International Journal of Advanced Technology and Engineering Exploration, Vol 10(98)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2021.876300

Research Article

Optimal feature selection for cricket talent identification

Naveed Jeelani Khan, Gulfam Ahamad*, Nahida Reyaz and Mohd Naseem
Department of Computer Sciences, Baba Ghulam Shah Badshah University, Rajouri, Jammu and Kashmir, India

Received: 11-June-2022; Revised: 26-January-2023; Accepted: 27-January-2023

Abstract
Cricket talent identification (TiD) is a methodical process that aims to find the young athletes possessing a potential to excel in the cricket sport at an early age. The sports scientists have identified a set of twenty-eight parameters that determine the cricket TiD. In order to realize the objective of computational efficiency by reducing the feature set, we perform an optimal feature selection for cricket TiD using nine different feature selection techniques Viz. mutual information, information gain ratio, correlation, chi square, univariate root mean square error (RMSE), receiver operating characteristic (ROC) with decision tree classifier, reliefF, boruta and oneR. The individual results obtained from the feature selection techniques are provided along with the individual ranking. We aggregate the results using two different rank aggregation techniques namely average ranking aggregation and majority vote ranking aggregation. The aggregation results show a significant agreement between the two schemes. Fourteen out of twenty-eight features are selected using a threshold of 0.52– the value selected on recommendation of four different domain experts. 71.4% of the selected features are sport-centric and only 28.6% of the selected features are from the cognitive ability category. To the best of our knowledge, this is first such attempt to identify the talent in cricket using this methodology.

Keywords
Sports talent identification, Feature selection, Cricket talent identification, Applied decision sciences, Feature reduction.

1.Introduction
Talent identification (TiD) in sports is a process of identifying the potential future athletes. The TiD models are expected to predict the talent of potential athletes at an early age [1]. With the worldwide race between the countries to excel in sports, the employment of sports TiD models to assess and predict the athletic talent is increasing day by day [2, 3]. The term "talent identification" refers to the estimation of enthusiasts who are in some way interested in physical activities and may be prospective athletes for a certain sport. There are two primary ways used to identify sporting talent, namely natural and scientific [4]. A common strategy employed by coaches or experts is natural selection, in which the expert selects talent based only on his or her judgment or assessment, without the use of any scientific methods, tests, or assessments. But the scientific selection follows a rigorous approach. Its function is based on the essential characteristics that set the best athletes in each sport apart.

A computational model is used to process the results after various testing techniques are used to gather the data for the enthusiasts [5]. The coaches then use the findings to assess and pinpoint each person's potential talent. These methods help an athlete's innate potential to shine since they emphasize the areas that still require refinement. The scientific methods for identifying cricket talent have previously been developed by the researchers [6–9]. The TiD models predict the future talent of the candidates on the basis of a set of characteristics (variables or features). Since the talent is not one dimensional rather a multi-dimensional entity, the features are usually large in number. The features are usually from the following diverse categories like health, motor skills, functional skills, physiological skills, anthropometric skills, psychological skills, sociological skills, cultural structure, game intelligence skills, technical/tactical skills and genetics. Moreover, there exist no consensus on the standard feature set for the talent prediction as the characteristics of talent vary from one sport to another e.g., the features able to predict the swimming talent may not be the same for the football sport. Even in a single sport, the features vary because of the perception and domain

*Author for correspondence

knowledge of the experts or coaches. As the dimensions (features) of the data tend to increase, the data growth escalates as well [10].

This ever-growing abundance of data nowadays makes the data difficult to be comprehended. The large number of features in the dataset leads to a number of genuine problems like the curse of dimensionality, loss of explain-ability (Occam's razor), poor accuracy, time consuming, high computational cost, risk of overfitting and poor visualization experience. Feature engineering can help in overcoming the aforementioned issues. The feature selection is a crucial step in any predictive model including the machine learning models. It helps us to select a relevant and most important subset of features from the feature space and remove (or lessen the contributed impact of) irrelevant and redundant features, maintaining the accuracy of the model [11, 12]. The feature selection techniques are also being utilized to rank and weight the features of our dataset.

In cricket, the TiD programs are relatively new [13], although the field is emerging. To the best of our knowledge, [14] developed the first TiD model for cricket. Ahamad et al. [15] have identified 28 attributes that are sufficient to predict the potential talent of young enthusiasts. Selection of a lesser but relevant number of attributes of the cricket TiD model will certainly improve the computational accuracy of the final model. Our study aims to reduce this feature space of 28 features using various feature selection techniques, in order to achieve the computational excellence in the cricket TiD. We performed the experiment upon the dataset of the same study with as different as nine feature selection techniques. The final results are compared and evaluated using the majority vote and mean rank aggregation strategies.

Our work makes the following contributions:

1. Dataset Creation: We use the data collected by Ahamad et al. [15]. The dataset is small for some of the feature selection algorithms that are based on machine learning. In order to overcome this hurdle, we created an R based data synthesizer tool. It takes the original data as seed and replicates it intelligently, keeping the important characteristics of the data (like upper bound, lower bound, mean, deviation, etc.) in consideration. However, the dataset remains representative of the original dataset only. The dataset contains 1726 entries for the male gender.
2. Application of feature selection techniques in cricket TiD: To the best of our knowledge, this is a first attempt in the domain of cricket TiD. We came across no study from literature attempting the same.
3. Feature reduction: The results claim that out of the 28 features, 14 features contribute the most and are indicative for the purpose of cricket TiD.

The rest of this article is organized as follows. Section 2 provides the background knowledge. Section 3 discusses the employed feature selection methods briefly and provides the individual results. Section 4 discusses the experimental results. Finally, Section 5 concludes the paper.

2.Literature review

Sports talent is being examined since a long time ago [16]. Sports TiD is a process to identify the potential superiorly talented athletes at a young age. Prompt recognition of talent saves a lot of resources like time and effort. The developed countries like United States, China, Russia, Germany and United Kingdom deploy the scientific models for talent discovery. These methods also help to detect the early inclination of a child towards different sports. Moreover, the selected talent can also be enhanced by deploying such models. The scientific modelling involves the identification of key parameters that are decisive for the TiD. Each sport has its own different set of contributing TiD parameters. The data for respective parameters is collected using various testing mechanisms, and the findings are processed using a scientific model, which is then utilized by specialists to analyze and discover the potential talent in the candidates. We, in our latest published work [5] provide a landscape of sports TiD techniques. Taha et al. [17] discusses the importance of utilization of the game specific variables in TiD process for identifying the optimal performers in Archery. The authors identify the candidates into high performance archers and low performance archers. Noori and Sadeghi [18] studied and provided a fuzzy based TiD model for the volleyball sports. The model works on the following parameters: height, lower extremity length, upper extremity length, palm-size, shoulder width, agility, endurance, strength, flexibility, power, self-confidence, motivation, focus, goal directed behavior, imagination, aerobic endurance, anaerobic endurance, specific endurance, lactic acid tolerance, spike, serve, forearm pass and overhead pass. The model generates an output of crisp values from which the final TiD is performed. The results show the percentage of the

following categories against each participant: unmatched, semi-matched, matched, brilliant and rare. Kusnanik et al. [19] developed a TiD model for the sprinter. The model is based on the discriminant factor analysis. The parameters used are standing height, sitting height, body mass, leg length, speed, agility, power and endurance. Rozi et al. [20] proposed an identification model for swimming. Two similarity models are developed for male and female athletes. The model uses the following parameters: Weight, height, arm span, leg length, palm width, foot length, sit and reach, pull up, sit up, standing broad jump, kick board 15 meters, 15 meters swimming and 15 minutes swimming. Similarly, many other authors like Dwivedi et al. [21], Matrasid et al. [22], Huang et al. [23], etc. have developed the TiD models for different sports based on a wide variety of features.

The efficient solution for any computational problem demands employment of the most relevant features from the identified feature set for the model and the same goes true for the sports TiD models as well. The identification of irrelevant or redundant features is a challenging task. Feature selection methods are used to choose a relevant subset of parameters that is sufficient to symbolize the original data [24]. The model with a reduced feature set will perform better at the time complexity. It is impractical to use n number of features, if the same accuracy (approximately) can be attained by only n-k features where $n>1$, $k<n$ and $k>0$. *Table 1* shows the parameters along with the measurement descriptions, employed by Ahamad et al. [15] for cricket TiD.

Removing the noise from data reduces the number of computations to be performed and hence minimizes the computational complexity. Moreover, it reduces the overfitting. One such technique to remove noisy (or redundant) data is feature selection. It is a pre-processing procedure that selects a sub-set of features from an original feature set. The feasibility of the selected subset is evaluated using certain criteria [25]. Nominates the four steps of the feature selection process as subset generation, subset evaluation, stopping criterion and validation. In the first step subsets are generated on the basis of a search strategy [26, 27].

The evaluation of the same is performed against a criterion in order to select a best subset that outperforms in the group. Both processes continue until a stopping condition is satisfied. After the best subset is generated, it needs to be validated using some validation mechanism. There are three broad types of feature selection procedures namely filter wrapper and embedded methods [28, 29].

Filter methods use the statistical characteristics of data, like correlation, to select a feature subset. These methods do not use any mining/learning algorithm to perform the subset selection. The modelling/classifier algorithm plays no role in the feature subset selection. The methods are pretty straight forward and computationally inexpensive. The features are ranked and selected on the basis of the statistical scores for each feature. Pearson correlation, mutual information, chi square and ReliefF are some of the prominent examples of filter methods.

Wrapper methods use particular algorithms to evaluate the selected features. Unlike the filter-based methods, wrapper methods take into account the effect of selected feature subsets on the algorithms used. These methods typically follow a two-step process wherein the feature subsets are generated using some search strategy and the same are evaluated using a particular algorithm. The process continues until the desired output is generated. Wrapper based approaches are more accurate than the filter-based approaches however they need more computational power than the later [30]. Genetic algorithm, ant colony optimization and swarm optimization-based methods are typical examples of the wrapper feature selection.

Embedded methods reap the benefits of both filter-based as well as wrapper-based methods. These methods are usually implemented by the learning algorithms that include the built-in feature selection option. The machine learning algorithms take advantage of the feature selection procedures during the process of classification eliminating the need for a separate learning process and hence reducing the time complexity. Lasso and ridge regression are the examples of embedded feature selection.

Ahamad et al. [31] proposed an ordered weighted averaging aggregation (OWA) operator based cricket TiD model. This model is the first cricket TiD of its kind [14]. The end results of this model when compared with the manual TiD system proved to be justifiable and promising. This model takes twenty-eight input features to identify the talent. The features along with the corresponding details are mentioned in *Table 1*. Like any other sport, these features mainly belong to the categories of physical, anthropometric and physiological factors [32]. Our study aims to

reduce the number of parameters by only selecting the features having a significant impact on the outcome.

The key findings from the literature are:
a)A stagnancy at the computational level is perceived as the new tide models continue to use the same conventional procedures. Only a few studies like [15, 21, 33, 34] were found that used the advanced computational techniques. It was observed that a small amount of research exists for the pure computational part of the TiD process.

b) Despite the fact that female participation in sports is increasing day by day, a significantly lesser number of TiD studies (among reviewed articles) were explicitly conducted for the female gender. Only 4% of individual studies for females were found during the survey, which is significantly low than the figure found for the male gender, i.e., 38%. The same issue has been detected by many other studies [35].

c) The reviewed articles originated from sixteen different countries all over the world, namely Croatia, China, Iran, UK, India, Japan, Turkey, Australia, Serbia, Indonesia, Malaysia, Slovenia, South Africa, The Netherlands, Israel and Belgium. Most of the Articles about TiD in sports were from Croatia, i.e., 17%.

d) In the domain of Cricket TiD, no study was found that have attempted to reduce the feature set using the feature selection methods.

Table 1 Identified features for cricket TiD

S. No.	Parameter name	Corresponding test name	Category	Description
1	Speed (T1)	speed test	physical/ motor ability	time taken to run the 30-meter distance
2	Agility (T2)	illion's agility test	physical/ motor ability	test the running agility in a 10 × 5 m setup with 4 cones at 3.3 m apart.
3	Endurance (T3)	step up and down test	physical/ motor ability	calculate the heartrate in bpm after stepping up and down on a bench for 3 minutes.
4	Stress (T4)	sports competitive anxiety test	cognitive ability tests	based on the sport's competitive anxiety test quiz, record the scores. more the score, more the stress.
5	Self-Motivation (T5)	self-motivation test quiz	cognitive ability tests	the results are calculated on the basis of scores recorded from a self-motivation quiz
6	Upper Body Strength (T6)	push up test	physical/ motor ability	number of push-ups without losing the normal form
7	Lower Body Power (T7)	hop run test	physical/ motor ability	average time of the left and right leg - 25 m hop
8	Reaction Time (T8)	ruler catching test	cognitive ability tests	distance between base of the ruler and the tip of thumb at which it has been caught
9	Flexibility (T9)	sit and reach test	physical/ motor ability	distance reached by the hand in the position prescribed by the sit and reach test
10	Fatigue Index (T10)	running based anaerobic sprit (RAST) test	physical/ motor ability	power is calculated using the weight, distance run and time of the candidate. then fatigue index is calculated from maximum and minimum power and total time taken for the sprints.
11	Bowler Accuracy (T11)	bowler accuracy test	physical/ motor ability	number of times a target is hit out of 10.
12	Throw Catching Accuracy (T12)	throw catching accuracy test	physical/ motor ability	number of times the candidate's ball gets close to the gloves of wicket keeper out of 20.
13	Under Arm Throw Accuracy (T13)	under arm through accuracy test	physical/ motor ability	number of times ball hits the stumps for an underarm throw out of 20.
14	Catching Ability (T14)	catching ability test	physical/ motor ability	number of catches caught (from 30-yard distance) out of 10.
15	Ground Fielding (T15)	assessment of clean pick-ups.	physical/ motor ability	number of clean pick-up returns to the wicket keeper in the distance range of 10 yard
16	Vo2 Max (T16)	maximum oxygen up taken test	anthropometric tests	maximum oxygen taken

S. No.	Parameter name	Corresponding test name	Category	Description
17	Body Mass Index (T17)	weight/(height)2	anthropometric tests	$BMI = Weight/(height)^2$
18	Hand Eye Coordination (T18)	catching and throwing the ball in cyclic order with hands.	cognitive ability tests	count of catches of a tennis ball when thrown to a wall with the left hand and caught with the right hand and vice versa.
19	Creativity (T19)	creativity test quiz	cognitive ability tests	score (0-100) as recorded from the creativity test quiz
20	Decision Making (T20)	decision making ability test quiz	cognitive ability tests	score (0-100) as recorded from the decision-making ability test quiz
21	Self-Control And Self-Monitoring (T21)	self-control and self-monitoring test quiz	cognitive ability tests	score (0-100) as recorded from the self-control and self-monitoring test quiz
22	Will Power (T22)	will power test quiz	cognitive ability tests	score (0-100) as recorded from the will power test quiz
23	Self-Confidence (T23)	self-confidence test quiz	cognitive ability tests	score (0-100) as recorded from the self-confidence test quiz
24	Integrity And Work Ethic (T24)	integrity and work ethic ability test quiz	cognitive ability tests	score (0-100) as recorded from the integrity and work ethic ability test quiz
25	Shoulder Flexibility (T25)	static flexibility test based on physical action	physical/ motor ability	the shoulder measurements are subtracted from the distance between thumb tips (in the stature as prescribed by the specific test)
26	Balance (T26)	beam test for balance	physical/ motor ability	the scores are taken on basis of how the enthusiast walks on a beam.
27	Balance In Static Form (T27)	standing stork test	physical/ motor ability	the maximum time is recorded for which an enthusiast can hold on with the sole on one feet against the side knee cap of other (for both legs)
28	Concentration And Focus Monitoring (T28)	concentration and focus skill test quiz	cognitive ability tests	score (0-100) as recorded from the concentration and focus skill test quiz

3.Methods

The feature selection techniques help in improving the performance of models by selecting the most informative features. In this section the methodology is discussed along with the working.

3.1Working

We calculate the feature ranking by employing the feature selection techniques discussed in section 3 for our dataset. For each feature selection technique to follow, an individual ranking table with results is provided in this section. As shown in *Figure 1*, the dataset was considered as the input to the mentioned feature selection techniques. Each technique provides us with a reduced subset of the features of the original 28 features that are optimal. Each technique uses a different methodology to calculate the ranking and importance of the features. The results need to be aggregated for the interpretability. We aggregate the results using two different rank aggregation techniques, namely average ranking aggregation and majority vote ranking aggregation. The aggregation results show a significant agreement between the two schemes. In order to choose a threshold for selecting the number of features from the top category, we normalize the mean ranking. Fourteen out of twenty-eight features are selected using a threshold of 0.52– the value selected on recommendation of four different domain experts. The ranking values determine the feature importance. Each method has its own way to calculate the importance and hence the number scale differs but for the different feature selection methods, the ranking numbers (first, second, third, etc.) remain same. In order to avoid the problems posed by the imbalanced datasets, we used the k (k=10) fold cross validation to partition the dataset into subsets for the training purpose of machine learning based methods. In this method, all of the data is used for training and testing purposes using the moving bucket strategy.

3.2Experimental setup

The rankings were calculated using Python and R on a 64-bit system with Intel® Core™ i7 10510U Processor, NVIDIA® GeForce® MX350 Graphics Processor, 8GB 2666MHz of DDR4 RAM Memory, and 512GB PCI Express Gen 3 NVMe SSD Storage.

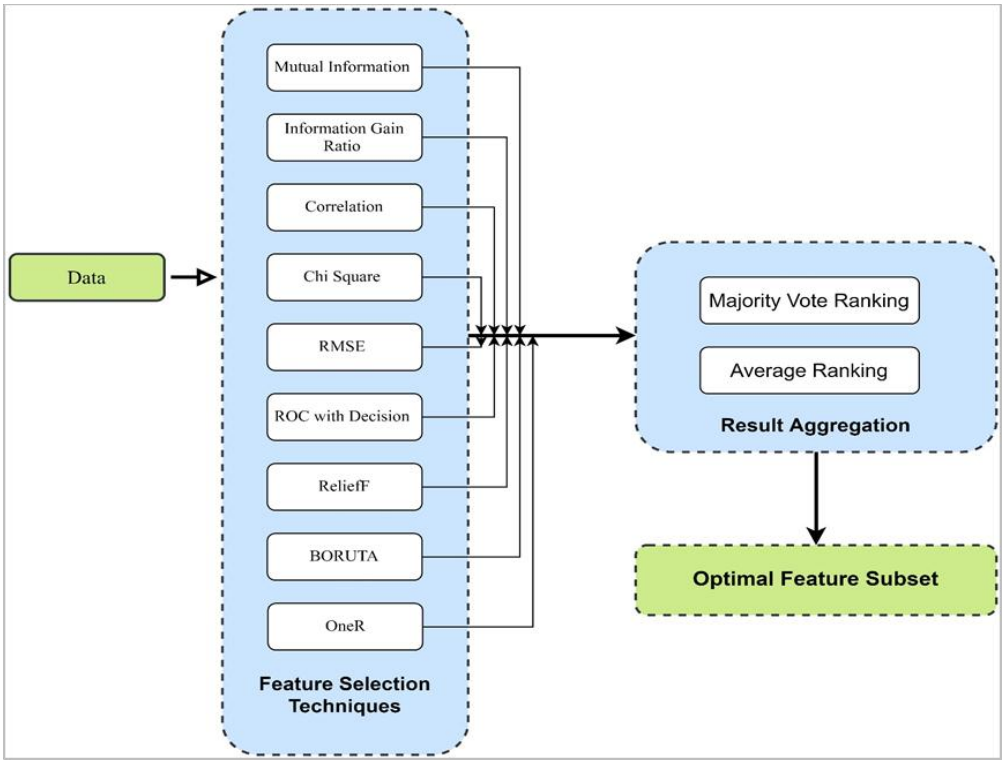

Figure 1 Feature selection methodology

3.3Selection procedure

The selection of feature selection approaches depends on several factors. We considered the following factors for selection of feature selection approaches for our study:

Type of data: Different feature selection approaches may be more suitable for different types of data. Since our data is numbed and the input variables are continuous as well as discrete in nature, therefore the relevant approaches were selected.

Type of problem: The problem type and the goals of the study determine the type of feature selection approaches that are most appropriate. In our case the problem is classification type and hence the selected methods were relevant.

Algorithm: The algorithms being used also influence the selection of feature selection approach. For example, we used the decision tree algorithm, then an importance-based feature selection such as ReliefF was more appropriate. Computational cost and time constraint: Some feature selection approaches, such as wrapper methods, are computationally expensive and might not be suitable when working with large datasets. We chose the computationally efficient methods. We left out the methods that required high computational costs. Moreover, some feature selection approaches are time-consuming as well, so the time constraint should be considered when selecting an approach.

The size of the dataset: We considered the size of our dataset as some feature selection approaches are not suitable for large datasets so we had to balance out between the size of the dataset and feature selection methods.

After surveying the literature and careful evaluation of the pros and cons of different feature selection approaches, we chose the following ones that fit best for our purpose:

3.3.1Mutual information

Mutual information is a measure of the amount of information shared between two random variables [36]. In feature selection, mutual information is used to quantify the relationship between a feature and the target variable. Features with high mutual information are considered to be more informative and relevant for the task at hand, and are more likely to be selected. Zero mutual information value indicates that the two features are independent. This method provides an intuitive interpretation. Moreover, it can detect the non-linear relationships between the parameters. The Mutual information is given by Equation 1 [37]:

$$I(x;y) = \sum_{i=1}^{n}\sum_{j=1}^{n} p(x(i),y(j).\log\left(\frac{p(x(i),y(j))}{p(x(i).p(y(j))}\right) \quad (1)$$

Where x and y are two random features,
$p(x,y)$ is the joint probability density function of x and y,
$p(x)$ is the marginal probalilty density function of x,
And $p(y)$ is the marginal probability density function of y.

3.3.2Information gain ratio

Gain ratio is an entropy-based feature selection method. It adds a refinement to the information gain in order to overcome the large number of trivial partitioning for discrete features. This method improves the performance when applied to the parameters with large number of unique values. Since our dataset had same (for catching accuracy, throw accuracy, etc.), this method is found to be useful in our context. The process begins by calculating the entropy of the target variable. Entropy is a measure of the amount of disorder or randomness in the target variable. Next, the entropy of each feature is calculated and the information gain of each feature is determined by subtracting the entropy of the feature from the entropy of the target variable. The feature with the highest information gain is selected as the most informative feature. This process is repeated for the remaining features until a desired number of features have been selected. The selected features are then used as input to a model for prediction. Information gain-based feature selection can also help to improve the overall performance of a model by reducing overfitting, increasing the interpretability of the model and reducing the computational cost of prediction.

The information gain ratio is given by Equation 2 [38, 39]:

$$IGainRatio(x) = \frac{InfoGain(x)}{H(\hat{p}_x)} \quad (2)$$

Where $InfoGain(x) = H(\hat{p}) - \sum_{i=1}^{S_k}\frac{n_i}{n}H(\hat{p}_{ik})$

and $H(\hat{p}_x) = -\sum_{i=1}^{S_k}\frac{n_i}{n}\log_2\frac{n_i}{n}$

3.3.3Correlation

Correlation is a classic method of measuring the dependency of one variable on another. Correlation based feature selection is a method used in data mining to select the most relevant features for a given dataset. The goal is to select a subset of features that are highly correlated with the target variable and are not highly correlated with each other. This technique begins by calculating the correlation between each feature and the target variable. The feature with the highest correlation with the target variable is selected as the most relevant feature. Next, the correlation between each remaining feature and the selected feature is calculated. If two features are highly correlated with each other, one of them is removed from the dataset. This process is repeated until a desired number of features have been selected. The selected features are then used as input to the model prediction. This method is useful in eliminating irrelevant features and the computational cost of this method is low in comparison to other methods. For features X and Y, the correlation is calculated as given in Equation 3 [40]:

$$r_{xy} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{(x_i-\bar{x})^2}.\sqrt{(y_i-\bar{y})^2}} \quad (3)$$

Where x_i and y_i are the corresponding values of x and y in the sample,
$\bar{x}$ and $\bar{y}$ are mean values of x and y.

3.3.4Chi square

Chi Square is a famous hypothesis testing method. It is used to study the relationship between two entities. Chi Square feature selection selects the features with high dependency on the response feature. It is based on the Chi Square statistical test, which measures the association between two categorical variables. The process calculates the Chi Square value of each feature with respect to the target variable. The Chi Square value represents the degree of association between the feature and the target variable. The higher the Chi Square value, the greater the association between the feature and the target variable. The feature with the highest Chi Square value is selected as the most relevant feature. This process is repeated for the remaining features until a desired number of features have been selected. Smaller chi value means the intensity of dependence between the two features is low and hence close to the independence. This method follows simple univariate selection strategy. It is used where the data can be clubbed into the frequencies.

Chi Square [41] is calculated as given in Equation 4:

$$\chi_c^2 = \sum\frac{(O_i-E_i)^2}{E_i} \quad (4)$$

Where c = degrees of freedom, O_i = observed values of the sample and E_i = expected values of the sample

3.3.5Univariate RMSE

Root mean square error (RMSE) is another useful method to observe the deviation between the variables. It is also known as Root mean square deviation (RMSD). The Univariate RMSE feature selection ranks the features on basis of the dependency of features on the target feature. RMSE based feature selection is used to select the most relevant features for a given dataset. It is based on the

RMSE metric, which measures the difference between predicted and actual values. The technique begins by with all the features in the dataset. The model's performance is evaluated using the RMSE metric. Next, each feature is removed from the dataset one by one and the model is retrained and evaluated using the RMSE metric. The feature that results in the lowest RMSE value is selected as the most relevant feature. This process is repeated for the remaining features until a desired number of features have been selected. The selected features are then used as input to a model for prediction.

RMSE [42] is calculated as given in Equation 5:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \widehat{x_i})^2}{N}} \quad (5)$$

Where x_i = the actual observation, $\widehat{x_i}$ = the estimated observation and N = the number of observations.

3.3.6Receiver operating characteristic (ROC) with decision tree classifier

Receiver operating characteristic (ROC) curve is a graphical plot obtained by plotting the true positive rate vs the false positive rate of a classifier. The area under the ROC curve is called AUC. ROC is being used to evaluate the supervised machine learning models [43]. It shows the tradeoff between sensitivity and specificity. The properties of ROC make it a viable method for feature selection. ROC based feature selection is based on the graphical representation of the performance of a binary classification model. The technique uses all the features in the dataset. The model's performance is evaluated using the ROC curve, which plots the true positive rate against the false positive rate. Next, each feature is removed from the dataset one by one and the model is retrained and evaluated using the ROC curve. The feature that results in the highest area under the curve (AUC) value is selected as the most relevant feature. This process is repeated for the remaining features until a desired number of features have been selected. ROC based feature selection is commonly used in binary classification tasks, where the goal is to distinguish between two classes. It is a useful method for selecting the most relevant features from a large dataset, and it can improve the performance of a model by increasing the discrimination power of the model and by increasing the interpretability of the model. The True positive and false positive rates are calculated as given in Equation 6 and Equation 7 respectively:

$$True\ Positive\ Rate\ (TPR) = \frac{TP}{TP+FN} \quad (6)$$

$$False\ Positive\ Rate\ (FPR) = \frac{FP}{FP+TN} \quad (7)$$

Where TP = true positive, FN = false negative, FP = false positive and TN = true negative

3.3.7ReliefF

The ReliefF algorithm is a more robust version of the original relief feature selection method for it can deal with the incomplete and noisy data [44]. Moreover, it is not limited to the one or two class problems. Urbanowicz et al. [45] refer to the reliefF as the best known variant of relief method. ReliefF feature selection based on the ReliefF algorithm, which uses a nearest-neighbor approach to identify the most relevant features for a given dataset. It begins by initializing the feature weights to zero for all features. Next, for each sample in the dataset, the algorithm identifies the nearest hit and nearest miss for the sample. A hit is a sample that belongs to the same class as the current sample, and a miss is a sample that belongs to a different class. The feature weights are then updated based on the difference between the current sample and its nearest hit and miss. The feature with the highest weight is selected as the most relevant feature. This process is repeated for the remaining features until a desired number of features have been selected. ReliefF based feature selection is commonly used in classification tasks, where the goal is to distinguish between multiple classes.

ReliefF calculates the feature weights [40] as given in Equation 8 as follows:

$$J_{rx}(y,x) = \frac{GS_x}{(1-S_y)S_y} \quad (8)$$

Where $S_x = \sum_{i=1}^{k} P(x_i)^2$; $S_y = \sum_{j=1}^{M_y} P(y_j)^2$

And

$$G = \sum_j P(y_j)\left(1 - P(y_j)\right) - \sum_{i=1}^{k}\left\{\frac{P(x_i)^2}{S_x}\sum_j P(y_j|\, x_i)(1 - P(y_j|\, x_i))\right\}$$

3.3.8Boruta

Boruta [46] is a random forest based feature selection algorithm. All the variables are replicated and the replicated values are randomized, these are called shadow features. Random forest is executed for the same and the feature importance are calculated for each run. If the value is higher than the maximal importance of all randomized variables, the feature is deemed as important. It employs maximal importance of the random attributes (MIRA) statistical test. Boruta can deal with classification as well as regression problems. Moreover, it can efficiently detect the non-linear relationships between the variables. Boruta feature selection is based on the

boruta algorithm, which uses random forests to identify the most relevant features for a given dataset. It begins by creating a set of shadow features, which are random permutations of the original features. Next, a random forest model is trained using both the original and shadow features. The feature importance values are calculated for each feature and are used to rank the features. The features that have a higher importance value than the shadow features are considered as relevant features. After that, the algorithm iteratively removes the features with the lowest importance values, retrains the random forest model and calculates new importance values. The process stops when all features have been removed or when a stopping criterion is met. The remaining features are considered as the most relevant features. Boruta based feature selection is commonly used in classification and regression tasks.

3.3.9OneR

OneR is a rule-based learning algorithm, it constructs a classification while creating a one level decision tree. One individual attribute is tested at a time. It creates a frequency table and identifies the important predictors in the dataset. Based on the important predictors, a classification rule is constructed for the classification purpose [47]. OneR feature selection is based on the OneR algorithm, which is a simple and efficient method for feature selection. The process begins by training a simple rule-based classifier, called OneR, on each feature individually. OneR classifier is based on the idea of finding the feature that has the lowest error rate. Next, the error rate of each feature's classifier is calculated. The feature with the lowest error rate is selected as the most relevant feature. This process is a simple but effective method for feature selection, especially when the dataset is large and has many features. It is commonly used in classification tasks, where the goal is to distinguish between multiple classes. OneR based feature selection can improve the performance of a machine learning model by reducing overfitting and increasing the interpretability of the model, making it easier to understand the relationships between the features and the target variable.

4.Results

This section provides the individual results of the feature selection techniques as discussed in the section 3.

4.1Dataset description

Source: The already published data is taken against the 29 (including the label) parameters from the various cricket enthusiasts under the guidance of coaches by [15]. The data was collected using various tools like timer, stopwatch, ruler, tennis ball, beam and relevant psychological tests as shown in *Table 1*. For the physical tests, an average of 3 trials was taken into consideration. The Screenshot of the dataset is shown in *Figure 2*.

Data synthesis: In order to overcome the challenge of data scarcity for some of the feature section methods, we created an R based data synthesizer tool. In order to carry out the data synthesis process in a responsible and accurate way and ensuring that the resulting dataset is reliable and suitable for analysis, we kept the following things in consideration:

Data quality: Before synthesizing data, we made sure that the data used is of high quality and suitable for the analysis. This included checking for missing values, outliers, and other potential issues with the data.

Data compatibility: we made sure that the data used in the synthesis is compatible and can be combined in a meaningful way. This includes checking for the same data format, units, and time frames.

Data bias: When synthesizing data, we considered potential biases that may be introduced. This included checking for any differences in data collection methods, sampling techniques, and measurement errors.

Validation: We validated the synthesized data by comparing it with the original data by using cross-validation.

Our R program takes the original seed data and replicates it intelligently, keeping the important characteristics of the data (like upper bound, lower bound, mean, deviation, etc.) in consideration. It checks for the maximum and minimum values of all the parameters. Keeping in consideration the data distribution, it generates new values between the limits. The dataset remains representative of the original dataset only. The new data points contain no individual significance.

Number of Observations: 1726 entries.
Type of Problem: Classification
Class Distribution: 1131 entries for non-talented and 594 for talented enthusiasts.
Gender of Subjects: Male.
Location: New Delhi, India.
Variables/Parameters: Twenty-eight input variables and one output variable (i.e., label) viz. speed (T1), agility (T2), endurance (T3), stress (T4), self-motivation (T5), upper body strength (T6), lower body power (T7), reaction time (T8), flexibility (T9), fatigue index (T10), bowler accuracy (T11), throw catching accuracy (T12), under arm throw accuracy

(T13), catching ability (T14), ground fielding (T15), vo2 max (T16), body mass index (T17), hand eye coordination (T18), creativity (T19), decision making (T20), self-control and self-monitoring (T21), will power (T22), self-confidence (T23), integrity and work ethic (T24), shoulder flexibility (T25), balance (T26), balance in static form (T27), concentration and focus monitoring (T28) and one label variable. The output label contains the binary values 0 or 1. 0 stands for non-talented and 1 stands for the 'potentially talented' candidate. The labelling is manually performed by the respective coaches.

Data Cleaning/Preprocessing: In order to make the dataset accurate and reliable for analysis. The following steps were followed:

Duplicate entries: We compared the values of each column in the dataset and removed any rows that had the same values.

Missing Values: Some of the columns in the dataset were missing, we cleaned the data from such entries as well.

Outliers: We identified and removed the outliers in our dataset by calculating the Z-score of each data point and removed any points that were above the threshold of 3.

Normalization: Since the 28 features used are altogether different in nature and have different range for values, in order to get them on one common scale, the data values were normalized using the 0 to 1 Minmax Scaler.

T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	T16	T17	T18	T19	T20	T21	T22	T23	T24	T25	T26	T27	T28	label
0.54	0.53	0.5	0.32	0.36	0.68	0.33	0.47	0.3	0.8	0.4	0.3	0.4	0.5	0.6	0.6	0.29	0.64	0.47	0.54	0.35	0.65	0.67	0.63	0.31	0.5	0.8	0.6	0
0.49	0.59	0.47	0.29	0.23	0.58	0.28	0.38	0.57	0.6	0.2	0.4	0.3	0.5	0.6	0.64	0.3	0.62	0.42	0.47	0.62	0.65	0.34	0.73	0.4	0.3	0.73	0.45	0
0.31	0.5	0.4	0.46	0.7	0.67	0.09	0.32	0.68	0.6	0.5	0.6	0.5	0.6	0.9	0.97	0.79	0.79	0.89	0.71	0.76	0.92	0.94	0.81	0.3	0.8	0.96	0.85	1
0.36	0.54	0.45	0.44	0.35	0.67	0.29	0.46	0.41	0.7	0.4	0.4	0.2	0.5	0.7	0.56	0.57	0.57	0.54	0.65	0.48	0.87	0.09	0.65	0.36	0.5	0.37	0.62	0
0.44	0.48	0.63	0.61	0.64	0.59	0.2	0.32	0.68	0.7	0.4	0.3	0.2	0.5	0.6	0.35	0.36	0.52	0.76	0.64	0.6	0.49	0.2	0.69	0.3	0.5	0.59	0.72	0
0.48	0.7	0.61	0.52	0.56	0.42	0.31	0.39	0.63	0.5	0.4	0.5	0.4	0.6	0.6	0.66	0.49	0.68	0.48	0.47	0.68	0.86	0.74	0.74	0.43	0.7	0.83	0.42	1
0.46	0.69	0.59	0.42	0.39	0.56	0.28	0.39	0.46	0.5	0.3	0.4	0.5	0.5	0.3	0.63	0.25	0.66	0.74	0.68	0.36	0.8	0.32	0.59	0.38	0.5	0.36	0.7	0
0.36	0.44	0.63	0.44	0.63	0.43	0.41	0.34	0.31	0.8	0.4	0.3	0.3	0.7	0.3	0.73	0.52	0.59	0.56	0.47	0.28	0.75	0.74	0.55	0.48	0.3	0.73	0.76	0
0.54	0.48	0.58	0.39	0.24	0.63	0.28	0.34	0.37	0.8	0.3	0.4	0.4	0.3	0.3	0.53	0.6	0.71	0.61	0.55	0.54	0.57	0.64	0.39	0.45	0.5	0.81	0.73	0
0.42	0.5	0.55	0.43	0.24	0.6	0.4	0.37	0.59	0.7	0.6	0.4	0.5	0.3	0.8	0.61	0.34	0.6	0.46	0.62	0.32	0.52	0.53	0.75	0.4	0.5	0.52	0.81	0
0.36	0.47	0.47	0.4	0.39	0.59	0.41	0.46	0.65	0.8	0.3	0.3	0.3	0.6	0.7	0.49	0.74	0.62	0.41	0.77	0.42	0.76	0.19	0.6	0.5	0.4	0.33	0.56	0
0.5	0.5	0.66	0.66	0.32	0.47	0.4	0.42	0.54	0.6	0.6	0.4	0.2	0.4	0.2	0.72	0.23	0.65	0.48	0.54	0.31	0.71	0.75	0.56	0.32	0.7	0.42	0.77	0
0.38	0.59	0.46	0.34	0.57	0.65	0.27	0.28	0.4	0.7	0.3	0.3	0.3	0.4	0.5	0.69	0.32	0.56	0.75	0.71	0.59	0.44	0.41	0.77	0.4	0.4	0.39	0.5	0
0.46	0.67	0.54	0.41	0.53	0.55	0.31	0.34	0.49	0.6	0.2	0.4	0.4	0.5	0.5	0.73	0.5	0.67	0.39	0.8	0.64	0.87	0.8	0.78	0.46	0.6	0.58	0.61	1
0.41	0.45	0.52	0.39	0.3	0.47	0.3	0.44	0.67	0.7	0.4	0.3	0.3	0.5	0.7	0.39	0.65	0.57	0.42	0.4	0.71	0.59	0.42	0.55	0.48	0.7	0.64	0.71	0
0.43	0.59	0.44	0.47	0.52	0.42	0.4	0.39	0.3	0.5	0.5	0.4	0.3	0.2	0.2	0.67	0.4	0.71	0.65	0.48	0.58	0.91	0.27	0.67	0.48	0.7	0.6	0.65	0
0.5	0.72	0.47	0.46	0.41	0.49	0.25	0.39	0.63	0.7	0.5	0.5	0.2	0.6	0.5	0.55	0.46	0.64	0.74	0.61	0.78	0.75	0.78	0.61	0.32	0.7	0.72	0.53	1
0.39	0.44	0.47	0.46	0.36	0.49	0.38	0.3	0.51	0.6	0.5	0.5	0.4	0.3	0.2	0.42	0.74	0.57	0.57	0.72	0.75	0.62	0.76	0.64	0.48	0.3	0.83	0.45	0
0.35	0.56	0.49	0.48	0.75	0.67	0.13	0.33	0.52	0.4	0.5	0.8	0.8	0.5	0.7	0.96	0.49	0.68	0.56	0.73	0.41	0.52	0.86	0.62	0.18	0.6	0.95	0.55	1
0.5	0.61	0.62	0.54	0.26	0.7	0.23	0.39	0.38	0.7	0.5	0.4	0.3	0.7	0.4	0.63	0.59	0.54	0.78	0.52	0.52	0.78	0.43	0.48	0.37	0.7	0.53	0.74	0
0.43	0.51	0.65	0.67	0.44	0.7	0.39	0.47	0.53	0.6	0.5	0.5	0.3	0.7	0.2	0.75	0.3	0.63	0.43	0.4	0.77	0.54	0.44	0.75	0.53	0.4	0.67	0.39	0
0.4	0.59	0.5	0.3	0.44	0.45	0.23	0.44	0.38	0.7	0.3	0.4	0.4	0.3	0.2	0.45	0.51	0.63	0.8	0.43	0.31	0.45	0.27	0.46	0.49	0.6	0.58	0.64	0
0.51	0.51	0.52	0.67	0.31	0.59	0.19	0.43	0.64	0.8	0.5	0.4	0.5	0.7	0.8	0.36	0.26	0.63	0.47	0.7	0.49	0.51	0.19	0.43	0.41	0.5	0.72	0.64	0
0.37	0.26	0.56	0.49	0.6	0.66	0.32	0.26	0.64	0.2	0.9	0.6	0.6	0.6	0.7	0.61	0.72	0.66	0.94	0.89	0.58	0.56	0.59	0.86	0.3	0.7	0.78	0.85	1
0.43	0.38	0.58	0.36	0.57	0.76	0.19	0.37	0.58	0.4	0.9	0.7	0.6	0.8	0.9	0.72	0.45	0.8	0.73	0.87	0.53	0.54	0.63	0.73	0.2	0.9	0.47	0.93	1
0.43	0.38	0.6	0.52	0.73	0.7	0.18	0.23	0.69	0.4	0.7	0.7	0.8	0.7	0.6	0.88	0.81	0.81	0.8	0.97	0.5	0.45	0.88	0.57	0.36	0.6	0.5	0.92	1
0.4	0.6	0.51	0.32	0.31	0.51	0.25	0.27	0.54	0.6	0.6	0.4	0.5	0.3	0.7	0.66	0.23	0.73	0.55	0.4	0.62	0.59	0.33	0.63	0.44	0.7	0.38	0.38	0
0.42	0.43	0.6	0.37	0.33	0.44	0.38	0.41	0.35	0.5	0.2	0.4	0.4	0.6	0.7	0.55	0.58	0.61	0.43	0.47	0.59	0.68	0.49	0.71	0.49	0.5	0.66	0.4	0
0.5	0.63	0.59	0.59	0.62	0.47	0.22	0.45	0.48	0.5	0.4	0.3	0.2	0.7	0.8	0.79	0.75	0.71	0.47	0.51	0.33	0.4	0.22	0.58	0.41	0.6	0.37	0.78	0
0.54	0.41	0.55	0.65	0.64	0.46	0.2	0.4	0.46	0.6	0.3	0.4	0.4	0.4	0.7	0.46	0.6	0.69	0.45	0.65	0.62	0.93	0.8	0.58	0.46	0.4	0.32	0.78	1
0.51	0.52	0.63	0.64	0.24	0.6	0.23	0.27	0.54	0.5	0.4	0.4	0.3	0.4	0.7	0.42	0.28	0.71	0.68	0.48	0.44	0.47	0.56	0.77	0.3	0.6	0.65	0.6	0
0.53	0.47	0.52	0.36	0.62	0.61	0.21	0.41	0.36	0.7	0.4	0.4	0.3	0.5	0.4	0.77	0.24	0.57	0.47	0.59	0.48	0.71	0.82	0.38	0.38	0.6	0.48	0.42	0
0.43	0.48	0.46	0.31	0.64	0.69	0.23	0.47	0.42	0.8	0.4	0.3	0.3	0.5	0.5	0.36	0.77	0.57	0.39	0.81	0.37	0.75	0.37	0.69	0.34	0.5	0.78	0.6	0
0.41	0.39	0.35	0.32	0.41	0.82	0.14	0.18	0.4	0.2	0.6	0.9	0.9	0.7	0.6	0.52	0.93	0.72	0.88	0.5	0.62	0.63	0.64	0.67	0.3	0.7	0.9	0.85	1
0.4	0.49	0.62	0.44	0.3	0.63	0.38	0.35	0.42	0.6	0.3	0.4	0.2	0.7	0.3	0.77	0.2	0.54	0.4	0.53	0.44	0.47	0.05	0.54	0.53	0.4	0.79	0.47	0

Figure 2 Screenshot of the cricket TiD dataset

4.2Feature selection results

In order to calculate the mutual information scores, we used the mutual_info_classif() function from the sklearn.feature_selection library. With the 70:30 train: test split, we developed the model with the following parameters: discrete_features='auto', n_neighbors=3, copy=True and random_state=None. We calibrated the train test ratio for maximizing the accuracies. For our dataset the 70:30 ratio was found to be more accurate and the same was selected for all the methods, in order to retain the uniformity. *Table 2* shows the top fourteen features (based on results from mutual information scores) are as follows: T12, T13, T14, T11, T26, T8, T23, T18, T6, T1, T15, T19, T25 and T5. The ranks are provided according to the descending score values.

Table 2 Mutual information ranking

Feature name	Mutual information score	Rank	Feature name	Mutual information score	Rank
T12	0.368	1	T7	0.146	15
T13	0.307	2	T10	0.146	16
T14	0.250	3	T24	0.142	17
T11	0.247	4	T16	0.140	18
T26	0.225	5	T17	0.129	19
T8	0.204	6	T27	0.128	20
T23	0.197	7	T2	0.125	21
T18	0.193	8	T21	0.116	22
T6	0.191	9	T3	0.114	23
T1	0.175	10	T9	0.114	24
T15	0.164	11	T28	0.108	25
T19	0.163	12	T20	0.089	26
T25	0.158	13	T4	0.009	27
T5	0.156	14	T22	0.000	28

For the information gain ratio implementation, we used the SelectKBest() function from sklearn.feature_selection library. In order to get the scores of all the features, we kept k='all' and used the score_func=mutual_info_classif. The data was split into 70:30 train: test ratio. *Table 3* shows the top fourteen features (based on results from information gain ratio scores) are as follows: T12, T13, T11, T26, T14, T15, T10, T18, T1, T8, T6, T23, T7 and T25. For the Correlation feature selection, we used the corrcoef() function from NumPy library. After obtaining the correlation score vector, we sorted the same. The data was split into 70:30 train: test ratio. *Table 4* shows that the top fourteen features (based on Correlation scores) are as follows: T12, T13, T14, T26, T11, T15, T10, T18, T1, T8, T23, T6, T3 and T7. For the Chi Square feature selection implementation, we used the SelectKBest() function from sklearn.feature_selection library with the 'chi2' option. The data was split into 70:30 train: test ratio. *Table 5* shows the top fourteen features (based on results from Chi Square scores) are as follows: T13, T23, T12, T14, T11, T15, T26, For the RMSE feature selection implementation, we used the mean_squared_error () function from sklearn.metrics library with the DecisionTreeRegressor() from sklearn.tree library. The data was split into 70:30 train: test ratio. *Table 6* shows the top fourteen features (based on results from RMSE scores) are as follows: T12, T13, T11, T14, T26, T23, T18, T6, T1, T25, T8, T15, T2 and T10. In case of RMSE, lower score means better. For the ROC feature selection, we used the roc_auc_score() function from sklearn.metrics library with the DecisionTreeClassifier() from sklearn.tree library. We get a true positive rate (sensitivity) of 1.0 and a false positive rate (1-specificity) of 0.0. The AUC score is 1 meaning the classifier accurately classifies the instances of positive and negative labels. *Figure 3* shows the confusion matrix for the dataset. The data was split into 70:30 train: test ratio. *Table 7* shows the top fourteen features (based on results from ROC - decision tree scores) are as follows: T13, T12, T11, T14, T26, T23, T18, T6, T15, T1, T19, T10, T8 and T16.

Table 3 Information gain ratio values

Feature name	Information gain score	Rank	Feature name	Information gain score	Rank
T12	0.214	1	T3	0.042	15
T13	0.167	2	T19	0.040	16
T11	0.130	3	T5	0.039	17
T26	0.127	4	T24	0.039	18
T14	0.117	5	T2	0.038	19
T15	0.079	6	T17	0.035	20
T10	0.073	7	T16	0.035	21
T18	0.062	8	T27	0.034	22
T1	0.056	9	T28	0.032	23
T8	0.056	10	T21	0.032	24
T6	0.053	11	T9	0.032	25
T23	0.052	12	T20	0.027	26
T7	0.046	13	T4	0.010	27

Feature name	Information gain score	Rank	Feature name	Information gain score	Rank
T25	0.044	14	T22	0.004	28

Table 4 Correlation values

Feature name	Correlation score	Rank	Feature name	Correlation score	Rank
T12	0.3078	1	T25	0.0459	15
T13	0.2406	2	T5	0.0447	16
T14	0.1795	3	T24	0.0427	17
T26	0.1793	4	T19	0.0422	18
T11	0.1744	5	T27	0.0409	19
T15	0.1528	6	T2	0.0397	20
T10	0.1246	7	T9	0.0396	21
T18	0.0754	8	T28	0.039	22
T1	0.0613	9	T16	0.0376	23
T8	0.0572	10	T17	0.0372	24
T23	0.056	11	T21	0.0333	25
T6	0.0533	12	T20	0.0331	26
T3	0.0506	13	T4	0.0234	27
T7	0.0473	14	T22	0.0206	28

Table 5 Chi square values

Feature name	Chi square score	Rank	Feature name	Chi square score	Rank
T13	42.705	1	T6	6.622	15
T23	39.613	2	T21	6.504	16
T12	39.070	3	T16	6.096	17
T14	32.629	4	T7	5.462	18
T11	30.804	5	T25	5.202	19
T15	22.696	6	T28	4.905	20
T26	21.642	7	T1	4.479	21
T17	11.476	8	T18	4.051	22
T5	9.872	9	T20	4.041	23
T10	8.480	10	T9	4.020	24
T27	8.461	11	T2	3.777	25
T8	7.796	12	T3	2.438	26
T19	6.983	13	T22	0.295	27
T24	6.951	14	T4	0.036	28

Table 6 RMSE values

Feature name	RMSE score	Rank	Feature name	RMSE score	Rank
T12	0.0851	1	T3	0.1786	15
T13	0.1066	2	T16	0.1793	16
T11	0.1288	3	T19	0.1793	17
T14	0.1369	4	T7	0.1874	18
T26	0.1390	5	T24	0.1874	19
T23	0.1519	6	T28	0.1915	20
T18	0.1554	7	T9	0.1919	21
T6	0.1571	8	T17	0.1934	22
T1	0.1624	9	T5	0.1944	23
T25	0.1661	10	T20	0.1946	24
T8	0.1677	11	T21	0.1954	25
T15	0.1754	12	T27	0.1960	26
T2	0.1759	13	T4	0.2253	27
T10	0.1761	14	T22	0.2387	28

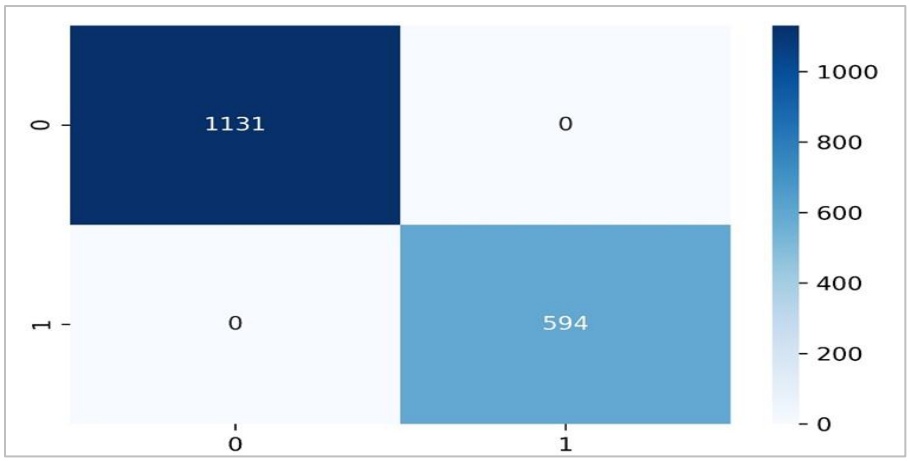

Figure 3 Confusion matrix calculated for the use of ROC feature selection

Table 7 ROC with decision tree values

Feature name	ROC score	Rank	Feature name	ROC score	Rank
T13	0.8796	1	T2	0.7155	15
T12	0.8778	2	T25	0.7151	16
T11	0.8621	3	T3	0.7146	17
T14	0.8524	4	T24	0.7137	18
T26	0.8408	5	T5	0.7109	19
T23	0.8266	6	T27	0.7107	20
T18	0.8008	7	T28	0.7056	21
T6	0.7916	8	T17	0.6991	22
T15	0.7779	9	T9	0.6990	23
T1	0.7614	10	T7	0.6779	24
T19	0.7450	11	T21	0.6773	25
T10	0.7416	12	T20	0.6651	26
T8	0.7369	13	T4	0.5662	27
T16	0.7296	14	T22	0.5249	28

For the ReliefF feature selection, we used the RFE() function from the sklearn.feature_selection library with the import RandomForestClassifier() of sklearn.ensemble library. We kept the n_estimators=100. The data was split into 70:30 train: test ratio. *Table 8* shows the top fourteen features (based on results from ReliefF scores) are as follows: T12, T13, T14, T26, T11, T15, T10, T18, T8, T1, T7, T5, T24 and T23.

For the BORUTA feature selection, we used the BorutaPy () function with RandomForestClassifier(), with parameters like n_jobs=-1, class_weight='balanced', max_depth=5, n_estimators='auto' and verbose=2. The data was split into 70:30 train: test ratio. *Table 9* shows the top fourteen features (based on results from BORUTA scores) are as follows: T12, T13, T11, T23, T27, T14, T26, T17, T20, T6, T1, T8, T5 and T18. For the OneR feature selection, we used the Decision Tree to create a tree of level 1. The Accuracies were calculated on basis of error rates. The data was split into 70:30 train: test ratio. *Table 10* shows the top fourteen features (based on results from One R scores) are as follows: T12, T13, T11, T14, T8, T1, T26, T6, T18, T7, T25, T3, T2 and T15.

Table 8 ReliefF values

Feature name	ReliefF score	Rank	Feature name	ReliefF score	Rank
T12	0.1942	1	T6	0.0147	15
T13	0.1645	2	T25	0.0144	16
T14	0.1525	3	T19	0.0142	17
T26	0.1379	4	T27	0.0137	18
T11	0.1353	5	T9	0.0126	19

Feature name	ReliefF score	Rank	Feature name	ReliefF score	Rank
T15	0.0932	6	T17	0.0112	20
T10	0.0779	7	T2	0.0096	21
T18	0.0348	8	T28	0.0095	22
T8	0.0292	9	T3	0.0070	23
T1	0.0271	10	T21	0.0064	24
T7	0.0221	11	T16	0.0041	25
T5	0.0198	12	T20	0.0029	26
T24	0.0170	13	T22	0.0004	27
T23	0.0158	14	T4	-0.0027	28

Table 9 BORUTA values

Feature name	BORUTA score	Rank	Feature name	BORUTA score	Rank
T12	24.7746	1	T7	11.1607	15
T13	18.5303	2	T25	10.5151	16
T11	17.1673	3	T19	10.1904	17
T23	16.9925	4	T21	9.6724	18
T27	16.5306	5	T28	9.6674	19
T14	15.2203	6	T15	9.6035	20
T26	15.1156	7	T24	9.5513	21
T17	14.7324	8	T16	8.9521	22
T20	13.6655	9	T2	8.1372	23
T6	13.6618	10	T22	7.7961	24
T1	12.9734	11	T3	7.4403	25
T8	12.7888	12	T10	7.1550	26
T5	12.3231	13	T9	6.6655	27
T18	11.8234	14	T4	1.0863	28

Table 10 One R values

Feature name	One R score	Rank	Feature name	One R score	Rank
T12	90.667	1	T10	74.667	15
T13	85.913	2	T24	73.449	16
T11	81.333	3	T19	73.159	17
T14	79.246	4	T17	72.638	18
T8	79.073	5	T20	72.464	19
T1	79.015	6	T5	72.406	20
T26	79.015	7	T21	72.406	21
T6	78.145	8	T16	72.232	22
T18	77.565	9	T28	71.826	23
T7	76.928	10	T23	71.420	24
T25	76.754	11	T9	71.246	25
T3	75.710	12	T27	70.493	26
T2	75.536	13	T4	67.073	27
T15	74.725	14	T22	63.826	28

5.Discussion

From the experimental results (*Table 2* to *Table 10*) it is prevalent that there is a significant agreement between the methods for the top 14 features. However, the ranking positions and scores vary due to the fact that different methods use different characteristics to calculate the importance scores. In order to combine the individual results, we used the majority vote and mean rank aggregation methods. In majority vote aggregation [48], we consider the rank indicated by the majority of procedures. *Figure 4* shows the majority vote ranking results. Lower the number of rank, higher is the weight e.g., rank 1 weighs more than rank 4.

Figure 5 shows the average ranking results. In this rank-aggregation scheme [49], a mean of the ranks (by different procedures) is calculated and the numbers are rounded off. The two aggregation schemes agree for the greatest number of ranks with a

slight variation in ranking and there is a perfect agreement for seven parameters (viz. T12, T13, T14, T26, T25, T19 and T4). The Four of the perfectly agreed parameters fall into the top 14 category. Moreover T11, T1 and T7 only vary by the decimals for the two schemes. The overall agreement for all parameters in the top – 14 category varies by positions 2-3. This shows a good agreement between the two aggregation schemes implying the parameters to be indicative for the cricket TiD.

In order to choose a relevant threshold for selecting the relevant features, we normalized the mean rankings as shown in *Table 11*. The top ranked features have low mean values and least ranked features have high mean values. Since there is no thumb rule to guide on how many features should be taken exactly, we relied on the human expert knowledge. In consultation with four domain experts and keeping in consideration the results, a threshold of 0.52 was chosen. The features scoring more than 0.52 are dropped and the features scoring less or equal to 0.52 of the normalized mean ranking are selected. *Figure 6* shows the graph of normalized mean rankings. The individual results from feature selection algorithms also reveal the same.

In our case, the top indicators with this threshold were found to be the following: T12 (throw catch accuracy), T13 (under arm throw), T11 (bowler accuracy), T14 (catching ability), T26 (balance), T15 (ground fielding), T23 (self-confidence), T10 (fatigue index), T6 (upper body strength), T18 (hand eye coordination), T1 (speed), T8 (reaction time), T5 (self-motivation) and T7 (lower body power). The results from the different feature selected methods upon aggregation, as shown in the results section, show that these 14 features are the strongest indicators for predicting the outcome. We compared the two aggregation schemes (majority vote and average ranking) and found a good and feasible agreement of the same. It can also be visualized from *Figure 4* and *Figure 5.*

Table 11 Normalized mean ranking results

Feature name	Score	Rank	Feature name	Score	Rank
T12	0.000	1	T24	0.533	15
T13	0.019	2	T25	0.548	16
T11	0.100	3	T19	0.598	17
T14	0.103	4	T17	0.625	18
T26	0.146	5	T27	0.667	19
T15	0.314	6	T16	0.670	20
T23	0.318	7	T28	0.690	21
T10	0.322	8	T2	0.724	22
T6	0.337	9	T3	0.782	23
T18	0.345	10	T9	0.793	24
T1	0.352	11	T21	0.835	25
T8	0.452	12	T20	0.847	26
T5	0.521	13	T4	0.996	27
T7	0.521	14	T22	1.000	28

In the final list of fourteen features, only four parameters (i.e., self-confidence, hand eye coordination, reaction time and self-motivation) fall from the cognitive ability category. The rest of ten features are from the physical/motor category. The top six of the features (i.e., throw catch accuracy-T12, under arm throw-T13, bowler accuracy test-T11, catching ability-T14, balance-T26 and ground fielding-T15) are game centric and are exclusively related to the cricket game itself. The Feature to rank number one was found to be the throw catch accuracy (T12). While measuring the value of this feature, the player has to pick up the ball and throw it to the wicket keeper, the number of times (out of 20) ball gets into the keeper gloves is the score. One logical explanation for throw catch accuracy (T12) to top the list may be that it already requires the components of the other features like speed, agility, reaction time, upper body strength, reaction time, flexibility, hand-eye coordination, etc. The results seem promising as the majority of the selected features fall into the physical/ sports domain. Such parameters are more imperative of the TiD and have a deep effect on the player performance.

The reduction of feature set from twenty-eight (28) parameters to fourteen (14) will prove to be a computational boost for any predictive model. It will cut down the computational costs and ultimately improve the performance. Training time for

algorithms will reduce. For the sports stakeholders, it will be a less painful task to take the parameter values for the enthusiasts. For example, for cricket TiD, if we select some students to identify whether they can be athletes or not. We need to take the values of fourteen parameters instead of the 28 parameters for the subjects. They should be sufficient to identify the talent among the subjects.

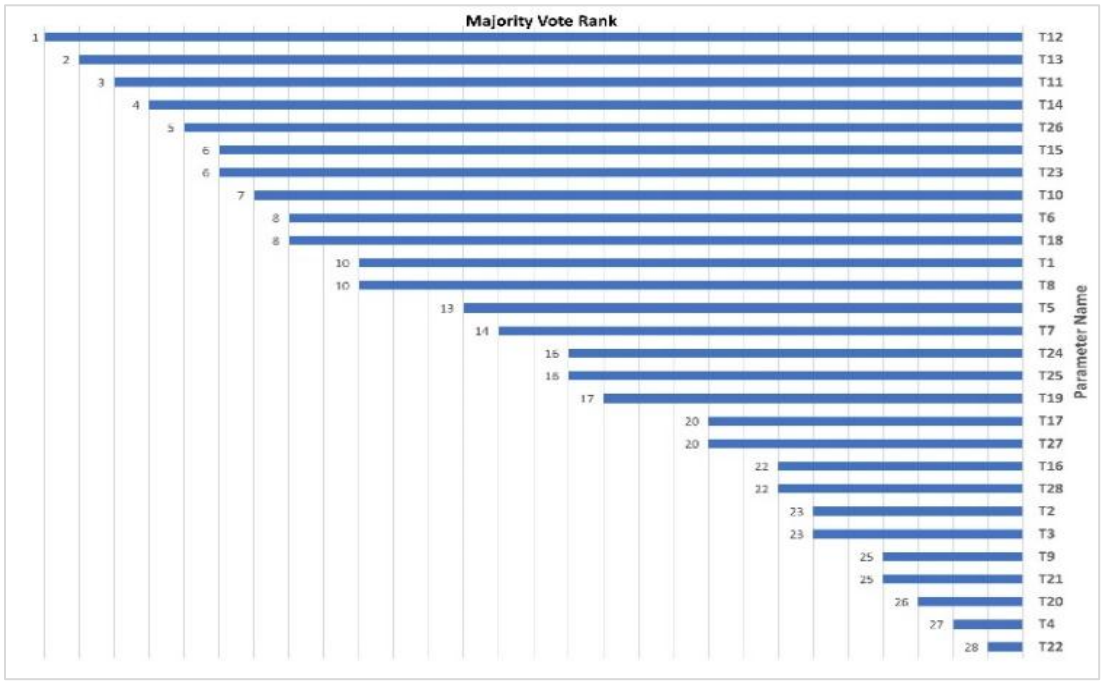

Figure 4 Majority vote results

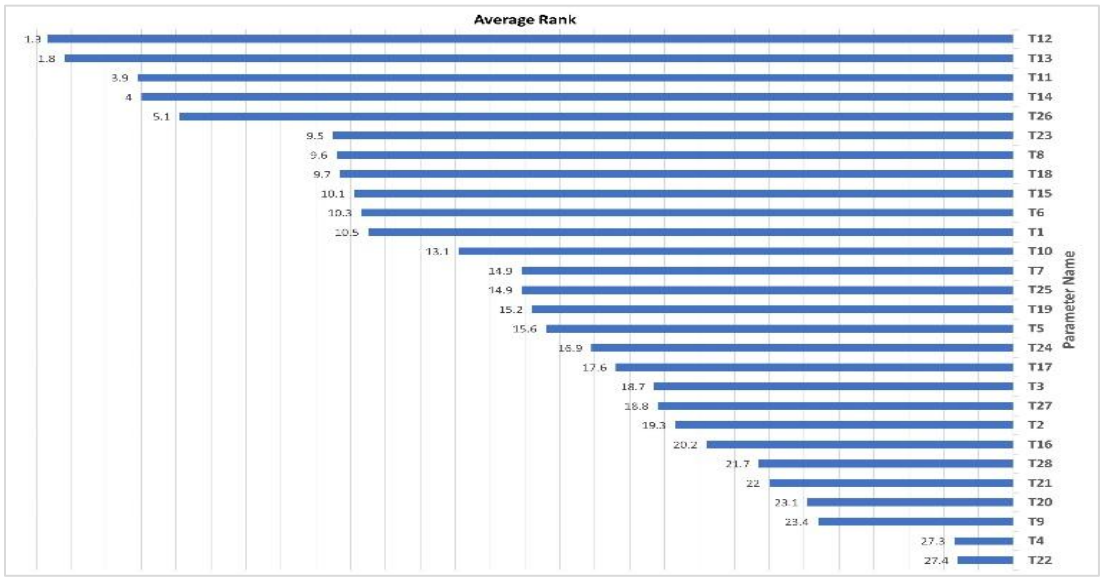

Figure 5 Average ranking

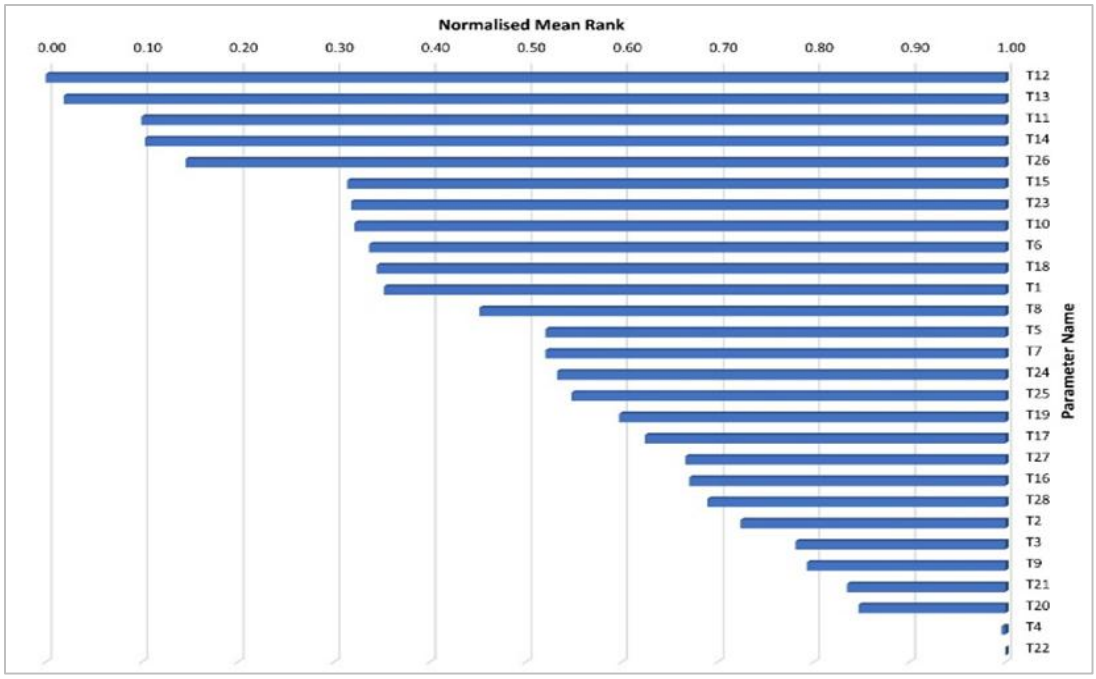

Figure 6 Normalised mean ranking results

5.1Limitations

The primary concern of feature reduction studies remains the fact that the notion of the importance of any feature is deduced from the data patterns and associations. This makes the feature selection a relative concept. Moreover, there is no thumb rule that can determine the exact number of features to be included. The same is overcome by including the expert knowledge into the system. The automation of the same remains an open challenge. One more challenge is that the different feature selection techniques work distinctively and hence the preference order of the feature varies as well. We have tried to overcome the same using rank aggregation algorithms as the solution was prevalent from literature. This study is cricket sport specific, since the dataset and the features are specifically taken from the same sport and hence the results cannot be generalized. However, the procedure of the study can be implemented for other sports as well. The absence of such study in the domain of cricket TiD is a hurdle to make it to some comparison scale. A complete list of abbreviations is shown in *Appendix I.*

6.Conclusion and future work

In this study, we perform an optimal feature selection for cricket TiD using nine different feature selection techniques. We calculate the individual rankings of all the feature selection methods. The results are aggregated using two different rank aggregation techniques, namely average ranking and majority vote ranking. The results show a significant agreement between the two schemes. A normalized mean ranking is calculated and on the basis of a threshold chosen in consultation with 4 domain experts, fourteen (14) features are found to be most contributing and indicative of the potential talent in the cricket sport. The selected features are: T12 (throw catch accuracy), T13 (under arm throw), T11 (bowler accuracy), T14 (catching ability), T26 (balance), T15 (ground fielding), T23 (self-confidence), T10 (fatigue index), T6 (upper body strength), T18 (hand eye coordination), T1 (speed), T8 (reaction time), T5 (self-motivation) and T7 (lower body power). 71.4% of the selected features are sport-centric and only 28.6% of the selected features are from the cognitive ability category. In future, the work will be practically tested and validated by using the full feature set and reduced feature set with a common TiD model. The comparative results will be insightful about the study. Moreover, the strategy may also be used for a variety of different datasets.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

Author's contribution statement

Mr. Naveed Jeelani Khan: Conceptualization, investigation, data curation, writing-original draft, analysis and interpretation of results. **Dr. Gulfam Ahamad:** Conceptualization, design, writing- review and editing, supervision. **Ms. Nahida Reyaz:** Data curation and study. **Dr. Mohd Naseem:** Technical support and supervision.

References

[1] Vaeyens R, Güllich A, Warr CR, Philippaerts R. Talent identification and promotion programmes of olympic athletes. Journal of Sports Sciences. 2009; 27(13):1367-80.

[2] Anshel MH, Lidor R. Talent detection programs in sport: the questionable use of psychological measures. Journal of Sport Behavior. 2012; 35(3): 239-66.

[3] Lidor R, Côté JE, Hackfort D. ISSP position stand: to test or not to test? the use of physical skill tests in talent detection and in early phases of sport development. International Journal of Sport and Exercise Psychology. 2009; 7(2):131-46.

[4] Bompa TO, Buzzichelli C. Periodization-: theory and methodology of training. Human Kinetics; 2018.

[5] Khan NJ, Ahamad G, Naseem M, Sohail SS. Computational efficiency in sports talent identification-a systematic review. International Journal of Applied Decision Sciences. 2022:1-34.

[6] Connor JD, Renshaw I, Farrow D. Defining cricket batting expertise from the perspective of elite coaches. PLoS One. 2020; 15(6):1-20.

[7] Premkumar P, Chakrabarty JB, Chowdhury S. Key performance indicators for factor score based ranking in one day international cricket. IIMB Management Review. 2020; 32(1):85-95.

[8] https://dtai.cs.kuleuven.be/events/MLSA16/slides/06_Madan_Gopal.pdf. Accessed 14 April 2022.

[9] Manage AB, Kafle RC, Wijekularathna DK. Classification of all-rounders in limited over cricket-a machine learning approach. Journal of Sports Analytics. 2020; 6(4):295-306.

[10] Zare CMA. An effective method of feature selection in persian text for improving the accuracy of detecting request in Persian messages on telegram. Journal of Information Systems and Telecommunication. 2021; 4(32):249-62.

[11] Alice K, Natesan K, Dhanalakshmi B, Jaisharma K. Role of attribute selection on tuning the learning performance of Parkinson's data using various intelligent classifiers. International Journal of Advanced Technology and Engineering Exploration. 2021; 8(78):560-75.

[12] Wiharto W, Suryani E, Susilo M. Performance analysis of hybrid SOM and AdaBoost classifiers for diagnosis of hypertensive retinopathy. Journal of Information Systems and Telecommunication. 2021; 2(34):79-88.

[13] Barney EG. Preliminary stages in the validation of a talent identification model in cricket. Bangor University (United Kingdom); 2015:1-23.

[14] Ahamad G, Naqvi SK, Beg MS. OWA based model for talent selection in cricket. In advance trends in soft computing: proceedings of WCSC 2013, San Antonio, Texas, USA 2014 (pp. 229-39). Springer International Publishing.

[15] Ahamad G, Naqvi SK, Beg MS. A model for talent identification in cricket based on OWA operator. International Journal of Information Technology & Management Information System. 2013; 4(2):40-55.

[16] Johnston FE. The physique of the olympic athlete, by JM Tanner, with the assistance of RH Whitehouse and Shirley Jarman. 126 pp., 6 tables, 80 figures, 118 plates. George Allen and Unwin, Ltd., London. 1964; 22(4):494-5

[17] Taha Z, Musa RM, Majeed AP, Alim MM, Abdullah MR. The identification of high potential archers based on fitness and motor ability variables: a support vector machine approach. Human Movement Science. 2018; 57:184-93.

[18] Noori M, Sadeghi H. Designing smart model in volleyball talent identification via fuzzy logic based on main and weighted criteria resulted from the analytic hierarchy process. Journal of Advanced Sport Technology. 2018; 2(1):16-24.

[19] Kusnanik NW, Hariyanto A, Herdyanto Y, Satia A. Talent identification model for sprinter using discriminant factor. In IOP conference series: materials science and engineering 2018 (pp. 1-6). IOP Publishing.

[20] Rozi F, Setijono H, Kusnanik NW. The identification model on swimming athletes' skill. Theory and Methodology of Physical Education and Sports. 2019; 27(4):30-5.

[21] Dwivedi P, Chaturvedi V, Vashist JK. Efficient team formation from pool of talent: comparing AHP-LP and TOPSIS-LP approach. Journal of Enterprise Information Management. 2020; 33(5):1293-318.

[22] Mat-rasid SM, Abdullah MR, Juahir H, Maliki AB, Kosni NA, Musa RM, et al. Applied multidimensional analysis for assessing youth performance in sports talent identification program. International Journal of Recent Technology and Engineering. 2019; 8(2S7):207-11.

[23] Huang X, Wang G, Chen C, Liu J, Kristiansen B, Hohmann A, et al. Constructing a talent identification index system and evaluation model for cross-country skiers. Journal of Sports Sciences. 2021; 39(4):368-79.

[24] Priscilla CV, Prabha DP. A two-phase feature selection technique using mutual information and XGB-RFE for credit card fraud detection. International Journal of Advanced Technology and Engineering Exploration. 2021; 8(85):1656-68.

[25] Liu H, Yu L. Toward integrating feature selection algorithms for classification and clustering. IEEE Transactions on Knowledge and Data Engineering. 2005; 17(4):491-502.

[26] Yu L, Liu H. Feature selection for high-dimensional data: a fast correlation-based filter solution. In proceedings of the 20th international conference on machine learning 2003 (pp. 856-63).

[27] Liu H, Setiono R. A probabilistic approach to feature selection-a filter solution. In ICML 1996 (pp. 319-27).
[28] Kohavi R, John GH. Wrappers for feature subset selection. Artificial Intelligence. 1997; 97(1-2):273-324.
[29] Das S. Filters, wrappers and a boosting-based hybrid for feature selection. In ICML 2001 (pp. 74-81).
[30] Karegowda AG, Jayaram MA, Manjunath AS. Feature subset selection problem using wrapper approach in supervised learning. International Journal of Computer Applications. 2010; 1(7):13-7.
[31] Ahamad G, Naqvi SK, Beg MS, Ahmed T. A web based system for cricket talent identification, enhancement and selection (C-TIES). Procedia Computer Science. 2015; 62:134-42.
[32] https://era.ed.ac.uk/handle/1842/1952. Accessed 14 April 2022.
[33] Mann DL, Dehghansai N, Baker J. Searching for the elusive gift: advances in talent identification in sport. Current Opinion in Psychology. 2017; 16:128-33.
[34] Xian S, Guo H, Chai J, Wan W. Interval probability hesitant fuzzy linguistic analytic hierarchy process and its application in talent selection. Journal of Intelligent & Fuzzy Systems. 2020; 39(3):2627-45.
[35] Curran O, Macnamara A, Passmore D. What about the girls? exploring the gender data gap in talent development. Frontiers in Sports and Active Living. 2019; 1(3):1-7.
[36] Cover TM, Thomas JA. Information theory and statistics. Elements of Information Theory. 1991; 1(1):279-335.
[37] Vergara JR, Estévez PA. A review of feature selection methods based on mutual information. Neural Computing and Applications. 2014; 24:175-86.
[38] Torkkola K. Information-theoretic methods. Feature Extraction: Foundations and Applications. 2006:167-85.
[39] Salzberg SL. C4. 5: programs for machine learning by J.Ross quinlan. Morgan Kaufmann Publishers. 1993:235-40.
[40] Duch W. Filter methods. Feature Extraction: Foundations and Applications. 2006:89-117.
[41] Jin X, Xu A, Bie R, Guo P. Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene expression profiles. In data mining for biomedical applications: PAKDD 2006 workshop, BioDM 2006, Singapore proceedings 2006 (pp. 106-15). Springer Berlin Heidelberg.
[42] Embrechts MJ, Bress RA, Kewley RH. Feature selection via sensitivity analysis with direct kernel PLS. Feature Extraction: Foundations and Applications. 2006:447-62.
[43] Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition. 1997; 30(7):1145-59.
[44] Robnik-sikonja M, Kononenko I. Theoretical and empirical analysis of ReliefF and RReliefF. Machine Learning. 2003; 53:23-69.
[45] Urbanowicz RJ, Meeker M, La CW, Olson RS, Moore JH. Relief-based feature selection: introduction and review. Journal of Biomedical Informatics. 2018; 85:189-203.
[46] Kursa MB, Jankowski A, Rudnicki WR. Boruta–a system for feature selection. Fundamenta Informaticae. 2010; 101(4):271-85.
[47] Sujatha M, Devi L. Feature selection techniques using for high dimensional data in machine learning. International Journal of Engineering Research & Technology. 2013; 2(9):2909-16.
[48] Bauer E, Kohavi R. An empirical comparison of voting classification algorithms: bagging, boosting, and variants. Machine Learning. 1999; 36:105-39.
[49] Wald R, Khoshgoftaar TM, Dittman D, Awada W, Napolitano A. An extensive comparison of feature ranking aggregation techniques in bioinformatics. In 13th international conference on information reuse & integration 2012 (pp. 377-84). IEEE.

Mr. Naveed Jeelani Khan has completed his Bachelor of Sciences (BSc) and Masters in Information Technology (MSC IT) from the University of Kashmir, Srinagar J&K - India. He has one year of teaching and three years of research experience. He is currently pursuing his Doctoral Degree from the Department of Computer Sciences BGSBU, Rajouri J&K – India. He has published several articles in International Journals. His Research Interests include Sports Science, Applied Soft Computing, Artificial Intelligence, Nature Inspired Optimisation, Machine Learning and IoT.
Email: naveedjeelani@bgsbu.ac.in

Dr. Gulfam Ahamad has completed his B.Sc. from Chaudhary Charan Singh (CCS) University Meerut, Uttar Pradesh, India and PhD from Jamia Milla University, New Delhi - India in the field of Applied Soft Computing. He has 10 years of teaching and research experience. He has published a number of quality papers in various well reputed international journals. He is currently working as an Assistant Professor in the Department of Computer Sciences, BGSBU Rajouri, J&K – India. His Research Interest include Sports Science, Artificial Intelligence, Soft Computing, Fuzzy Computing, Internet of Things, etc.
Email: gulfamahamad@bgsbu.ac.in

Ms. Nahida Reyaz has completed her Bachelors in Computer Applications (BCA) and Masters in Computer Applications (MCA) from the University of Kashmir, Srinagar J&K - India. She is currently pursuing her Doctoral Degree from the Department of Computer Sciences BGSBU, Rajouri J&K – India. Her Research Interests include Data Science, Machine Learning and Internet of Things.
Email: beignahida9995@gmail.com

Dr. Mohd Naseem has completed his Master of Computer Applications (M.C.A), from the Department of Computer Science, Aligarh Muslim University, Aligarh, Uttar Pradesh, India and PhD degree from Indian Institute of Technology, Indian School of Mines, Dhanbad, (IIT (ISM) Dhanbad), Jharkhand, India. He has 6 years of teaching and research experience. He has published a number of quality papers in various well reputed international journals. He is currently working as an Assistant Professor in the Department of Computer Sciences, BGSBU Rajouri, J&K – India. His Research Interest include Network Security, Internet of Things, Internet of Vehicles, etc.
Email: naseemmohd@bgsbu.ac.in

Appendix I

S. No.	Acronym	Definition
1	AUC	Area Under the Curve
2	MIRA	Maximal Importance of the Random Attributes
3	OWA	Ordered Weighted Averaging
4	RAST	Running Based Anaerobic Sprit
5	RMSD	Root Mean Square Deviation
6	RMSE	Root Mean Square Error
7	ROC	Receiver Operating Characteristic
8	TiD	Talent Identification