Science and Medicine in Football

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/rsmf20

Optimising classification in sport: a replication study using physical and technical-tactical performance indicators to classify competitive levels in rugby league match-play

Victor Elijah Adeyemo, Anna Palczewska, Ben Jones, Dan Weaving & Sarah Whitehead

To cite this article: Victor Elijah Adeyemo, Anna Palczewska, Ben Jones, Dan Weaving & Sarah Whitehead (2024) Optimising classification in sport: a replication study using physical and technical-tactical performance indicators to classify competitive levels in rugby league match-play, Science and Medicine in Football, 8:1, 68-75, DOI: 10.1080/24733938.2022.2146177

To link to this article: https://doi.org/10.1080/24733938.2022.2146177

Published online: 14 Nov 2022.

SCIENCE AND MEDICINE IN FOOTBALL
2024, VOL. 8, NO. 1, 68–75
https://doi.org/10.1080/24733938.2022.2146177

Routledge
Taylor & Francis Group

RESEARCH ARTICLE

Optimising classification in sport: a replication study using physical and technical-tactical performance indicators to classify competitive levels in rugby league match-play

Victor Elijah Adeyemo [a,b,c,d], Anna Palczewska[a], Ben Jones[b,c,d,e,f], Dan Weaving[b,d] and Sarah Whitehead [b,d,g]

[a]School of Built Environment, Engineering and Computing, Leeds Beckett University, Leeds, UK; [b]Carnegie Applied Rugby Research (CARR) Centre, Carnegie School of Sport, Institute for Sport, Leeds Beckett University, Leeds, UK; [c]England Performance Unit, Rugby Football League, Leeds, UK; [d]Leeds Rhinos Rugby League Club, Leeds, UK; [e]School of Science and Technology, University of New England, Armadale, VIC, Australia; [f]Division of Exercise Science and Sports Medicine, Department of Human Biology, Faculty of Health Sciences, The University of Cape Town and the Sports Science Institute of South Africa, Cape Town, South Africa; [g]Leeds Rhinos Netball, Leeds, UK

ABSTRACT

Determining key performance indicators and classifying players accurately between competitive levels is one of the classification challenges in sports analytics. A recent study applied Random Forest algorithm to identify important variables to classify rugby league players into academy and senior levels and achieved 82.0% and 67.5% accuracy for backs and forwards. However, the classification accuracy could be improved due to limitations in the existing method. Therefore, this study aimed to introduce and implement feature selection technique to identify key performance indicators in rugby league positional groups and assess the performances of six classification algorithms. Fifteen and fourteen of 157 performance indicators for backs and forwards were identified respectively as key performance indicators by the correlation-based feature selection method, with seven common indicators between the positional groups. Classification results show that models developed using the key performance indicators had improved performance for both positional groups than models developed using all performance indicators. 5-Nearest Neighbour produced the best classification accuracy for backs and forwards (accuracy = 85% and 77%) which is higher than the previous method's accuracies. When analysing classification questions in sport science, researchers are encouraged to evaluate multiple classification algorithms and a feature selection method should be considered for identifying key variables.

ARTICLE HISTORY
Accepted 7 November 2022

KEYWORDS
Performance analysis; team sport; rugby league; machine learning; feature selection

Introduction

Sports analytics is a rapidly growing area under the broader scope of data science. This involves the use of sport-data and the application of various mathematical and/or statistical techniques, methods and algorithms (Morgulev et al. 2018). In the field of sports science, researchers and practitioners are faced with several analytical problems including visualization, regression, and classification. For example, visualization problems include displaying technical behaviours of Australian Football League players across multiple seasons by applying a non-metric multidimensional scaling technique (Woods et al. 2018). Regression problems include understanding the differences in technical and physical performance profiles between successful and less-successful professional rugby league teams via linear mixed models (Kempton et al. 2017). Classification problems in sport science have included the development of injury prediction models based on training load data by applying logistic regression (Carey et al. 2018) to classify injury occurrence.

One area relevant to classification analysis is the classification of players into competitive levels and the determination of the key physical and technical-tactical performance indicators (Burgess and Naughton 2010; Whitehead et al. 2021). This is important since young players are required to progress to senior competition as part of their development or compete at a higher level as a replacement for injured senior players. Through the use of microtechnology devices (Cummins et al. 2013; Whitehead et al. 2018) and notational analysis (Woods et al. 2018), match-play characteristics across different playing pathways can be quantified by their physical characteristics (e.g., total distance, maximum velocity, average speed) (Whitehead et al. 2019) and technical-tactical performance indicators (e.g., line breaks, defensive errors, try, missed tackles, play-the-ball wins) (Kempton et al. 2017; Gabbett and Hulin 2018).

In sports science, it is common for research designs that aim to address a classification problem to include multiple predictor variables. Therefore, it becomes important to evaluate the construct validity and reliability of each predictor variable included before analysis. Often, researchers and practitioners are still left with high dimensional and colinear variables following this process. To overcome multidimensional and multicollinearity of predictor variables (i.e., identify key predictor variables),

CONTACT Victor Elijah Adeyemo v.adeyemo@leedsbeckett.ac.uk School of Built Environment, Engineering & Computing, Leeds Beckett University, Headingley Campus, Leeds LS6 3QS, UK

studies typically conduct multiple univariate analyses by investigating each predictor and target variable values separately (Gabbett 2013). For example, Gabbett (2013) investigated the difference in external loads among rugby leagues players across two different competitive levels (i.e., the National Youth Competition and National Rugby League) using a repeated-measures analysis of variance on physical performance indicators. However, such an approach is limited as it doesn't consider the covariance of the data and the multiple models produced could increase classification models' error rates.

Alternatively, machine learning variable importance methods can be used to identify key predictor variables by selecting only the variables which are relatively important to the target variable values (Thornton et al. 2017). This approach has been implemented when establishing the important training load indicators to predict injury status (Thornton et al. 2017) and in establishing the importance of seven sleep components to the Pittsburgh Sleep Quality Index score (Halson et al. 2021). However, using machine learning variable importance methods is reported to be suboptimal in identifying key predictors to the target variable values (Williamson et al. 2021) and it affects classification accuracy. For example, Whitehead et al. (2021) identified key predictor variables by using a single random forest model to establish variable importance of technical-tactical and physical performance indicators to classify rugby league players into two competitive levels (i.e., senior and academy) based on their playing positions (i.e., backs and forwards) using another Random Forest classification model. Whitehead et al. (2021) reported 83% accuracy for backs and 68% for forwards. These accuracies can be improved using other classification techniques.

A more reliable and robust method to Whitehead et al. (2021) method is aggregating repeated random forest variable importance results which involve using a different number of variables per attempt (Calhoun et al. 2021). However, this method is computationally expensive and may still produce a suboptimal classification model. Alternatively, key physical and technical-tactical performance indicators can be identified by applying a feature selection method that possesses no bias to a specific classification algorithm (Mabayoje et al. 2016). All feature selection methods exist as Filter feature ranking, Filter feature subset selection, Wrapper-based and Embedded methods (Balogun et al. 2020; Chih-wen et al. 2020).

Filter feature ranking methods generate a ranked score for every variable based on statistical properties found in the data as computed by the method. Filter feature subset-selection implements heuristic and search methods to evaluate multiple subsets of variables to produce the best subset of key predictor variables as related to target variable values (Balogun et al. 2020; Chih-wen et al. 2020). Importantly, the sets of variables produced by filter feature ranking and filter feature subset-selection methods do not have a bias towards any classification method. On the other hand, the wrapper-based method is based on a computational greedy search method of the variable space for finding variables that improve the predictive performance of a particular classification algorithm. Similarly, embedded methods are intrinsic to machine learning algorithms that find the best features for split decisions while fitting a predictive model (Balogun et al. 2020; Chih-wen et al. 2020). Both wrapper-based and embedded feature selection methods are for improving the performance of a specific machine learning algorithm, as partially implemented by Whitehead et al. (2021) to optimize the Random Forest model for player-level prediction. In this study, a filter feature subset-selection method is considered because it outputs the best subset of key predictor variables and without bias to any classification algorithm.

Consideration should also be given to the training-test split method for developing classification models, which has been used in sport science for the prediction of football league tables and the performance of players (Pantzalis and Tjortjis 2020). Whitehead et al. (2021) also used this technique, which is limited as it performs model fitting and evaluation only once on the given data. A *k*-fold cross-validation method is an alternative that allows training and testing to be performed *k*-times and outputs the average score of any selected evaluation metric (e.g., accuracy). Moreover, there are many applicable machine learning algorithms to solve classification problems. They are broadly categorized by their learning schemes, such as conditional probability, functions, decision trees, neural networks, and instance-based learning (Witten et al. 2011).

Therefore, this study aims to introduce feature selection methods to optimize classification models performances in sports analytics and demonstrate this by improving the classification accuracy of rugby league seniors and academy players through the application of filter feature subset-selection method (i.e., Correlation-based feature subset using best-first search method) on the same data used by Whitehead et al. (2021) and evaluating multiple classification algorithms (i.e., Logistic Regression, Multi-Layered Perceptron, Naïve Bayes, Support Vector Machine, Random Tree and k-Nearest Neighbour) to find the best classification model.

Methods

Design

Whitehead et al. (2021) data were used in this study. This included 157 physical and technical-tactical variables and two target variable values (i.e., Academy and Senior). The physical indicators were derived from microtechnology data (Catapult S5, Catapult Innovations, Melbourne, Australia) while the technical-tactical indicators were expertly coded by analysts from filmed matches. See Whitehead et al. (2021) for a full description of the variable names, descriptions, and methods for collection. As per Whitehead et al. (2021), the dataset was divided into two positional groups (i.e., backs and forwards) across two competitive levels (i.e., Academy and Senior). The backs dataset contained 453 match observations (Academy = 220; Senior = 233). The forwards dataset contained 527 match observations (Academy = 251; Senior = 276). Two phases of data analyses were conducted using the datasets. Phase 1 analysis involved identifying key performance variables while Phase 2 involved developing improved classification models.

Framework for data analyses

The framework shown in Figure 1 captures the two phases of data analysis. It was applied to backs and forwards positioning groups respectively.

In phase I, the 'Correlation-based feature subset' (Cfs) feature selection method was applied on the 157 physical and technical-tactical performance indicators to identify key ones.

The Correlation-based feature subset is an example of a filter feature subset-selection method that output the subset of variables with the highest score according to the heuristic evaluation function (Hall 1999; Ali et al. 2020). The score is calculated as follows:

$$M_s = \frac{l\overline{t_{cf}}}{\sqrt{l + l(l-1)\overline{t_{ff}}}}, \quad (1)$$

where M_s holds the score after evaluating a subset of S consisting of l variables, $\overline{t_{cf}}$ is the average correlation values between subset variables and target variable values, and $\overline{t_{ff}}$ is the average correlation values between subset variables (Hall 1999).

A dataset was extracted based on the output subset of performance indicators identified through this process and referred to as the 'reduced dataset'. More so, the original dataset with all 157 performance indicators is referred to as the 'full dataset'.

In phase II, classification models were developed and evaluated using both full and reduced datasets. Six classification algorithms were chosen based on their learning method, namely: Random Tree, Naive Bayes, Logistic Regression, Multilayered Perceptron, Support Vector Machine, and k-Nearest neighbour.

The Random tree algorithm uses a divide and conquer learning method. It constructs an unpruned decision tree by randomly choosing certain numbers of variables at each (split) node while it allows the estimation of class probabilities through a process called backfitting (Khabat et al. 2020).

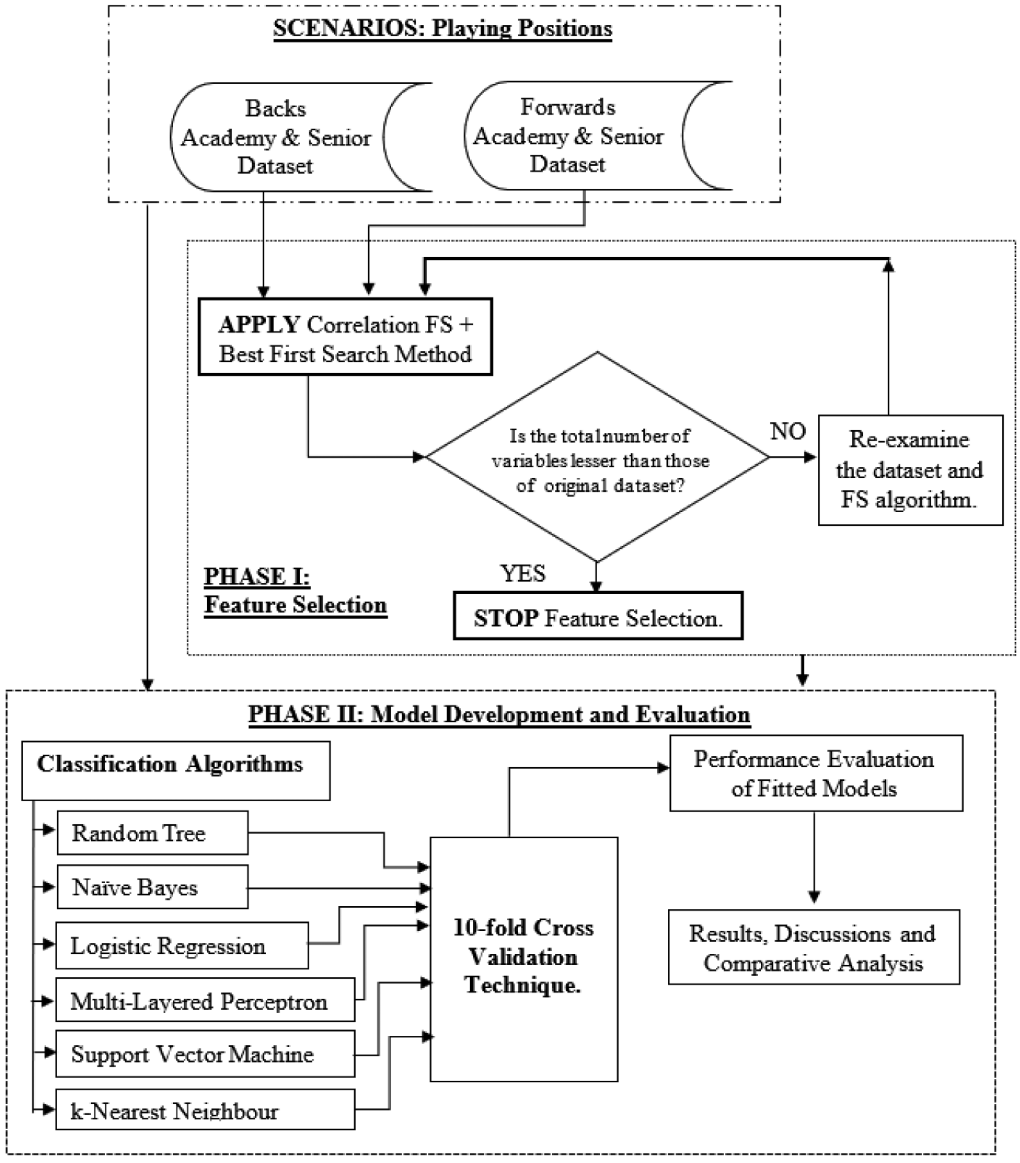

Figure 1. Experimental framework.

Naïve Bayes is a conditional probability-based classification machine learning algorithm. It produces predictive models based on the Bayes theorem that infers that all variables are independent of themselves (Elijah et al. 2019; Shengle et al. 2020). Logistic Regression is a statistical analysis technique used for predictive modelling such that the vectors of the independent variables are used to predict the target variable values (Balogun et al. 2019; Wilkens 2021). The fitting of a Logistic Regression model is achieved through maximum likelihood where the optimal vectors and a constant is being determined.

Multi-Layered Perceptron classification machine learning algorithm is based on Artificial Neural Network (ANN) and it represents black-box learning method. It is implemented as layers of input, (multiple) hidden and output interconnected neurons (Mabayoje et al. 2016; Sharma et al. 2019). Support Vector Machine is another black-box learning method but differs from the Multi-Layered Perceptron algorithm as it implements functional margin for discrimination of observations between target variable values (Gauthama Raman et al. 2020; Wilkens 2021). The performance of a support vector machine model is usually optimized by applying a suitable kernel function. k-Nearest Neighbour is a lazy learning and instance-based classification machine learning algorithm. It applies distance metrics (i.e., Manhattan, Euclidean, Jaccard etc.) to separate two instances in a set of *k* observations (Mabayoje et al. 2019; Kasongo and Sun 2020). The parameter *k* is used for determining the number of closest instances of the observation whose target variable value is to be predicted. The *k* parameter was set to 5 for this study.

In this study, the target variable values refer to player level (i.e., Senior and Academy) and the modelling task is to predict which level a player belongs to. The models were developed using 10-fold cross-validation (Alsariera et al. 2020) (Figure 1). The 10-fold cross-validation technique splits data into 10 subsets, nine subsets are used for training models while the remaining one subset is used for testing. It is repeated 10 times using each subset as a testing set. The results are averaged over 10 iterations. Classification models are evaluated using the following evaluation metrics: time taken, kappa value, confusion matrix and Area under Curve (AUC). Kappa value is the measurement of chance. It is the subtraction of agreement expected by chance from the observed agreement and divided by the maximum possible agreement. Kappa is calculated as follows:

$$Kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \quad (2)$$

The confusion matrix (Niyaz et al. 2016) for each model contains values of correctly and incorrectly classified instances for the values of the target variable (Figure 2). Several model evaluation metrics can be obtained from the confusion matrix, the main ones used in this study are presented with details in Table 1.

Confusion Matrix		
	Senior	Academy
Senior	TP	FN
Academy	FP	TN

Figure 2. A typical confusion matrix.

Area Under Curve is the classification model degree of separability of the majority and minority target variable values (Adeyemo et al. 2020). All analyses were conducted using Waikato Environment for Knowledge Analysis (WEKA) GUI software version 3.8.5 on an Intel Core i5 CPU with 8GB RAM. Parameter optimization was not considered for any of the algorithms selected in this study. All experiments are reproducible using the WEKA software.

Results

Phase I

The Correlation-based feature subset filter method identified 15 key performance indicators out of the original 157 variables for senior and academy backs (Table 2). The subset of the 15 key performance indicators had the highest score (0.277) out of all 2955 subsets.

The Correlation-based feature subset filter method identified 14 key performance indicators out of the original 157 variables for senior and academy forwards (Table 3). The subset of the 14 key performance indicators had the highest score (0.22) out of all 2683 subsets.

Seven variables were common (High-Speed Distance active, Collision active, Absolute average acceleration [120 s], Tackle Duration [240 s], Player Load 2D, Defensive Collision: Collision Lost, and Defensive Collision: Dominant Hit) between forwards and backs.

Phase II

The performances of the six classification models applied on the reduced dataset to improve the classification accuracy for senior and academy rugby league backs were better than the performances of the same classification models when applied on the full dataset (Table 4). The comparative analysis of the six different classification models' performances reveals 5-nearest neighbour as the best performing classification model for classifying senior and academy between rugby league backs

Table 1. Performance metrics extracted from the confusion matrix.

Metric	Equation	Description
True Positive Rate (TPR)	$TP = \frac{TP}{TP+FN}$	The score of correctly classified senior players.
True Negative Rate (TNR)	$TN = \frac{TN}{TN+FP}$	The score of correctly classified academy players.
False Positive Rate (FPR)	$FP = \frac{FP}{FP+TN}$	The score of academy players being incorrectly classified as senior players.
False Negative Rate (FNR)	$FN = \frac{FN}{FN+TP}$	The score of senior players being incorrectly classified as academy players.
Accuracy	$\frac{TP+TN}{TP+FP+TN+FN} X100$	The percentage of senior and academy players that are correctly predicted as senior and academy players.

Table 2. List of optimal performance indicators of backs.

S/N	Ranked List of Optimal Variables	Technical-tactical or physical variable	Senior (mean ± SD)	Academy (mean ± SD)	Cohen's Effect Size
1	High speed distance active	Physical	584.45 ± 178.46	467.9 ± 189.39	0.634
2	Relative collision count	Physical	33.95 ± 12	25.82 ± 12.74	0.635
3	Sprint Meters Per Min	Physical	0.71 ± 0.5	0.83 ± 0.72	−0.198
4	Absolute acceleration average (120 seconds)	Physical	0.72 ± 0.07	0.69 ± 0.06	0.519
5	Tackle Duration (240 seconds)	Physical	17.57 ± 6.25	14.93 ± 5.32	0.454
6	Player Load™	Physical	749.26 ± 160.91	620.18 ± 158.67	0.808
7	Player Load 2D™	Physical	475.33 ± 101.95	379.16 ± 98.16	0.96
8	Player Load Slow™	Physical	285.67 ± 60.2	247.94 ± 62.08	0.617
9	Carry: run	Technical	12.3 ± 7.81	7.79 ± 6.07	0.642
10	Missed tackle	Technical	0.95 ± 1.2	1.41 ± 1.9	−0.292
11	Positive Offload	Technical	0.66 ± 0.93	0.28 ± 0.69	0.463
12	Carry type: hit up	Technical	2.11 ± 2.69	0.81 ± 1.45	0.601
13	Defensive collision: collision lost	Technical	0.85 ± 1.4	0.51 ± 0.78	0.295
14	Defensive collision: dominant hit	Technical	0.75 ± 1.23	0.27 ± 0.59	0.498
15	Quick play-the-ball	Technical	2.21 ± 2.4	0.65 ± 0.93	0.85

Table 3. List of optimal performance indicators of forwards.

S/N	Ranked List of Optimal Variables	Technical-tactical or physical variable	Senior (mean ± SD)	Academy (mean ± SD)	Cohen's Effect Size
1	High speed distance active	Physical	278.77 ± 157.12	234.59 ± 111.25	0.32
2	Relative collision count	Physical	43.89 ± 15.99	36.6 ± 12.49	0.51
3	Meters Per Min (i.e., average match-speed)	Physical	79.32 ± 8.49	81.89 ± 9.94	0.28
4	Absolute acceleration average (60 seconds)	Physical	0.92 ± 0.22	0.83 ± 0.072	0.52
5	Absolute acceleration average (120 seconds)	Physical	0.73 ± 0.1	0.65 ± 0.5	0.53
6	Tackle Duration (240 seconds)	Physical	21.79 ± 7.5	19.65 ± 7.04	0.29
7	Absolute acceleration average (300 seconds)	Physical	0.58 ± 0.05	0.56 ± 0.05	0.55
8	Number of collisions within high-speed running Distance (480 seconds)	Physical	6.09 ± 3.33	4.84 ± 2.66	0.42
9	Player Load 2D™	Physical	373.18 ± 134.43	336.74 ± 118.22	0.29
10	Carry type: run	Technical	1.85 ± 2.38	0.91 ± 1.36	0.48
11	Defensive collision: collision lost	Technical	2.48 ± 2.53	0.94 ± 1.29	0.76
12	Defensive collision: dead stop	Technical	0.36 ± 0.63	0.47 ± 0.96	0.14
13	Defensive collision: dominant hit	Technical	1.49 ± 1.7	0.9 ± 1.23	0.39
14	Defensive play-the-ball loss	Technical	4.43 ± 3.87	1.66 ± 1.87	0.9

(Table 4). The 5-nearest neighbour classification model developed on the reduced data had the highest accuracy of 84.55%, highest correctly classified senior players (i.e., 0.81 TPR), highest correctly classified academy players (i.e., 0.88 TNR), lowest misclassification of academy players as senior players (i.e., 0.12 FPR), lowest misclassification of senior players as academy players (i.e., 0.19 FNR), highest kappa score of 0.69 and highest AUC score of 0.92.

The performances of the six classification models applied on the reduced dataset to improve the classification accuracy for senior and academy rugby league forwards were better than the performances of the same classification models when applied on the full dataset (Table 5). Also, the comparative analysis of the six different classification models' performances reveals 5-nearest neighbour as the best performing classification model for classifying rugby league forwards into senior and academy (Table 5). The 5-nearest neighbour classification model developed on the reduced data had the highest accuracy of 77.42%, highest correctly classified senior players (i.e., 0.76 TPR), lowest misclassification of senior players as academy players (i.e., 0.24 FNR), and highest kappa score of 0.55. However, the Multi-Layered Perceptron classification model developed on the reduced dataset had the highest correctly classified academy players (i.e., 0.82 TNR), lowest misclassification of academy players as senior players (i.e., 0.18 FPR), and highest AUC score of 0.84.

The most accurate classification model among the selected six machine-learning methods is the 5-nearest neighbour

Table 4. Summative performance of all classification models for backs using all variables and those identified through the correlation feature subset method.

Models	Dataset	Accuracy (%)	True Positive Rate	False Positive Rate	True Negative Rate	False Negative Rate	Kappa	AUC	Time (Secs)
Logistic Regression	Full	75.28	0.74	0.23	0.77	0.26	0.51	0.8	0.26
	Reduced	81.46	0.8	0.17	0.83	0.2	0.63	0.89	0.01
Multi-Layered Perceptron	Full	78.37	0.76	0.19	0.81	0.24	0.57	0.85	2.84
	Reduced	83.22	0.81	0.14	0.86	0.19	0.67	0.9	.0.07
Naïve Bayes	Full	70.86	0.77	0.36	0.64	0.23	0.42	0.79	0.01
	Reduced	78.15	0.77	0.21	0.79	0.23	0.56	0.86	0.01
Support Vector Machine	Full	73.73	0.75	0.27	0.73	0.25	0.47	0.73	0.08
	Reduced	74.39	0.72	0.23	0.77	0.28	0.49	0.75	0.06
Random Tree	Full	64.68	0.65	0.36	0.65	0.35	0.29	0.65	0.01
	Reduced	77.48	0.8	0.25	0.75	0.2	0.55	0.77	0.01
5-Nearest Neighbour	Full	76.82	0.78	0.24	0.76	0.22	0.54	0.84	0.01
	Reduced	**84.55**	**0.81**	**0.12**	**0.88**	**0.19**	**0.69**	**0.92**	0.01

Table 5. Summative performance of all classification models for forwards using all variables and those identified through the correlation feature subset method.

Models	Dataset	Accuracy (%)	True Positive Rate	False Positive Rate	True Negative Rate	False Negative Rate	Kappa	AUC	Time (Secs)
Logistic Regression	Full	74.57	0.77	0.28	0.72	0.23	0.49	0.8	1.41
	Reduced	74	0.72	0.24	0.76	0.28	0.48	0.83	0.04
Multi-Layered Perceptron	Full	74.95	0.71	0.21	0.79	0.29	0.5	0.84	5.49
	Reduced	75.9	0.7	**0.18**	**0.82**	0.3	0.52	**0.84**	0.3
Naïve Bayes	Full	68.88	0.75	0.38	0.62	0.25	0.37	0.78	0.03
	Reduced	74.19	0.71	0.22	0.78	0.29	0.49	0.84	0.02
Support Vector Machine	Full	72.3	0.74	0.3	0.7	0.26	0.44	0.72	0.12
	Reduced	74	0.67	0.19	0.81	0.33	0.48	0.74	0.05
Random Tree	Full	61.29	0.63	0.41	0.59	0.37	0.22	0.61	0.01
	Reduced	70.59	0.74	0.3	0.67	0.26	0.41	0.7	0.01
5-Nearest Neighbour	Full	70.97	0.68	0.26	0.74	0.32	0.42	0.78	0.01
	Reduced	**77.42**	**0.76**	0.21	0.79	**0.24**	**0.55**	0.83	**0.01**

classification model developed on the reduced datasets for backs and forwards respectively.

Discussion

This study solves a common classification problem in sport regarding identifying key physical and technical-tactical performance indicators that help classify between senior and academy rugby league players without bias to any classification machine learning algorithm. Through the obtained results, fifteen key performance indicators were identified as key performance indicators for differentiating between senior and academy levels in the backs and 14 key performance indicators were identified for forwards (Tables 2 and 3).

Significant differences in seniors and academy players were observed in the key performance indicators identified by the Correlation-based feature subset filter method to improve classification accuracy. Senior rugby league backs observed more high-speed distance, more relative collision count, accumulated more player load, completed more 240 s tackle duration and performed more carries than academy among others, while academy rugby league players only observed more sprint meters per min than seniors. Cohen's effect size analysis of the identified key variables reveals nine (9) of the 15 identified key variables for backs have a large effect size between senior and academy backs while three key variables had moderate effect size between both positional groups (Table 2). Two of the three key indicators with small effect size between seniors and academy backs was seen between academy and seniors (Table 2).

For rugby league forwards, senior players performed greater workload than academy players such as increased high-speed distance, relative collision, 60-s absolute average, 240 s tackle duration, defensive collision dead-stop, etc., while academy players recorded more meters per minute than senior players (Table 3). Six key performance indicators have a large effect size between senior and academy forwards while four indicators had moderate effect size between the positional groups (Table 3). Two of the four key performance indicators with small effect sizes occurred between academy forwards and senior forwards (Table 3).

The study (Woods et al. 2018) compared elite youth and senior Australian National Rugby Leagues game-play characteristics and reported that elite youth are usually not exposed to higher physical demands (e.g., tackling capacity) compared to senior players. Gabbett (Gabbett 2013) also reported higher physical demands among National Rugby League players during competitive matches than National Youth competition players. These studies (Gabbett 2013; Woods et al. 2018) further validate the key performance indicators identified by the Correlation-based feature subset filter method.

Whitehead et al. (2021) identified nine key variables for backs and three variables for forwards. For the backs, two of the fifteen performance indicators identified in phase I was common to those identified by Whitehead et al. (2021) (i.e., Player Load 2D and Player Load Slow). For forwards, there were no common performance indicators identified between both studies despite analysing the same data. This is because the variables identified by Whitehead et al. (2021) are specifically to increase the predictive performance of the Random Forest classification algorithm whereas the Correlation-based feature subset filter method applied in this study is not specific to any classification algorithm. This highlights the importance of applying a feature selection method rather than applying variable importance for identifying key variables. The study (Thornton et al. 2017) that applied the variable importance method to identify key training load variables to predict injury status and the study (Halson et al. 2021) that identified key seven sleep components to the Pittsburgh Sleep Quality Index score suffer similar limitation to the study (Whitehead et al. 2021), which can be resolve by applying feature selection method.

Having developed and comparatively evaluated six classification models, the classification models developed using the reduced datasets outperformed those of the full dataset despite including fewer predictor variables in the models (Tables 4 and 5). This is due to the removal of performance indicators that are strongly correlated among themselves and those not strongly correlated to the target variable values through the application of the Correlation-based feature subset-selection method. The best classification model of this study involved using Correlation-based feature subset-selection method to identify key performance indicators and 5-nearest neighbour algorithm to develop a classification model with improved accuracy.

On the other hand, Whitehead et al. (2021) involved using a single random forest model variable importance method to identify key performance indicators and another random forest classification model to classify between senior and academy players via a single attempt of a train-test split method for model development. Whitehead et al. (2021) reported a classification accuracy of 82.0% for backs and 67.5% for forwards. In contrast, the 5-nearest neighbour model in this

study, fitted on the reduced dataset produced an accuracy of 84.55% for classifying backs and an accuracy of 77.42% for classifying forwards. Nonetheless, there are other classification models from this study fitted on the reduced dataset that outperformed Whitehead et al. (2021) reported accuracy. The Multi-Layered Perceptron had a classification accuracy of 83.22% for backs and 75.9% for forwards, and all six classification models for forwards. The performances of this study's methods are directly linked to the underlying performance indicators used in classification model development. Therefore, the overall findings of the current study suggest that studies should avoid using the classification model variable importance method to identify key performance variables for generic use and to avoid using a single train-test split method for fitting classification model for sports analytics,

Conclusions

This study fulfilled its aim by improving the classification accuracy of senior and academy rugby league players, in comparison to a previously published study by Whitehead et al. (2021). Correlation-based feature subset-selection using the best-first search method as a feature selection method identified key physical and technical-tactical performance indicators for improving classification accuracy of rugby league senior and academy levels. The development of multiple classification models experimentally produced the best performing model with better predictive ability than the existing method.

In the attempt to identify key performance indicators to classify senior and academy players backs, a balanced set of physical and technical-tactical performance indicators were discovered for backs. Whereas more physical performance indicators were identified than technical indicators for forwards.

Based on the findings of this study, it is recommended that the application of a feature selection method is used before classification model development, evaluation, and improvement. Also, we encourage the development of classification models using various classification machine learning algorithms from different categories before selecting and presenting the best-performing methods. It is also recommended to develop a classification model via a 10-fold cross-validation method.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The author(s) reported there is no funding associated with the work featured in this article.

ORCID

Victor Elijah Adeyemo http://orcid.org/0000-0002-8398-3609
Sarah Whitehead http://orcid.org/0000-0002-6105-3160

Informed consent

The study got the ethics approval of the Institutions Ethics Committee and written informed consent was obtained from all participants who are completely anonymized and cannot be identified through this study.

References

Adeyemo VE, Balogun AO, Mojeed HA, Akande NO, Adewole KS. 2020. Ensemble-based logistic model trees for website phishing detection. International Conference on Advances in Cyber Security. 1347 (February):627–641. doi:10.1007/978-981-33-6835-4_41.

Ali A, Qadri S, Mashwani WK, Brahim Belhaouari S, Naeem S, Rafique S, Jamal F, Chesneau C, Anam S. 2020. Machine learning approach for the classification of corn seed using hybrid features. Int J Food Prop. 23 (1):1097–1111. doi:https://doi.org/10.1080/10942912.2020.1778724.

Alsariera YA, Adeyemo EV, Balogun AO. 2020. Phishing website detection: forest by penalizing attributes algorithm and its enhanced variations. Arab J Sci Eng. 45(12):10459–10470. doi:10.1007/s13369-020-04802-1.

Balogun AO, Basri S, Abdulkadir SJ, Adeyemo VE, Imam AA, Bajeh AO. 2019. Software defect prediction: analysis of class imbalance and performance stability. J Eng Sci Technol. 14(6):3294–3308.

Balogun AO, Basri S, Mahamad S, Abdulkadir SJ, Almomani MA, Adeyemo VE, Al-Tashi Q, Mojeed HA, Imam AA, Bajeh AO. 2020. Impact of feature selection methods on the predictive performance of software defect prediction models: an extensive empirical study. Symmetry. 12 (7):1147. doi:https://doi.org/10.3390/sym12071147.

Burgess DJ, Naughton GA. 2010. Talent development in adolescent team sports: a review. Int J Sports Physiol Perform. 5(1):103–116. doi:10.1123/ijspp.5.1.103.

Calhoun P, Levine RA, Fan J. 2021. Repeated measures random forests (RMRF): identifying factors associated with nocturnal hypoglycemia. Biometrics. 77(1):343–351. doi:10.1111/biom.13284.

Carey DL, Ong K, Whiteley R, Crossley KM, Crow J, Morris ME. 2018. Predictive modelling of training loads and injury in Australian football. Int J Comput Sci Sport. 17(1):49–66. doi:10.2478/ijcss-2018-0002.

Chih-wen C, Tsai Y, Chang F, Lin W. 2020. Ensemble feature selection in medical datasets: combining filter, wrapper, and embedded feature selection results. Expert Syst. 37(5):e12553.

Cummins C, Orr R, O'Connor H, West C. 2013. Global positioning systems (GPS) and microtechnology sensors in team sports: a systematic review. Sports Med. 43(10):1025–1042. doi:10.1007/s40279-013-0069-2.

Elijah AV, Abdullah A, JhanJhi NZ, Supramaniam M, Balogun Abdullateef O. 2019. Ensemble and deep-learning methods for two-class and multi-attack anomaly intrusion detection: an empirical study. Int J Adv Comput Sci Applic. 10(9):520–528. doi:10.14569/ijacsa.2019.0100969.

Gabbett TJ. 2013. Influence of playing standard on the physical demands of professional rugby league. J Sports Sci. 31(10):1125–1138. doi:10.1080/02640414.2013.773401.

Gabbett TJ, Hulin BT. 2018. Activity and recovery cycles and skill involvements of successful and unsuccessful elite rugby league teams: a longitudinal analysis of evolutionary changes in National Rugby League match-play. J Sports Sci. 36(2):180–190. doi:10.1080/02640414.2017.1288918.

Gauthama Raman MR, Somu N, Jagarapu S, Manghnani T, Selvam T, Krithivasan K, Shankar Sriram VS. 2020. An efficient intrusion detection technique based on support vector machine and improved binary gravitational search algorithm. Artif Intell Rev. 53:3255–3286. Springer Netherlands. doi:10.1007/s10462-019-09762-z.

Hall MA. 1999. Correlation-based feature selection for machine learning. Doctoral dissertation, The University of Waikato.

Halson SL, Johnston RD, Appaneal RN, Rogers MA, Toohey LA, Drew MK, Sargent C, Roach GD. 2021. Sleep quality in elite athletes: normative values, reliability and understanding contributors to poor sleep. Sports Med. 52: 417–426. doi:10.1007/s40279-021-01555-1.

Kasongo SM, Sun Y. 2020. Performance analysis of intrusion detection systems using a feature selection method on the UNSW-NB15 dataset. J Big Data. 7(1). doi:10.1186/s40537-020-00379-6.

Kempton T, Sirotic AC, Coutts AJ. 2017. A comparison of physical and technical performance profiles between successful and less-successful

professional rugby league teams. Int J Sports Physiol Perform. 12 (4):520–526. doi:10.1123/ijspp.2016-0003.

Khabat K, Cooper JR, Daggupati P, Pham BT, Bui DT. 2020. Bedload transport rate prediction: application of novel hybrid data mining techniques. J Hydrol. 585:124774. doi:10.1016/j.jhydrol.2020.124774.

Mabayoje MA, Balogun AO, Ameen AO, Adeyemo VE. 2016. Influence of feature selection on multi-layer perceptron classifier for intrusion detection system. Comput Inf Syst Dev Inf Allied Res J. 7(4):87–94.

Mabayoje MA, Balogun AO, Jibril HA, Atoyebi JO, Mojeed HA, Adeyemo VE. 2019. Parameter tuning in KNN for software defect prediction: an empirical analysis. Jurnal Teknologi Dan Sistem Komputer. 7(4):121–126. doi:10.14710/jtsiskom.7.4.2019.121-126.

Morgulev E, Azar OH, Lidor R. 2018. Sports analytics and the big-data era. Int J Data Sci Anal. 5(4):213–222. doi:https://doi.org/10.1007/s41060-017-0093-7.

Niyaz Q, Sun W, Javaid AY. 2016. A deep learning based DDoS detection system in software-defined networking (SDN). ICST Trans Secur and Saf. 4(12):153515. November. 2016. 10.4108/eai.28-12-2017.153515.

Pantzalis VC, Tjortjis C. 2020. Sports analytics for football league table and player performance prediction. 11th International Conference on Information, Intelligence, Systems and Applications (IISA) IEEE. p, 1–8.

Sharma J, Giri C, Granmo OC, Goodwin M. 2019. Multi-layer intrusion detection system with extra trees feature selection, extreme learning machine ensemble, and softmax aggregation. Eurasip J Inf Secur. 2019 (1). doi:10.1186/s13635-019-0098-y.

Shengle C, Webb GI, Liu L, Ma X. 2020. A novel selective naïve Bayes algorithm. Knowl-Based Syst. 192:105361. doi:10.1016/j.knosys.2019.105361.

Thornton HR, Delaney JA, Duthie GM, Dascombe BJ. 2017. Importance of various training-load measures in injury incidence of professional rugby league athletes. Int J Sports Physiol Perform. 12(6):819–824. doi:10.1123/ijspp.2016-0326.

Whitehead S, Till K, Jones B, Beggs C, Dalton-Barron N, Weaving D. 2021. The use of technical-tactical and physical performance indicators to classify between levels of match-play in elite rugby league. Sci Med Football. 5(2):0(0. doi:https://doi.org/10.1080/24733938.2020.1814492.

Whitehead S, Till K, Weaving D, Hunwicks R, Pacey R, Jones B. 2019. Whole, half and peak running demands during club and international youth rugby league match-play. Sci Med Football. 3(1):63–69. doi:10.1080/24733938.2018.1480058.

Whitehead S, Till K, Weaving D, Jones B. 2018. The use of microtechnology to quantify the peak match demands of the football codes: a systematic review. Sports Med. 48(11):2549–2575. doi:10.1007/s40279-018-0965-6.

Wilkens S. 2021. Sports prediction and betting models in the machine learning age: the case of tennis. J Sports Anal. 7(2):99–117. doi:10.3233/JSA-200463.

Williamson BD, Gilbert PB, Carone M, Simon N. 2021. Nonparametric variable importance assessment using machine learning techniques. Biometrics. 77(1):9–22. doi:10.1111/biom.13392.

Witten IH, Frank E, Hall MA. 2011. Data mining - practical machine learning tools and techniques. Morgan Kaufmann Publishers; pp. 1–665.

Woods CT, Robertson S, Collier NF, Swinbourne AL, Leicht AS. 2018. Transferring an analytical technique from ecology to the sport sciences 2. Sports Med. 48(3):725–732. doi:10.1007/s40279-017-0775-2.

Woods CT, Robertson S, Sinclair WH, Till K, Pearce L, Leicht AS. 2018. A comparison of game-play characteristics between elite youth and senior Australian National Rugby League competitions. J Sci Med Sport. 21(6):626–630. doi:10.1016/j.jsams.2017.10.003.