classification model based on random forest to predict sporting injuries of soccer players. The learning based system reduces attributes that would be significantly impactful on the risk of injury, then apply the Random Forest based algorithm. Liu suggest using the apriori algorithm for feature extraction to gather strong relative features to injury risks.

Michalowska (2017) also assess the risk of injury by using artificial neural networks, this time the data is on knee injury risk of over 60 football players. The parameters used were divided into 4 groups and used for 5 feedforward ANNs, 22 parameters were used which contained in depth muscle peak torque of specific leg muscle as well as body weight, muscle acceleration etc. The ANNs were created with MATLAB in a neural network toolbox, the models were trained using the Levenberg-Marquardt Backpropagation algorithm. Over 0.5 for the output value was classified as injured anything less was no injury(RMSE). This research could be improved by identifying specific parameters that are linked to predicted injuries.

Kampakis (2016) looked at predictive modelling of football injuries in three different investigations. Neural networks were used to predict the recovery time of an injured player using European football association injury data along with other ML algorithms like naïve bayes, random forest, SVM, KNN and logistic regression to compare results . Neural network produced the best model for the integrated dataset, but before feature selection random forest produced the best performance, this confirms that feature selection and the choices made can severely impact the model's performance. In the other study, NN's were used to predict injury incidents occurring based off of GPS in training data with. supervised PCA is the best performing model here whereas neural networks do not perform well where PCA and random forest performed well due to ability to handle the high number of noisy features.

McCullagh (2013) investigated how artificial neural networks can be used to predict sports injuries occurring in the Australian football league. Players were either classified as a high or low injury risk, this was based off of parameters used in the ANN model like workload, flexibility, and durability. This was then used with the week of training as an input which would then result in the ANN model predicting a high risk of the player getting injured next week or that the player will resume match fitness for the coming week. Tenfold cross validation was used for assessing performance of the model which overall the injury classification achieved 82% whilst the injury prediction correctly predicted 94%. The author states that future research could improve the classifying of injury risk like incorporating medium as well as high and low.

Venturelli (2011) uses a cox regression model for the prediction of injuries in young football players. The multivariate survival model looks at factors that contribute to thigh muscle injuries specifically and identifies a correlation between previous injuries and future injury risk for young players. They looked at risk factors like age, height body mass and exposure and then previous thigh injuries on survival probability. Then the cox proportional hazard model was used to evaluate predictions of thigh strain injuries occurring. The model found that injury risk factors included positions of players like defenders and midfielders as well as the height of the player and if the player had previously gotten injured.

2.2 Literature review of injury prediction throughout global sports

Sarlis (2021) looks at the impact of injuries with team performance on basketball players using multiple machine learning techniques including random forest, linear and nonlinear regression models, ANNs, SVM, and naive bayes. In terms of accuracy, the XGBoost tree model was the highest with 99.9% accuracy, whilst SVM produced poor accuracy results at 31%. The study stats that the team performance is gravely impacted negatively when injuries occur to the players in that team, players who also return back from an injury perform poorly. This shows the need for injury prediction models in sports as the injury risk assessment of players could be the difference in a successful season or an abysmal season. This is why it is needed to accurately determine when players need resting due to high injury risk.

Song (2021) compares and contrasts machine learning and logistic regression models in the prediction of kidney injuries. Extensive exploratory analysis of the data was done using one way ANOVA tests and T-tests to calculate mean differences between ML models and logistic regression models. Gradient boosting was once again the best performing ML model which has been a trend in this literature review. Song suggests logistic regression is equally effective at predicting kidney injuries occurring as other ML techniques.

Luu (2020) echoes in their research that ML models such as XGBoost and random forest outperform logistic regression models like Song (2021). This research was carried out on hockey players for predicting injury occurrence with a total of 85 performance metrics with injury prior historic data on over 2000 players spanning 10 years. The research predicted injuries that would occur in the next seasons games and out of all model approaches XGBoost produced the highest accuracy and an AUC of 0.949 beating logistic regression model that had 0.937.

This review has shown a lack of uncertainty when it comes to the approach for injury prediction techniques and methodologies to follow, with a variety of techniques and approaches used on data that can vary from biased to non sensical for the models that have been fitted. The use of neural network models is quite novice for injury prediction and research carried out alters from high preforming models to less accurate models compared to techniques like XGBoost. Many papers contradict each other in terms of ML algorithms that are seen as the best approach for injury prediction.

3 Research Methodology

For this research, Knowledge Discovery Database (KDD) was followed throughout the project timeline. The first stage of KDD is the raw data source, in terms of this research was the data was acquired from Kaggle in the form of a csv file containing the football manager player performance statistics. Data is converted for the correct data type for this research whilst data pre-processing begins on the dataset, this is where the removal of unwanted values or missing values. In this dataset, Nan-s are scanned for and replaced with the appropriate most frequent values or if numeric value, simply replace with the mean value. The final step of pre-processing stage is to transform the data into the structure desired, once the data transformation is complete, the data mining stage can begin.

The data mining stage begins with the coding to discover patterns in the data via the models designed for the predictions to come from, once all the models have been implemented, the last stage of this methodology is to visualise the predictions to gather information from the visualise aid.

This research aims for implementing machine learning models for the prediction of injury risk level to a professional football player and to determine what approach in machine learning works best for the prediction of injury occurrence. Several important features in this approach that have significant impact on injury occurrence were used.

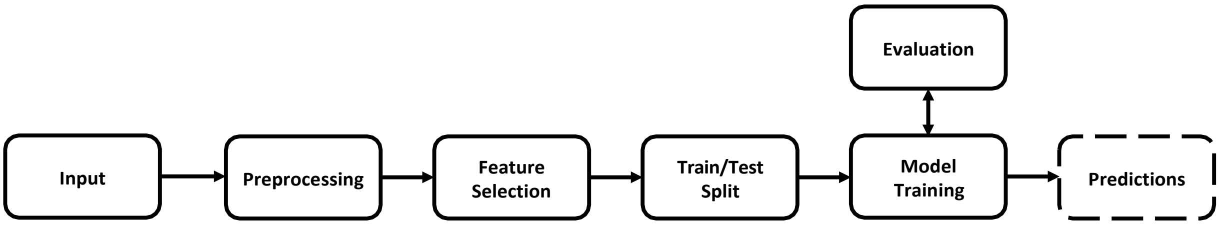

Figure 1: The overall framework approach of the research project.

The dataset consists of 84 features and 159, 541 examples. The label for each example reflects the specific attribute of the player. The attributes are ranged from 0 to 20. To facilitate the training process of the algorithms, the dataset is split into two distinct partitions with 80-20 ratio for train and test respectively. Table 1 shows the number of records in train and test splits.

Table 1: Dataset statistics

Split	No. of Examples
Train	127, 632
Test	31, 909
Total	159, 541

Two versions of the dataset have been prepared. (1) For regression task, where the label is a continuous value between 0-100, and (2) for classification task where the continuous label is discretized into three labels i.e., “Low Risk" ,“Medium Risk", and “High Risk" based on threshold.

Table 2: Dataset Class Distribution For Train and Test Splits in Classification Task

Split	Low Risk	Medium Risk	High Risk	Total
Train	65, 184	60, 677	1, 771	127, 632
Test	16, 296	15, 170	443	31, 909
Total	81, 480	62, 194	2, 214	159, 541

Feature selection is performed for both regression and classification versions of the dataset in order to remove the features that has negligible contribution towards correct predictions of the label. Four strategies to select features are applied. (1) First, selecting the top 50 features based on Random Forest feature importance. (2) Second, selecting features based on Support Vector coefficients values by training an SVM classifier. (3) Third, selecting features based on Linear Regression coefficients values by training a Linear Regression classifier. (4) Finally, using Principal Component Analysis (PCA) to select the components that count for 95% of the variance in the dataset. This way, there is 10 total variations of the dataset (i.e., 5 variations for regression and 5 variations for classification). Table 3 summarizes the number of features retained after each kind of feature selection method is applied.

Table 3: No. of features for each variation of the dataset.

Feature Selection Approach	Regression	Classification
No Selection	84	84
Random Forest	50	50
SVM	30	35
Linear Regression	32	36

PCA	8	8

3.1 Hyperparameters Tuning

Machine learning algorithms for regression and classification do not demand excessive amount of hyperparameters tuning. However, deep learning based models for classification and regression require hyperparameters to be tuned and it can have drastic effect on the performance of the models (Gorgolis, 2019). Thus, deciding to use the test data to set the hyperparameters for both DNN and CNN based models. This hyperparameter tuning is done using complete dataset (without any feature selection) on regression labels. Once the hyperparameters are selected, they are kept the same throughout the experiments. Table 4 shows the available choices for each model and the final selection. The grid search is used to find optimal hyperparameters. In terms of Dropout rate, it is observed that it has no significant impact on the performance of both DNN and CNN models. Thus, it is not used in the final model. With regards to activation function for hidden layers, ReLu turned out to be the best choice which is defined by max(0, x). For the choice of learning rate, 0.001 gave the highest performance paired with Adam optimizer. Specifically talking about CNN models, the first CNN layer has 64 filters while second CNN layer has 32 filters, which gave the lowest error values. For pooling strategy, Global Average Pooling is selected.

Table 4: Hyperparameters and Final choices selected (NNs)

Model Variation	Hyperparameter	Choices	Selected Value
DNN	Activation Function	ReLU, sigmoid, tanh	ReLU
	Dropout Rate	0.5, 0.4, 0.3, 0.2, 0.1, 0.0	0.0
	Optimizer	Adam, RMSProp, Adadelta	Adam
	Learning Rate	0.001, 0.002, 0.003, 0.005	0.001
	Number of Hidden Neurons	256, 128, 64	64
CNN	Activation Function	ReLU, sigmoid, tanh	ReLU
	Dropout Rate	0.5, 0.4, 0.3, 0.2, 0.1, 0.0	0.0
	Optimizer	Adam, RMSProp, Adadelta	Adam
	Learning Rate	0.001, 0.002, 0.003, 0.005	0.001
	CNN Filters for CNN Layer 1	256, 128, 64, 32	64
	CNN Filters for CNN Layer 2	256, 128, 64, 32	32
	Pooling Strategy	Avg Pooling, Max Pooling	Avg Pooling

4 Design Specification

The dataset consists of two versions i.e., for regression with continuous label and for classification with discretized label. Thus, examining both regression and classification models. Furthermore, for each kind of model, the use of machine learning as well as deep learning based models are applied. The detail of each model is presented in the subsequent sections. The inclusion of boosted models such as XGBoost, CatBoost, and gradient boost regression for this approach relates to the literature review where boosted models frequently outperformed other approaches like in the research carried out by Sarlis (2021) where XGBoost tree model produced 99.9% accuracy over SVM, naïve bayes and artificial neural networks. Luu (2020) and Song (2021) echoes this sentiment of XGBoost outperforming other model approaches. This research approaches the research question by including tree based models as seen by Oliver (2020) in which decision trees outperformed logistic regression approaches. The addition to ML approaches being used such as KNN and SVM came from a comparative study that Kampakis (2016) where deep neural networks outperformed these two models but only after feature selection, before feature selection PCA and KNN produced better accuracy, I will compare this in the approach to see any correlation between the different feature selections. The CNN model used for Haji(2021) played a key role in its inclusion for this research producing 97% accuracy compared to SVM for soccer injury prediction. The complete model approaches in both tasks are stated below in subsections.

4.1 Regression Models

The use of machine learning as well as deep learning based regression models and compare and contrast. The following subsection provides the list of models selected for the experiments on regression variations of the dataset.

4.1.1 Machine learning algorithms for Regression

The following ML algorithms were used:

1. Linear Regression LightGBM
2. XGBoost
3. CatBoost
4. Elastic Net Regression
5. Bayesian Ridge Regression
6. Gradient Boost Regression
7. Support Vector Regression

4.1.2 Deep Learning algorithms for Regression (Neural nets)

From deep learning family of algorithms, the development of deep learning neural network (DNN) architecture as well as convolutional neural network (CNN) architecture are used.

4.2 Classification Models

Similar to regression models, both machine learning and deep learning models for the classification task are used to compare how they do.

4.2.1 Machine learning algorithms for Classification

For the classification task, the following machine learning algorithms are used.

1. Support Vector Machine (SVM)
2. Guassian Naïve Bayes (GNB)
3. Stochastic Gradient Descent (SGD)
4. K nearest Neighbour (KNN)
5. Decision Tree
6. Gradient Boosting (GB)
7. LightGBM
8. XGBoost

9. CatBoost (CB)

4.2.2 Deep learning algorithms for classification (Neural nets)

From the deep learning family, DNN and CNN model are implemented, identical to the regression tasks. The only difference is the final layer where 3 neurons with SoftMax activation are used for classification (as there are three classes i.e., Low Risk, Medium Risk, High Risk of injury).

5 Implementation

The dataset consists of 84 features and 159, 541 examples. The label for each example reflects the players ability for that specific label, these labels are used in relation to proneness of injury. The dataset was then split into an 80/20 train and test split. Then, two versions of the dataset for the regression task which was continuous value between 0-100 and then dataset for the classification task which the continuous label is divided into three labels ranging from Low-High risk. Feature selection was then done on the data using random forest feature importance to identify the 50 most important features. SVM classifier, linear regression classifier and PCA for further feature selection based on coefficient values. Python version 3.7 was used for this research with google colab. The dataset was stored via google drive and linked by google colab library. All libraries used includes pandas, numpy, tensorflow, scikit-learn, matplotlib, and lightgbm. The machine that preformed this research configuration is: 8gb RAM, AMD RADEON R5 3 GHz, 64 bit OS.

6 Evaluation

For the regression models, mean absolute error (MAE) and R2 coefficient score for evaluating the performance of the metrics was used. These are standard metrics that are widely used in literature for measuring multi-class classification performances, these metrics are accuracy, precision, recall, and F1-score (Sokolova and Lapalme, 2009), where the latter 3 can be computed using micro-average or macro-average strategies. In micro-average strategy, each instance holds equal weight and outcomes are aggregated across all classes to compute a particular metric. This essentially means that the outcome would be influenced by the frequent class if class distribution is skewed. In macro-average however, metrics for each class are calculated separately and then averaged, irrespective of their class label occurrence ratio. This gives each class equal weight instead of each instance, consequently favoring the under-represented classes. To avoid this class distribution bias, macro-average values for precision, recall, and F1-score was chosen to report.

Table 5 presents the results on the regression task on each kind of dataset variation. The results are discussed in terms of mean absolute error (MAE) and R2 score. In terms of mean absolute error, LR selected data yields better performing models. The lowest MAE is achieved by DNN model when using LR selected data. This is closely followed by CNN model. However, the difference is not statistically significant. Support Vector Regression performs the worst in all experiments in terms of MAE, while simultaneously exhibiting worst performance in terms of R2 as well. When comparing the feature selection methods, LR based feature selection yields the least MAE when using the DNN model. It is evident that feature selection helps in case of regression models. For instance, on the complete data selection, DNN model yields 1.634 MAE and when select features uses Random Forest, the MAE is reduced to 1.625. However, when the LR based feature selection is done, the MAE is further reduced to 1.613. In terms of SVM based feature selection, DNN exhibits a slightly worse performance of 1.629. In case of PCA selected data, all regression models yield highest MAE as compared to other feature selection methods (including complete data). Similarly, R2 is also used to evaluate the decency of the models. It is observed that boosting based algorithms show a higher R2 score. The worst performance is seen by SVR when using features selected through random forest with R2 of −1.707. The best performance is achieved by LightGBM Regression with an R2 of 0.122 when complete data is used (without any feature selection method).

Table 6 presents the results on classification task for each kind of dataset variation. Results are discussed in terms of F1-score. However, for comprehensiveness's sake, accuracy, precision, recall, and ROC-AUC score are also provided. It is observed that highest performance is achieved using full complete data while PCA selected data yields lowest performance. Turning now to the model specific performance, The DNN model has the highest F1-score of 0.48, which is closely followed by all boosting based algorithms (GB, LightGBM, XGB). Boosting based algorithms also exhibiting an interesting behavior that these are consistently outperforming other models for all kinds of feature selection methods. In all experiments, the DNN model is the better model as compared to the CNN model by a slight margin. The worst performance is exhibited by SVM model. Specifically talking about feature selection methods, no concrete conclusion could be drawn. For some models, feature selection reduces the F1-score while for some models, the F1-score improves by a huge margin. For example, when using the complete data, DNN shows an F1-score of 0.48 but when using Random Forest based data selection, the performance is reduced to 0.45, which is 3% reduction. Similarly, when SVM based feature selection is used, DNN yields an F1-score of 0.46 and identical performance is achieved when LR selected data is used. Similar to the regression-based models, PCA based feature selection method yields least performance and the F1-score of DNN is reduced to 0.41. CNN based models achieve lower F1-score as compared to DNN models. Interestingly, Gradient Boosting based models achieve consistent performance across all feature selection methods (between 0.46 to 0.47), except feature selection-based method, which yields the lowest performance across all feature selection methods in general as well. Overall, across both tasks, DNN based model performs better as compared to CNN based model or any other model.

Table 6: Performance of Regression Models on Test Split

Dataset Variant	Model	Mean Absolute Error	R^2 Score
Complete Data	Linear Regression	1.737	0.080
	LightGBM Regression	1.684	0.122
	XGBoost Regression	1.699	0.102
	CatBoost Regression	1.687	0.121
	Elastic Net Regression	1.770	0.052
	Bayesian Ridge Regression	1.737	0.080
	Gradient Boost Regression	1.695	0.113
	Support Vector Regression	2.385	-0.603

	DNN	1.634	0.061
	CNN	1.686	0.042
Random Forest Selected Data	Linear Regression	1.741	0.078
	LightGBM Regression	1.686	0.121
	XGBoost Regression	1.702	0.098
	CatBoost Regression	1.687	0.121
	Elastic Net Regression	1.679	0.052
	Bayesian Ridge Regression	1.741	0.077
	Gradient Boost Regression	1.697	0.111
	Support Vector Regression	3.381	-1.707
	DNN	1.625	0.078
	CNN	1.654	0.059
SVM Selected Data	Linear Regression	1.732	0.073
	LightGBM Regression	1.706	0.091
	XGBoost Regression	1.721	0.072
	CatBoost Regression	1.709	0.087
	Elastic Net Regression	1.763	0.051
	Bayesian Ridge Regression	1.732	0.073
	Gradient Boost Regression	1.711	0.087
	Support Vector Regression	1.904	-0.159
	DNN	1.629	0.031
	CNN	1.654	0.033
LR Selected Data	Linear Regression	1.739	0.079
	LightGBM Regression	1.689	0.119
	XGBoost Regression	1.705	0.097
	CatBoost Regression	1.690	0.117
	Elastic Net Regression	1.767	0.053
	Bayesian Ridge Regression	1.739	0.079
	Gradient Boost Regression	1.698	0.110
	Support Vector Regression	1.835	-0.067
	DNN	1.613	0.056
	CNN	1.643	0.056
PCA Selected Data	Linear Regression	1.774	0.055
	LightGBM Regression	1.755	0.061
	XGBoost Regression	1.760	0.047
	CatBoost Regression	1.749	0.058
	Elastic Net Regression	1.771	0.052
	Bayesian Ridge Regression	1.773	0.054
	Gradient Boost Regression	1.752	0.060
	Support Vector Regression	2.211	-0.0350
	DNN	1.699	0.036
	CNN	1.707	0.042

Table 7: Performance of Classification Models on Test

Dataset Variation	Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC Score
Complete Data	Logistic Regression	0.60	0.62	0.41	0.41	0.72
	SVM	0.54	0.38	0.35	0.30	0.72
	GNB	0.49	0.41	0.52	0.38	0.68
	SGD	0.60	0.45	0.42	0.43	0.69
	KNN	0.59	0.42	0.40	0.40	0.62
	DT	0.59	0.42	0.43	0.42	0.58
	GB	0.67	0.58	0.47	0.47	0.77

	LightGBM	0.67	0.58	0.47	0.47	0.77
	XGB	0.67	0.62	0.47	0.47	0.76
	CB	0.67	0.62	0.46	0.46	0.77
	DNN	0.66	0.61	0.47	0.48	0.76
	CNN	0.64	0.45	0.44	0.42	0.74
Random Forest Selected Data	Logistic Regression	0.60	0.57	0.41	0.41	0.72
	SVM	0.59	0.43	0.42	0.41	0.72
	GNB	0.51	0.42	0.53	0.40	0.69
	SGD	0.62	0.42	0.42	0.41	0.70
	KNN	0.60	0.47	0.41	0.42	0.63
	DT	0.59	0.42	0.42	0.42	0.58
	GB	0.67	0.57	0.46	0.46	0.77
	LightGBM	0.67	0.60	0.47	0.47	0.77
	XGB	0.67	0.62	0.47	0.47	0.77
	CB	0.67	0.64	0.47	0.47	0.77
	DNN	0.66	0.58	0.45	0.45	0.76
	CNN	0.65	0.67	0.44	0.44	0.74
SVM Selected Data	Logistic Regression	0.61	0.54	0.41	0.42	0.71
	SVM	0.51	0.39	0.33	0.23	0.68
	GNB	0.55	0.41	0.48	0.41	0.68
	SGD	0.60	0.74	0.41	0.41	0.70
	KNN	0.60	0.46	0.41	0.42	0.62
	DT	0.59	0.41	0.41	0.41	0.58
	GB	0.67	0.57	0.46	0.46	0.76
	LightGBM	0.67	0.59	0.47	0.47	0.76
	XGB	0.67	0.58	0.46	0.46	0.76
	CB	0.67	0.62	0.46	0.46	0.76
	DNN	0.66	0.62	0.46	0.46	0.77
	CNN	0.65	0.64	0.45	0.44	0.75
LR Selected Data	Logistic Regression	0.60	0.58	0.41	0.41	0.72
	SVM	0.61	0.42	0.42	0.41	0.72
	GNB	0.54	0.42	0.51	0.41	0.70
	SGD	0.54	0.71	0.36	0.32	0.70
	KNN	0.60	0.45	0.41	0.41	0.63
	DT	0.59	0.42	0.42	0.42	0.58
	GB	0.67	0.58	0.47	0.47	0.76
	LightGBM	0.67	0.58	0.47	0.47	0.77
	XGB	0.66	0.63	0.47	0.47	0.76
	CB	0.67	0.64	0.46	0.46	0.77
	DNN	0.66	0.63	0.46	0.46	0.76
	CNN	0.65	0.78	0.44	0.44	0.75
PCA Selected Data	Logistic Regression	0.55	0.37	0.37	0.37	0.67
	SVM	0.52	0.35	0.35	0.35	0.65
	GNB	0.55	0.37	0.37	0.37	0.66
	SGD	0.53	0.38	0.37	0.37	0.55
	KNN	0.58	0.41	0.39	0.39	0.60
	DT	0.55	0.38	0.38	0.38	0.54
	GB	0.61	0.44	0.41	0.40	0.69
	LightGBM	0.61	0.43	0.42	0.40	0.70
	XGB	07.61	0.42	0.42	0.40	0.70
	CB	0.61	0.42	0.42	0.41	0.70
	DNN	0.62	0.43	0.42	0.41	0.70
	CNN	0.61	0.42	0.42	0.41	0.70

6.1 Discussion

Eetvelde (2021) in his research carried out states that machine learning approaches in sports injury prediction and prevention is scarce in terms of quality ad substance, mainly due to the lack of quality data and model interpretation. The research presented, shows that quality data on football players can be simplified to assessment level coaching standards of ranking player ability and fitness levels which can be applied to any professional club regardless of finances or advanced technology GPS tracking data, whilst maintaining advanced analysis with deep learning models that can outperform many ml techniques that have been used in this field. The neural network models were the best models in both classification and regression models respectfully, with boosting based algorithms also exhibiting an interesting behavior that these are consistently outperforming other models for all kinds of feature selection methods. DNN model was the best out of the two neural nets with CNN being 2nd overall and not too far behind. Answering one of the research questions on if deep learning approaches were better suited for injury prediction and prevention.

7 Conclusion and Future Work

In this research carried out, an empirical comparison of multiple regression and classification models on the task of player injury prediction. In the case of regression, the output is a continuous value representing player's proneness towards injury. While in case of classification, the injury proneness is discretized to get categorical labels of three intensities (lower risk, medium risk, and higher risk of injury). The model's result output concludes that DNN based models achieve a superior performance as compared to any other model in both cases (i.e., regression and classification task). Neural networks outperforming all other machine learning approaches and should be used in future work related to injury prediction, however all of the models produced good results with RMSE and MAE values in the regression task and data could play a role in specific models underperforming.

In terms of future work, one can examine the performance of LSTM, GRU, or transformer based models for prediction, especially with LTSM over the course of a season with data taken from specific league games. Furthermore, ensemble models can also be explored in order to combine the benefits of multiple models.

Acknowledgement

I would like to personally thank my supervisor and lecturer Dr Noel Cosgrave for his guidance, encouragement, and support over the past year in pushing me to reaching my full potential. I would like to thank my friends and classmates who gave me support and suggestions in relation to my research project. Finally, and most importantly, I would like to thank my mother for her constant love and support throughout my academic study over the past few years, I am forever grateful for this.

References

Ayala, F., Lopez, V., Alejandro, J., Prof. De Ste Croix, A. and Garcia, V. (2019). A prevention model for hamstring injuries in professional soccer: learning algorithms, 32(4): 279-300.

Azzam, M., Throckmorton, T., Smith, R. and Azar, F. (2015). The functional movement screen as a predictor of injury in professional basketball players, 26(6): 619-23.

Carey, D., Ong, K., Whiteley, R., Crossley, K. and Crow, J. (2017).Predictive modelling of training loads and injury in Australian football, 12(3): 165-190.

Eetvelde, H. V., Mendonca, L., Ley, C., Seil, R. and Tischer, T. (2021).Machine learning methods in sport injury prediction and prevention: a systematic review, 96(8): 39-55.

Ekstrand, J. (2016). Preventing injuries in professional football: thinking bigger and working together, 50(12):709-10.

Gorgolis, N., Hatzilygeroudis, I., Istenes, Z., and Gyenne, L. (2019).Hyperparameter optimization of lstm network models through genetic algorithm, pages 1-4.

Haji, A., Saraf, R., Pawade, D., Dalvi, A. and Siddavatam, I. (2021).Human body part detection and external injury prediction using convolutional neural network, 21(11): 109-130.

He, K. (2021).Prediction model of juvenile football players sports injury based on text classification technology of machine learning, 62(2): 1179-1208.

Huang, C. and Jiang, L. (2020). Data monitoring and sports injury prediction model based on embedded system and machine learning algorithm 54 (9): 1100-1239.

Hughes, T., Riley, R., Callaghan, M., and Sergeant, J. (2020).The value of preseason screening for injury prediction: The development and internal validation of a multivariable prognostic model to predict indirect muscle injury risk in elite football(soccer) players, 16(13): 554-600.

Kampakis, S. (2016). Predictive modelling of football injuries, 12(4): 49-88.

Liu, G., Sun, H., Bai, W., Li, H. and Zhang, Z. (2018). A learning-based system for predicting sport injuries, 130(9): 889-1090.

Luu, B., Wright, A., Haeberle, H., Karunta, J., and Makhni, E. (2020) Machine Learning Outperforms Logistic Regression Analysis to Predict Next-Season NHL Player Injury.

McCall, A., Dupont G., and Ekstrand, J. (2016). Injury prevention strategies, coach compliance and player adherence of 33 of the UEFA elite club injury study teams: a survey of team’s head medical officials.(12) 700-30.

McCullagh, J. and Whitfort, T. (2013).An investigation into the application of artificial neural networks to the prediction of injuries in sport, 46(1): 79-100.

Meeuwisse, W., Tyreman, H., and Hagel, B. (2007). A dynamic model of etiology in sport injury: the recursive nature of risk and caution, (3) 210-20

Michalowska, M., Walczak, T., Grabski, J. and Grygorowicz, M. (2017). Artificial neural networks in knee injury risk evaluation among professional football players, 29(10): 79-162.

Oliver, J., Ayala, F., Myer, G., and Lloyd, R. (2020). Using machine learning to improve our understanding of injury risk and prediction in elite male youth football players.

Rommers, N., Rossler, R., Verhagen, E., Verstockt, S., Hondt, E. and Vaeyens, R. (2020). A Machine Learning Approach to Assess Injury Risk in Elite Youth Football Players.

Rossi, A., Pappalardo, L., Cintia, P., Laia, F. and Medina, D. (2018). Effective injury forecasting in soccer with GPS training data and machine learning, 43(21): 889-1030.

Ruddy, J., Cormack, S., Whiteley., R. and Timmins, R. (2021). Modeling the risk of team sport injuries: A narrative review of different statistical approaches, 110(14): 569-770.

Sarlis, V., Chatziilias, V., Tjortjis., C. and Mandalidis, D. (2021). A data science approach analysing the impact of injuries on basketball players and team performance, 86(8): 2043-2120.

Sokolova, M. and Lapalme, G. (2009). A systemic analysis of performance measures for classification tasks. Information processing & management, (IPM), 45(4):427-437.

Song, X. Liu, X., Liu, F., and Wang, C. (2021). Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis 44: 11-32.

Venturelli, M. and Zanolla, L. (2017). Injury risk factors in young soccer players detected by a multivariate survival model, 43(7): 1157-1190.

Wiik, T. Johansen, H., Pettersen, S., Baptista, I., and Kupka, T. (2019). Predicting Peek Readiness-to-Train of Soccer Players Using Long Short-Term Memory Recurrent Neural Networks. 10 (2): 1310-1325.

Woods, C., Hawkins, R., Hulse, M., and Hodson A.(2002). The football Association medical research programme: an audit of injuries in professional football analysis of preseason injuries. (36)436-41.