Shahid Bahonar University of Kerman

Journal of New Studies in Sport Management

Online ISSN: 2717 - 4069

Homepage: https://jnssm.uk.ac.ir

Iranian Scientific Association of Sport Management

Performance Analysis of Elite Taekwondo Athletes by Data Clustering and Semi-Supervised Prediction Approaches

Sana Esmaeili Abharian[1], Neda Abdolvand[2], Zhaleh Memari[3], Tahereh Esmaeili Abharian[4]

1. Master's Degree in Information technology Management, Department of Management, Faculty of Social Sciences and Economics, Alzahra University, Tehran, Iran. Email: sana.abharian@gmail.com
2. Associate Professor, Department of Management, Faculty of Social Sciences and Economics, Alzahra University, Tehran, Iran. n.abdolvand@alzahra.ac.ir.
3. Associate Professor, Department of Sports Management, Faculty of Sport Science, Alzahra University, Tehran, Iran. Email: zh.memari@alzahra.ac.ir.
4. Master's Degree in Computer Engineering, Faculty of Electrical, Computer and Information Technology, Islamic Azad University, Qazvin Branch, Qazvin, Iran. Email: tahereh.abharian@gmail.com.

ARTICLE INFO

Article type:
Original article

Article history:
Received: 18 October 2024
Received in revised form: 15 January 2025
Accepted: 20 January 2025
Published online: 21 August 2025

Keywords:
Machine Learning
Performance Analysis
Semi-Supervised prediction
Sport Analytics
Taekwondo

ABSTRACT

The purpose of this paper was to analyze the performance of elite taekwondo athletes using machine learning approaches to achieve three objectives: categorizing performance into four clusters ranging from excellent to poor, identifying key physical characteristics influencing performance, and predicting medal-winning potential in world competitions. This study employs the National Olympic Academy dataset of Iranian taekwondo athletes' physical fitness and anthropometric records from 1996 to 2019 to develop descriptive and predictive models. In datasets comprising 999 female and 1560 male records, SOM-means and SOM-spectral clustering algorithms achieved average test efficiencies of 80%, for Silhouette, and 20% for Davies-Bouldin in the female dataset, and 79% and 34% respectively, in the male dataset, identifying four performance clusters based on physical attributes and medal distribution. A semi-supervised learning model with the CPLE-Learning algorithm demonstrated medal prediction capabilities, with accuracy rates of 68%, 59%, and 73% for predicting gold, silver, and bronze medals in the female dataset, and 58%, 61%, and 54% in the male dataset. These findings highlight the effectiveness of machine learning in sports performance analysis, offering valuable insights for managing taekwondo athletes and enhancing physical preparation strategies.

Introduction

The complexity and dynamics of modern sports necessitate systemic observation and measurement to deepen our understanding of athletic performance. Performance analysis has become a fundamental tool for creating reliable records that aid athletes and coaches in refining strategies and improving outcomes (O'Donoghue, 2010). Sports performance is influenced by the combination of

How to Cite: Esmaieli Abharian, S, Abdolvand, N., Memari, Z., & Esmaieli Abharian, T. (2025). Performance Analysis of Elite Taekwondo Athletes by Data Clustering and Semi-Supervised Prediction Approaches. *Journal of New Studies in Sport Management, 7*(1), 55-68. DOI: 10.22103/jnssm.2025.24225.1336

 Publisher: Shahid Bahonar University of Kerman

DOI: 10.22103/jnssm.2025.24225.1336

physical, mental, technical, and tactical skills, which are evaluated based on established standards and athletes' results (Zahradník & Korvas, 2018).

In high-intensity sports like taekwondo, achieving competition is contingent mainly upon physical fitness (Ziv & Lidor, 2010), with athletes preparing for the physical and psychological demands of combat (Bridge et al., 2009; Casolino et al., 2012). Physical fitness indicators such as speed, agility, aerobic and anaerobic power, muscular strength, and flexibility are essential, and to succeed, athletes are expected to excel across these domains (Bouhlel et al., 2006). With the exponential growth of data in sports, especially within competitive and professional contexts, distinguishing between successful and unsuccessful performances has become increasingly challenging (Cao, 2012). As a result, sports organizations worldwide, including those in Iran, are investing in data-driven approaches to improve training, strategic decision-making, and performance outcomes (Kostakis et al., 2017; Pelechrinis & Papalexakis, 2018; Tichy, 2016).

Advances in artificial intelligence (AI) and machine learning (ML) have allowed for a more nuanced analysis of the diverse factors affecting performance, encompassing physical, psychological, tactical, and technical attributes (Zahradník & Korvas, 2018).

In sports analytics, research into performance prediction models generally falls into two types: descriptive and predictive. Descriptive models often rely on clustering techniques to reveal patterns in athletes' physical and psychological traits, helping us understand how these traits affect their performance (Musa et al., 2019; Taha et al., 2018). Clustering is a method that sorts data into groups, or "clusters," based on shared characteristics- much like how a coach might group players by similar skill levels or physical strengths. For example, self-organizing maps (SOM), a popular robust clustering approach, organize data into meaningful clusters, allowing coaches and analysts to see patterns within the team. SOM is particularly helpful for sports teams, as it can reveal relationships among players and positions, helping coaches create more balanced and effective lineups (Takemura et al., 2018).

Takemura et al. (2014 & 2018) used SOM to study team dynamics in volleyball and rugby, organizing players based on their physical and mental strengths to better understand team roles and relationships. In rugby, for instance, SOM helped coaches identify key performance traits for specific positions (Croft et al., 2015; Zheng et al., 2020). SOM has also been paired with decision-making tools like the analytic hierarchy process (AHP), where the two can guide team-building strategies. For example, in baseball, clustering combined with AHP helped a team's management group players based on essential skills and strengths, making it easier to plan for the season and strengthen the roster (Kohara & Enomoto, 2018).

Semi-supervised learning techniques are also becoming popular in sports because they allow analysts to work with labeled and unlabeled data. In sports, collecting detailed, labeled data (like specific player statistics) can be time-consuming and expensive, so semi-supervised methods help fill in the gaps by making use of both labeled data (like player status) and unlabeled data (such as GPS tracking). These methods can work in stages: the model learns from labeled data and then independently makes sense of unlabeled data. For example, a semi-supervised approach in taekwondo could initially use key statistics, like speed and agility, to classify athlete performance. Then, adding more general data could refine these predictions to highlight emerging talents and focus training efforts more effectively.

Predictive models are designed to uncover hidden relationships and patterns among known data to enhance performance prediction accuracy. For example, Gu et al. (2019) integrated principal component analysis (PCA), nonparametric statistical analysis, support vector machine (SVM), and ensemble learning to develop predictive models for hockey player recruitment. In basketball, machine learning algorithms such as decision trees, neural networks, and naïve Bayes have been applied to predict game outcomes based on tactical variables (Bunker & Thabtah, 2019). New research in soccer has shown that deep learning and hybrid machine learning models effectively predict critical game events, with ensemble techniques combining neural networks and gradient boosting outperforming traditional methods (Yeung et al., 2024). Additionally, Sarlis and Tjortjis (2020) employed neural networks to analyze NBA statistics to predict awards like "Most Valuable Player" and "Defender of the Year," underscoring AI's utility in identifying high-impact performance indicators. Mendes-Neves et al. (2024) presented the Large Events Model (LEM), an

innovative deep-learning framework for soccer match simulation and analysis that predicts event probabilities and outcomes from specific game states.

While advanced machine learning techniques have been widely applied in particular team sports, their use in martial arts, including taekwondo, remains limited (Park et al., 2009). This gap is particularly significant for Iran, a nation with a strong tradition in taekwondo, where national teams constantly rank among the world's top competitors. However, sports performance studies in Iran have focused mainly on football, with attempts to apply predictive models to other sports. For instance, Memari et al. (2020) used neural networks and decision trees to analyze player valuation in the Iranian Premier League, identifying age and physical strength as key predictive factors in pricing. Similarly, Izadyar et al. (2016) applied ordinary least squares (OLS) regression to estimate player pricing, finding significant effects of tactics, technique, and club brand on valuations, with physical fitness inversely correlated with player value. These findings highlight the utility of predictive models in the Iranian sports setting but also underscore the limited application of such methods beyond football.

Given this gap, this study focuses on Iranian taekwondo athletes, seeking to develop a performance prediction model tailored to the unique demands of this martial within the Iranian sports context. Unlike football, where extensive data infrastructure and analytical models are already established, taekwondo lacks such resources despite its popularity and competitive success in Iran. This study addresses this need by applying data clustering and semi-supervised prediction approaches to evaluate performance indicators and predict the likelihood of Iranian taekwondo athletes winning medals. By building upon existing techniques and incorporating Iran-specific considerations- such as local training regimens, competition strategies, and cultural factors influencing athletes' development- this research aims to offer actionable insights for coaches, sports managers, and athletes in Iran. Moreover, this study will contribute to the broader field of sports analytics by providing a scalable model that can be adapted to other martial arts or individual sports, enriching the field with an Iran-centered perspective on performance prediction.

Methodology

Data sample

Data were collected from the National Olympic Academy dataset of physical fitness and physiological records of Iranian taekwondo players from 1996 to 2019. The dataset contains 2559 records, of which 999 belong to females and 1560 to males. The dataset covers a range of features such as taekwondo players' anthropometric, physical fitness, and general features. The anthropometric data on the biological characteristics of athletes has nine components. Twenty-four physical fitness features of the athletes were taken from their fitness tests, of which 12 features had almost 70% missing values. We eliminated them in the preprocessing stage, as they will not affect the final decision. Moreover, the Iran Taekwondo Federation provided general features such as the Year of sports activity and the competition results of the medal-winning taekwondo athletes to compare this dataset. The characteristics of the data were determined after analyzing the given dataset.

Table 1. Introduction of taekwondo athletes' features

Features	Description	Missing Values %
Year	The year of athlete's sport activity.	0%
Competition Rank	Represents the achieved rank of taekwondo players in the competitions as first, second, or third rank in three categories of World Cup, Asian championship and Olympic competitions.	0%
Age	Athlete's age.	8.4%
Weight	Body weight or mass.	7.3%
Standing height	Athlete's height in a standing position.	15.1%

Features	Description	Missing Values %
Sitting height	Athlete's height in a sitting position.	23.5%
Arm span	The physical measurement of the length from one side of the athlete's arms to the other.	55.5%
Fat	Athlete's body fat.	37.6%
VO2max	Measuring the maximum amount of oxygen consumed or absorbed during strenuous exercise.	46.1%
Hearth rate	The number of heart beats per minute.	59.6%
40-Yard	A physical fitness test measuring the athlete's acceleration and speed.	7.1%
Sit-up	A muscular endurance test when an athlete elevating his/her trunk from supine position until his/her elbows.	28.6%
Side-to-side Bounding	Measuring the muscle strength of the legs by jumping side to side using the throwing movements of the hands and opening the knees and straightening the stature.	18.0%
4*9-m shuttle run	Assessing speed, agility and coordination (motor skills) by running athletes four times at his/her maximum speed in a distance of 9.0 meters.	52.6%
Illinois agility run	An athlete's agility test by running athletes at his/her maximum speed along the path of 10 meters with several cone-shaped obstacles.	56.4%
Visual reaction time (VRT)	An indicator of the information processing speed that can affect the performance of sports skills.	14.8%
Grip strength	Measures the ability of the grip and toe to produce maximum muscle strength.	47.5%
Sit and reach	A measurement to evaluate athlete's lower back flexibility and hamstring.	15.8%
Trunk lift	The trunk lifts test measures the flexibility, strength along with endurance of an athlete's back.	65.7%
Hamstring flexibility	It tests the flexibility, strength and endurance of an athlete's hamstring muscle.	63.7%
Vertical jump	To assess the power of the lower body, the vertical height of an athlete's jump is measured.	25.4%
Ergo jump	Measures aerobic and anaerobic power at sub-maximum intensities.	13.4%

Data analysis

The primary data collected often requires preprocessing to be prepared for applying machine learning algorithms. In our case, the female dataset was first separated from the male athlete dataset. Therefore, the preprocessing and modeling steps were used separately.

After examining and recognizing each feature, the action was taken to remove the features whose variances were precisely the same, or 70% of the variance on their values did not change with the variance threshold function. Considering the missing values, it examines the data variance relative to the mean and deletes the data with no change, as they will not affect the final decision. As missing values pose a difficulty in the modeling, we used imputation methods to deal with the remaining missing values, as shown in Table 1. The imputation method has the advantage that it can be applied once to modify the data. Various imputation methods insert the missing values in each of the selected attributes according to the type of modeling, including mean, median, and nearest neighbor, as well as Iterative Imputer strategy, which contains regressions like BayesianRidge, KNeighborsRegressor, DecisionTreeRegressor, ExtraTreesRegressor. Next, the data was standardized with the Standard Scaler class to fit into a scale if the individual attributes did not look like standard normally distributed data.

Since the data are related to the physical factors of professional athletes, the values of the characteristics are very close to different athletes, and the data are very overlapping. Therefore, to better separate data and extract information from it, we need powerful feature extraction algorithms to make it by transferring data to another space. Algorithms such as Latent Dirichlet Allocation

(LDA), Singular Value Decomposition (SVD), and Fast Independent Component Analysis (Fast ICA) were tested to extract the features of this dataset according to the type of modeling. This stage of data preprocessing was implemented simultaneously with the modeling stage.

The modeling phase is then divided into two stages. The first phase involves data clustering, which is performed to identify similar data and determine the types of clusters. In the second phase, we employ semi-supervised classification techniques to predict the number of medals won by male and female taekwondo athletes in world competitions.

We combined the K-means algorithm and SOM neural network to cluster the female dataset. For data clustering, the K-means and SOM are two popular techniques for dealing with large-scale databases (Everitt et al., 2011). In the second step, SOM maps a large-scale dataset on its network topology and generates the topological coordinates of the prototypes for k-means clustering. The excellent performance of the SOM in the training of its network depends on setting several parameters: (1) Learning rate, (2) Number of iterations of the algorithms for training the network (epochs), (3) Dimensions of the SOM network (number of neurons). To achieve reasonable efficiency, we put the number of network neurons to 4*4, the number of epochs to 10000, and the learning rate to 0.1. After SOM training, we implemented a K-means algorithm to modify the obtained weights from the training SOM. A combination of SOM and spectral clustering algorithms is used in the male dataset. Since the SOM map can be considered a graph, and the node weights can be considered similarity measures, we utilized the spectral clustering method to divide the SOM map into clusters with closer nodes. The node weights can be thought of as similarity measures. In other words, we transformed the SOM map to the similarity matrix of the individual nodes. The SOM-spectral clustering algorithm, Fast ICA feature extraction algorithm, and Bayesian Ridge strategy for estimating missing values showed a good performance.

Furthermore, we used a semi-supervised perspective to gain deeper insight into the taekwondo athletes' performance evaluation to predict their medal prospects in international competitions. Semi-supervised classification aims to employ many unlabeled data to build a better classifier from the labeled data. It teaches a classifier using both labeled and unlabeled datasets. Since our dataset has labeled and unlabeled data categories, different methods of the semi-supervised approach have been used to achieve the best prediction algorithm for taekwondo athletes, such as Semi Boost, Self-learning, and CPLE (SVM-classifier). Initially, a cross-validation function was designed to split the test and training datasets, with 60% of the labeled data being put into the test sub-dataset. Then, 40% of the labeled data was divided into k-folds. At each stage of the training, unlabeled data enters the learning phase with the k-1fold category, and the k-fold is used for validation. We used ensemble methods for multi-class modes, such as "one class against the other classes." In different implementations associated with different feature extraction algorithms, such as NMF, SVD, and PCA, the best result was obtained after applying the CPLE algorithm with the SVM classifier and SVD feature extraction algorithm. The CPLE framework only allows the use of probability-based classifications. In each dataset, we built (training and testing) a multiclass classifier (3 classes) for gold, silver, and bronze classes by developing the SVM classifier as a binary classifier and the usage of RBF kernel function, which is the most popular kernel among all the kernels in SVM (Gopi et al., 2023). We used SVD as a feature extraction algorithm among CPLE-Learning (base model=SVM (kernel= 'RBF') in female and male datasets.

Results

Figure 1 (a) shows the excellent performance of clustering the female dataset using the SOM-k-means algorithm, which is based on two primary components of the Fast ICA feature extraction algorithm. We examined the number of medals achieved in each cluster. Four clusters were extracted, with clusters 0, 1, and 2 having the most medals, but cluster 3, with its main components in the lower range, has the fewest medals.

Figure 1 (b) shows the clustering of the male dataset using the SOM-spectral clustering algorithm based on the two main components of the Fast ICA feature extraction. For each cluster, we examined the number of medals won by athletes. Four clusters were extracted according to the

number of medals identified in each cluster: cluster 3 is excellent, Clusters 0 and 1 are intermediate clusters, and cluster 2 is weak.

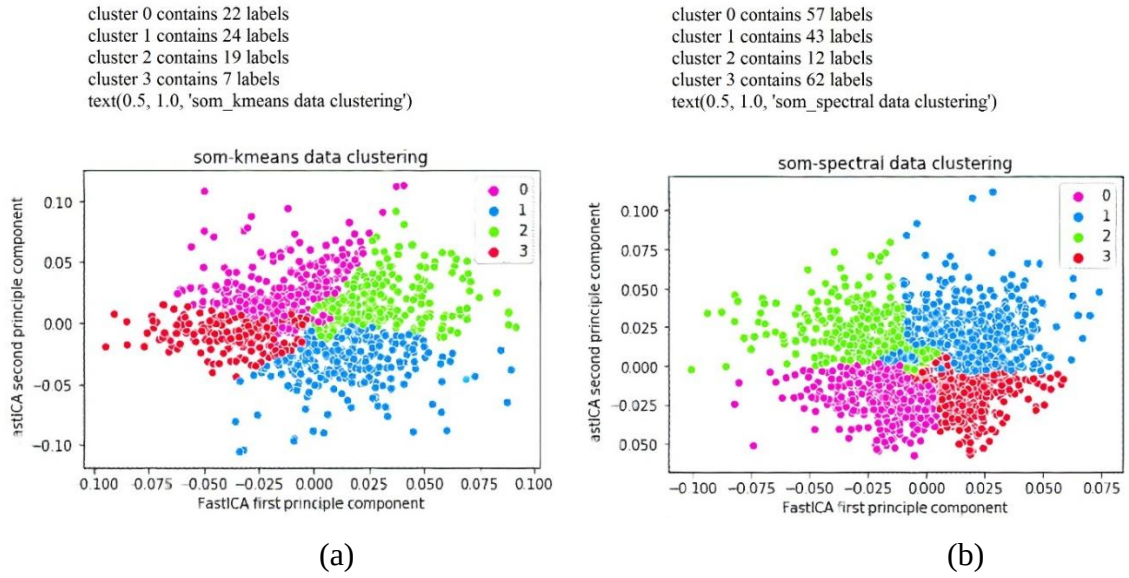

Figure 1. Clustering is based on the two main components of Fast ICA using the SOM-K-means algorithm in the female dataset (a) and the SOM-spectral clustering algorithm in the male dataset (b).

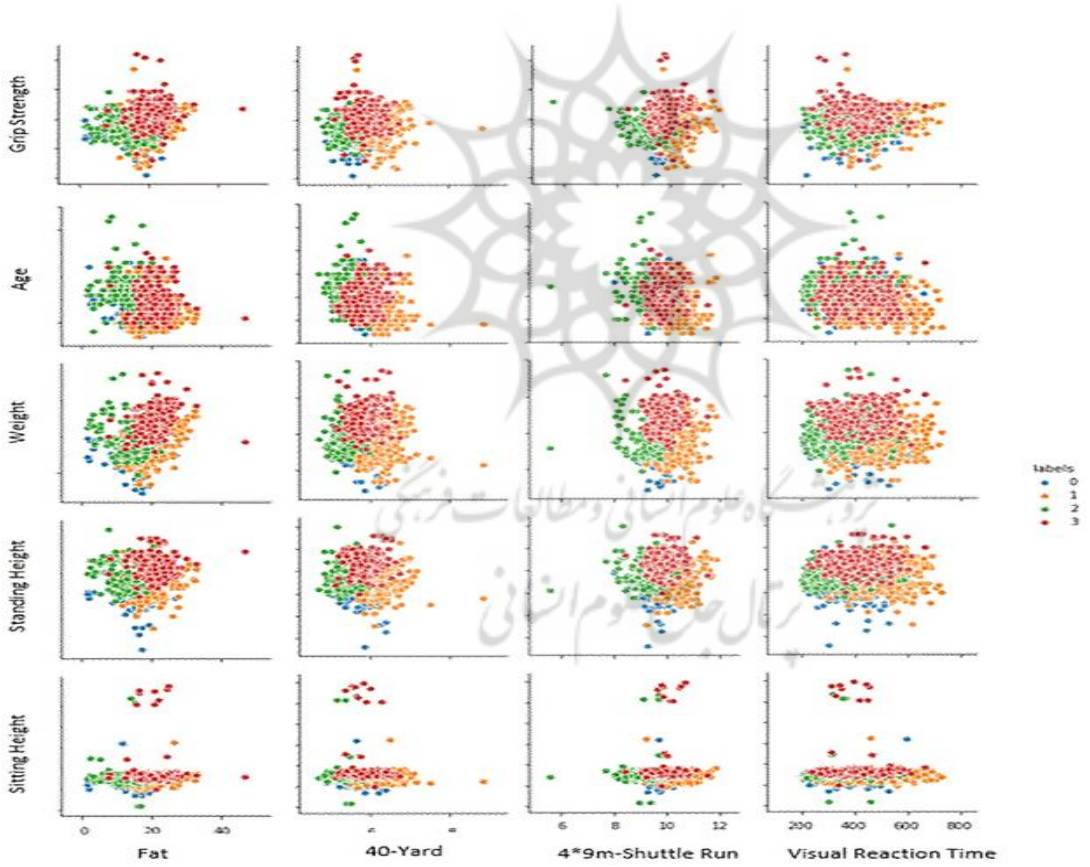

Figure 2. Clustering of the female dataset using the SOM-K-means algorithm based on the most practical features in calculating the first and second components of the fast ICA algorithm.

We extracted the most compelling features in the female and male datasets based on the calculations of the main components of the Fast ICA feature extraction algorithm. Then, we visualized the female and male clustering based on the most valuable features of the first and second components. In the female dataset, we found that features like **fat**, **40-yard**, **4*9-m shuttle run**, and **VRT** significantly affect the first component of Fast ICA feature extraction. For the

second component, features like **age**, **weight**, **standing height**, **sitting height**, and **grip strength** have the most significant effect. According to Figure 2, an average percentage of fat is one of the essential anthropometric variables for female athletes in cluster 1, who have the most medals. Most medalists (cluster 1) performed well in the 4*9-m shuttle run test, demonstrating their agility. Although they do not show high scores in the 40-yard feature, speed is an essential and influential performance indicator. They have a medium score in VRT and Grip strength tests. Most are young, unlike those with short-sitting heights and almost high-standing heights. They are in the range of medium to increased weight.

In the male dataset, features like **standing height**, **sitting height**, **40-yard**, **4*9-m shuttle run** have the most impact in the calculation of the first component of Fast ICA feature extraction, and features of **fat**, **40-yard**, **4*9-m shuttle run,** and **VRT** have the most impact in the measure of its second component. As shown in Figure 3, the athletes in clusters 3 and 0, who won the most medals, are in the younger age group with low weight. In terms of standing height and sitting height, they are almost tall. In this regard, their arm span is moderately long. Although the vertical jump in cluster 3 has a relatively low score compared to clusters 1 and 2, it is one of the most important features for good performance. Each cluster's medal winners cover a wide range of points in the VRT and the 9*4-m shuttle run tests. They score well in the Illinois agility run and 40-yard tests.

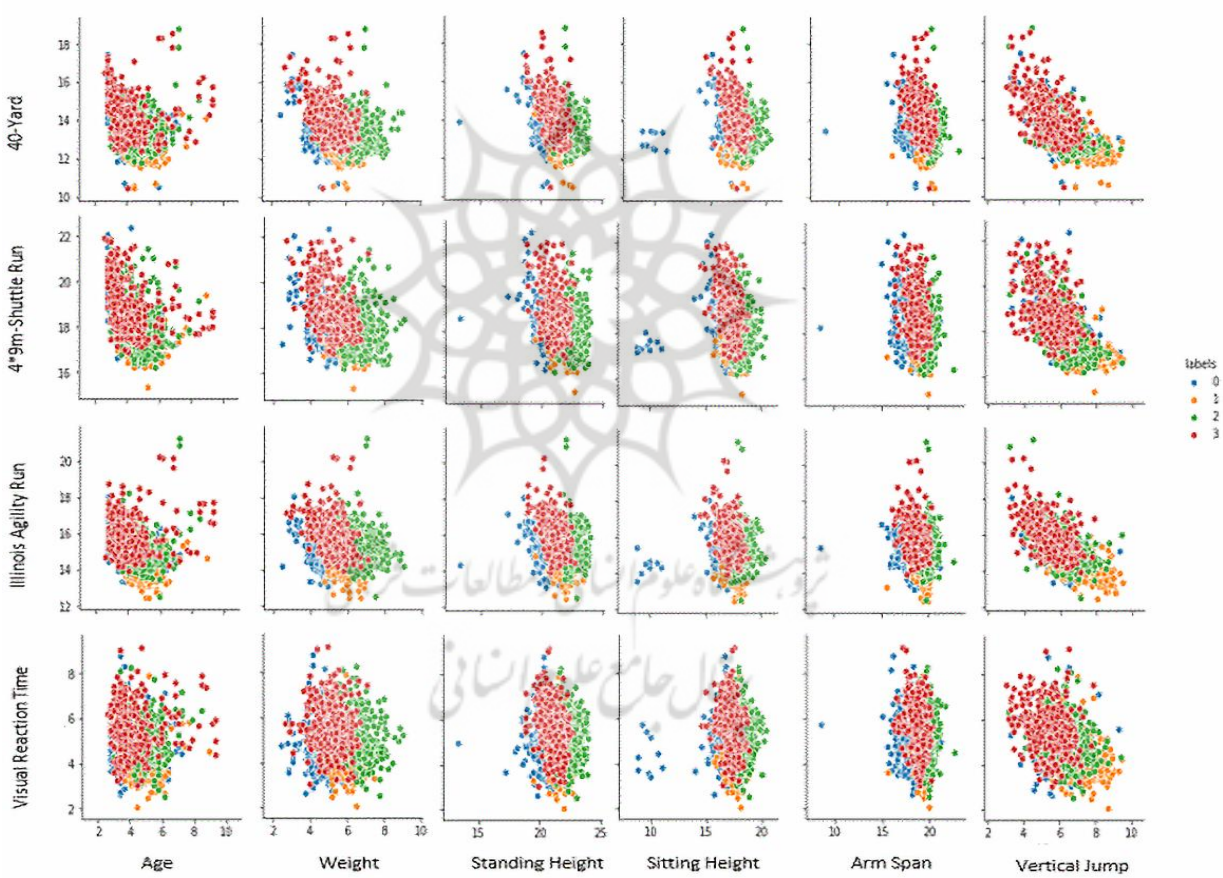

Figure 3. Clustering of the male dataset using SOM-spectral clustering algorithm based on the most practical features in calculating the Fast ICA algorithm's first and second principal components.

The clustering operation was performed through the 10fold CV, and the results of the efficiency of training and validation performances are recorded in Table 2. To record the test performance, 20% of the dataset was selected for testing ten times, and the average of these test performances was recorded as the test result to make the result more reliable. As shown in Table 2, the SOM-k-means algorithm provides the best performance in the female dataset, along with Fast ICA feature extraction methods and the Bayesian Ridge missing value estimator, among other performances. Accordingly, the average test efficiency for the two parameters, Silhouette and Davies-Bouldin, is calculated to be almost 80% and 20%, respectively. For the male dataset, the SOM-spectral

clustering algorithm with the same feature extraction and missing value strategy revealed the average test efficiency for Silhouette as 79% and for Davies-Bouldin as 34%.

Table 2. Clustering evaluation results of female and male datasets

	Feature Extraction	Clustering Algorithms	Cluster NO.	Missing Value Estimation technique	Ave. Training efficiency		Ave. Validation efficiency		Ave. Test efficiency (20% of total data)	
					Silhouette	Davies-Bouldin	Silhouette	Davies-Bouldin	Silhouette	Davies-Bouldin
Female taekwondo athletes' dataset	**Fast ICA**	**SOM-K-means**	**4**	**Bayesian Ridge**	**0.69**	**0.23**	**0.68**	**0.49**	**0.80**	**0.30**
	Fast ICA	SOM-spectral Clustering	4	Bayesian Ridge	0.69	0.49	0.70	0.48	0.78	0.38
	LDA	SOM-K-means	4	Decision Tree Regressor	0.73	0.51	0.69	0.51	0.81	0.37
	LDA	SOM-hierarchical	4	Decision Tree Regressor	0.68	0.52	0.68	0.51	0.80	0.45
	SVD	SOM-k-means	4	K Neighbors Regressor	0.71	0.48	0.73	0.48	0.70	0.40
Male taekwondo athletes' dataset	**Fast ICA**	SOM-K-means	4	Bayesian Ridge	0.67	0.54	0.66	0.54	0.78	0.35
	Fast ICA	**SOM-spectral clustering**	**4**	**Bayesian Ridge**	**0.67**	**0.49**	**0.67**	**0.51**	**0.79**	**0.34**
	Fast ICA	SOM-hierarchical	4	Bayesian Ridge	0.66	0.51	0.67	0.50	0.78	0.37
	LDA	SOM-K-means	4	Decision Tree Regressor	0.70	0.64	0.69	0.54	0.69	0.56
	LDA	SOM-spectral clustering	4	Decision Tree Regressor	0.66	0.54	0.67	0.55	0.69	0.46
	SVD	SOM-K-means	4	K Neighbors Regressor	0.68	0.52	0.68	0.52	0.69	0.52

In the female dataset, Figure 4, the left picture (a) shows the implementation of the CPLE (SVM) algorithm to predict the gold medal winners (label 1) in all labeled data (from silver and bronze medals (label 0)). The middle picture shows (b) the silver medal prediction (label 1) from the gold and bronze medals (label 0) in all labeled data, and the right (c) picture shows the bronze medal prediction (label 1) from the gold and silver medals (label 0) in all labeled data. Also, in the male dataset, Figure 4, the left picture (d) shows the gold medal prediction (label 1) in all labeled data (from silver and bronze medals (label 0)). The middle picture (e) shows the silver medal prediction (label 1) from the gold and bronze medals (label 0) in all labeled data. The right picture (f) shows the prediction of the bronze medal (label 1) from the gold and silver medals (label 0) in all labeled data. One of the proper methods to assess the performance of a classifier algorithm is the Receiver Operating Characteristic (ROC). The graphical plot of the valid positive rate draws the ROC curve) against false positive rate, which refers to the number of cells that are correctly classified and the number of incorrect ones. AUC is a numerical value between zero and one. Three ROC curves in Figure 5 represent females' medal-winning prediction accuracy. For the gold medal (a), the AUC is 0.68. For the silver medal (b), the AUC is 0.61. For the bronze medal, the AUC is 0.68. The ROC curve of the CPLE (SVM) classifier for medal-winning prediction of the male taekwondo athletes is

also plotted. Three ROC curves representing males' medal-winning prediction accuracy. For the gold medal (d), the AUC is 0.58. For the silver medal (e), the accuracy is 0.61; for the bronze medal, the AUC is 0.54 (f).

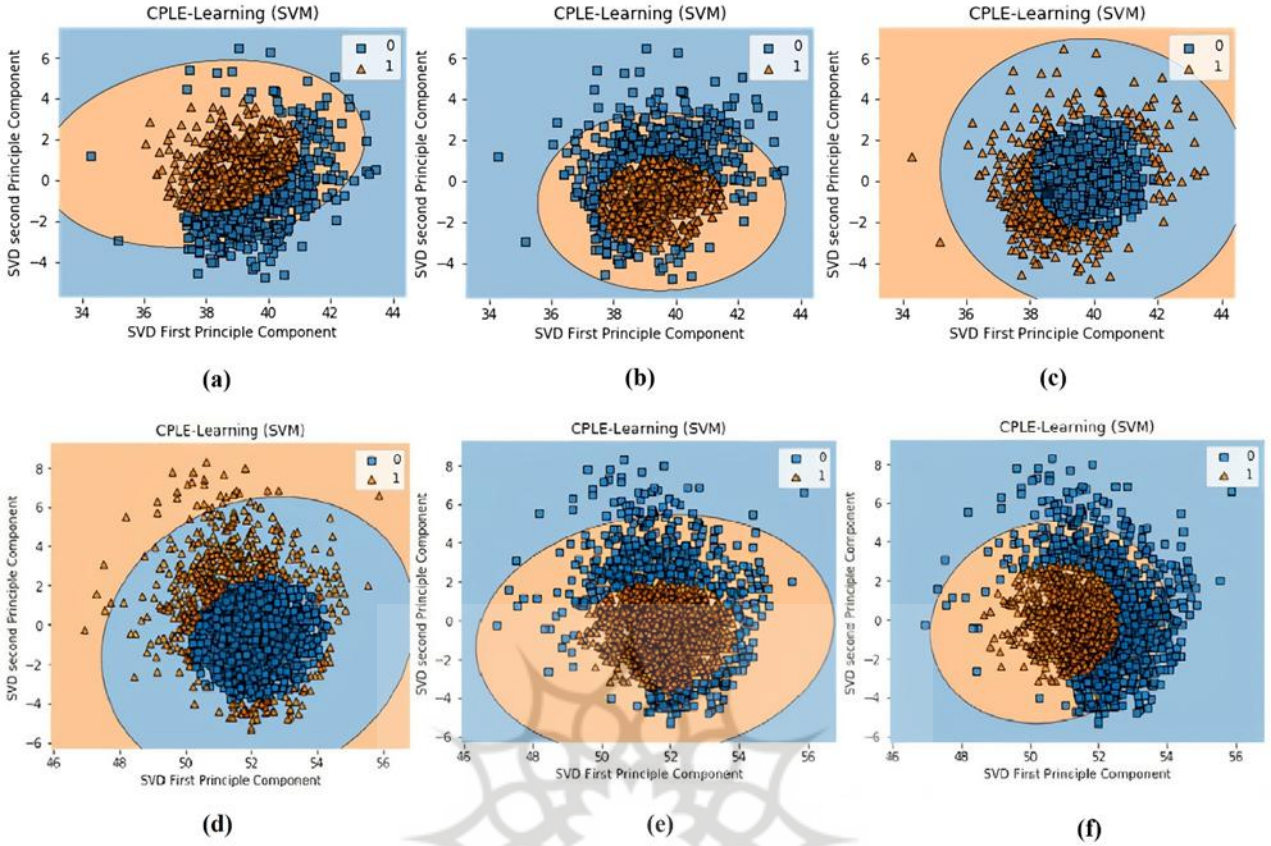

Figure 4. Medal prediction of female taekwondo athletes based on CPLE-Learning algorithm (SVM) for gold medal (a), Silver medal (b), and bronze medal (c). Medal prediction of male taekwondo athletes based on CPLE-Learning algorithm (SVM) for gold medal (d), Silver medal (e), and bronze medal (f).

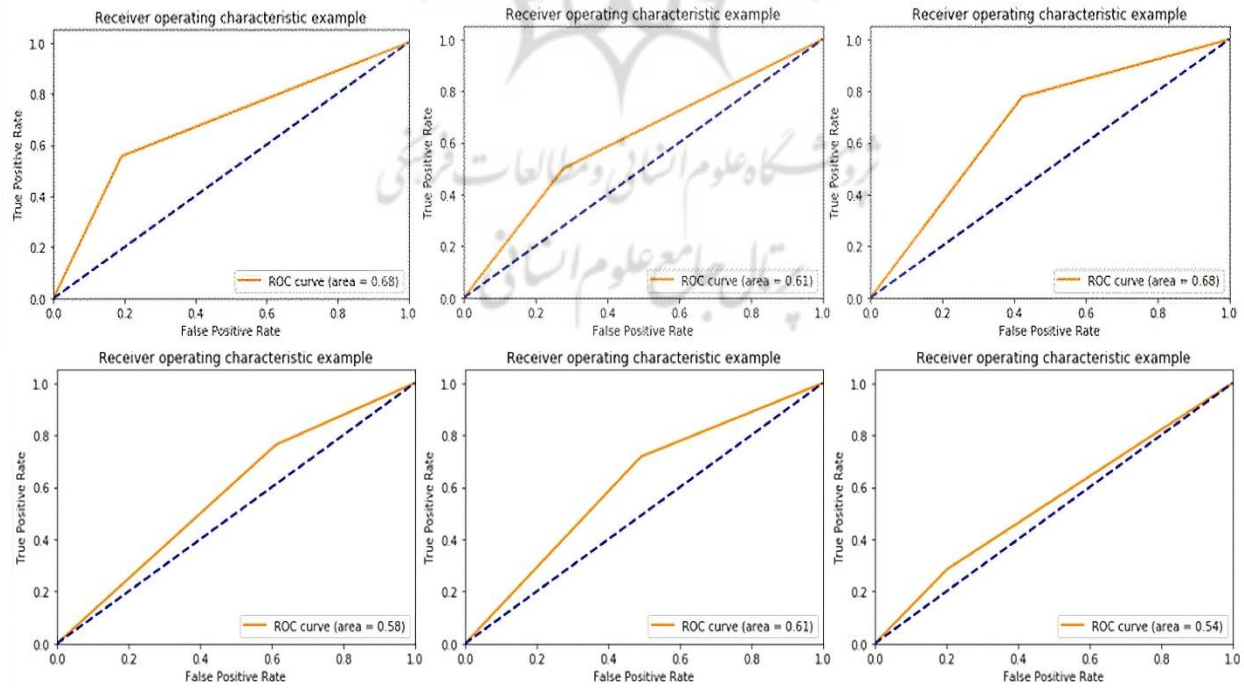

Figure 5. In the female dataset, three receivers operating characteristic curves (ROC) are plotted for the gold medal (a), silver medal (b), and bronze medal (c). In the male data test, three receivers operating characteristic curves (ROC) are shown for the gold medal (d), silver medal (e) and bronze medal (f).

To evaluate the effectiveness of the CPLE (SVM-classifier) model, experiments were conducted to compare its results with Semi Boost, Self-learning, and CPLE (logistic-classifier). The obtained accuracy of different performances, along with NMF, SVD, and PCA feature extraction, are shown in Table 3. As can be seen, in both female and male datasets, the best result is outperformed by the CPLE (SVM) algorithm with SVD feature extraction.

Table 3. Classification evaluation results of female and male taekwondo athletes' datasets

	Methods	Gold prediction	Silver prediction	Bronze prediction	Feature extraction
Female dataset	Semi Boost	70%	75%	34%	NMF
	CPLE (SVM)	68%	61%	68%	SVD
	CPLE (SVM)	68%	59%	73%	NMF
	Self-learning	61%	-	61%	NMF
Male dataset	Semi Boost	50%	56%	-	SVD
	CPLE (SVM)	58%	61%	54%	SVD
	CPLE (SVM)	50%	45%	63%	PCA
	Self-learning	53%	-	-	PCA

Discussion

To gain deeper insight into the performance of taekwondo athletes and provide objective information through a range of performance data, we used two descriptive and predictive analytical tools. Our modeling is designed for sports data that are generally decentralized and scattered. In cluster modeling, it considers the dispersion of data with almost 50% missing values and its extent during 20 years of collection. The combination of SOM and k-means and spectral clustering algorithms demonstrates its ability to reduce the dimensions of the data in the obtained dataset and cluster taekwondo athletes into four groups ranging from excellent to poor performance. The strength of the athletes' performances was based on the number of achieved medals in each cluster. Then, we found the anthropometric and physical fitness variables that most frequently resulted in successful athletic performance.

Because of the results, we observed that the speed index measured by the 40-yard test was one of the most critical characteristics in the sport performance of male and female taekwondo athletes. We also found that male athletes who were younger, lighter-weight, and taller, with a low to medium range of vertical jump, could run more quickly. Although female athletes achieved lower scores on the 40-yard test, it was still essential to their technical success. Another critical indicator was agility, measured by the 4*9-m shuttle run (for both male and female athletes) and the Illinois agility test (for male athletes only). Agility combines speed and coordination, as it measures athletes' physical ability to change direction quickly without disrupting their balance (Fachrezzy et al., 2021). In our work, younger, lighter male athletes with better standing and sitting height demonstrated greater agility than female athletes who were younger, taller, stronger in grip strength, and of average weight. Speed in taekwondo athletes requires agility, skill, and reaction time (Arabacı et al., 2010). Therefore, another essential feature in the performance of male and female taekwondo athletes in this study was reaction time, measured by the visual reaction time (VRT) test. This is a meaningful way to study information processing speed and coordinated environmental response for taekwondo athletes. Both the nature and analysis of taekwondo competitions show that controlling the change of one's body direction in the shortest time and with the best execution technique should be done with maximum power and speed. This combination leads to explosive leg power. The less time it takes athletes' legs to open and the more power and speed they apply to their opponent, the more superior they appear (Mirmohammadi, 2017). In this regard, the anaerobic ability of taekwondo practitioners is essential, as longer pauses of oxygen uptake are required to

attack the opponent more quickly. Furthermore, in our findings, the anaerobic strength index measured by the vertical jump test is only effective in the performance of male athletes. The same phenomenon happened in the previous study (Formalioni et al., 2020) in which men, compared to women, performed higher vertical jumps with greater anaerobic power.

Strength is another determining factor, as force is a powerful taekwondo movement that can be performed in the lower extremities to exert powerful blows (Chiodo et al., 2011). However, we have observed it to be an effective indicator of sports performance by the grip strength test only in the group of female athletes. In the cited study (Markovic et al., 2007), successful taekwondo players have greater explosive leg power than less successful ones, and their aerobic endurance and lateral agility are significantly higher. Moreover, they run faster and have a considerably higher anaerobic threshold than their less successful counterparts. Although the aerobic endurance index has not been seen as effective in athletes' sports performance in our study, the findings of our study confirm the results of previous ones.

Among anthropometric features, age was identified as one of the influential factors in the sports performance of both male and female athletes. In this study, we observed that almost all younger athletes were placed in the category of best performance. Male taekwondo practitioners in the excellent performance cluster generally weigh less than female athletes. It should be noted that the fat index was identified only in the female group as an essential indicator of sports performance. Gender is a determining factor in the percentage of fat accumulation in athletes. At the elite level, adult male taekwondo athletes have a lower rate of fat than females (Pieter & Falcó, 2011). We also identified the height index as an important factor that significantly impacted the success of both male and female athletes' performances. Generally, the cited research (Markovic et al., 2007), which examined the physical and psychological traits that make an elite taekwondo athlete more successful, confirmed these findings. They noted that these physical characteristics included high speed, explosive power, anaerobic power, agility, lower fat percentage, and height. Also, the research of Arazı et al. (2016) and Kazemi et al. (2006), which examined the athletic performances of female taekwondo athletes, confirmed our results.

In addition, we present a semi-supervised approach that yields relevant results without adequate labeled data in the given dataset. Our results demonstrate the superior performance of CPLE learning, one of the lesser-known algorithms in semi-supervised methods, with an SVM classifier to predict gold, silver, and bronze classes in competitions. Indeed, CPLE learning with the SVM classifier has excellent potential for future applications of semi-supervised learning in sports performance predictions. Since taekwondo is a high-performance sport and requires superior technical, tactical, physiological, and psychological skills and capabilities (Bridge et al., 2013), it isn't easy to establish a robust model for predicting taekwondo performance using only physical and anthropometric variables.

For practical applications, this research can be grouped with other pioneering studies to contribute to the martial arts community in developing and applying the scientific method for elite sports performance analysis. Such insights can help create an intelligent system to model taekwondo players' performance by first establishing a knowledge base for them. As competitions are categorized by age, sex, and rank, evaluating the physical fitness of taekwondo athletes based on these parameters can enhance their sporting profile and deepen their understanding of the sport's demands. One application of this modeling is that a sports coach or analyst can assess whether a new athlete is fit for the Olympics or World Championships by using training algorithms that incorporate their physical fitness data, which helps determine the athlete's cluster and predict their chances of victory. Analyzing past data and conducting complex analyses with machine learning algorithms can assist the National Olympic Committee and the National Taekwondo Committee better allocate resources to win medals. Furthermore, sports experts and coaches can design appropriate training programs to help athletes acquire essential skills and enhance their competitiveness.

Ethical Considerations

Compliance with ethical guidelines
All authors observed compliance with ethical guidelines.

Funding
No external funding was used for this work.

Authors' contribution
All authors have contributed to the design and implementation of this study.

Conflicts of interest
There is no conflict of interest.

Acknowledgments
We appreciate the National Olympic Committee and Taekwondo Federation of Iran for supporting this research and providing us with a database of Iranian taekwondo athletes' physical features.

References

Arabacı, R., Görgülü, R., & Çatıkkaş, F. (2010). Relationship Between Agility and Reaction Time, Speed and Body Mass Index in Taekwondo Athletes. *Sport Sciences*, *5*(2), 71-77. https://dergipark.org.tr/en/pub/nwsaspor/issue/20141/213818

Arazı, H., Hosseınzadeh, Z., & Izadı, M. (2016). Relationship between anthropometric, physiological and physical characteristics with success of female taekwondo athletes. *Turkish Journal of Sport and Exercise*, *18*(2), 69-75. https://doi.org/10.15314/tjse.94871

Bouhlel, E., Jouini, A., Gmada, N., Nefzi, A., Ben Abdallah, K., & Tabka, Z. (2006). Heart rate and blood lactate responses during Taekwondo training and competition. *Science & Sports*, *21*(5), 285-290. https://doi.org/10.1016/j.scispo.2006.08.003

Bridge, C., McNaughton, L. R., Close, G. L., & Drust, B. (2013). Taekwondo exercise protocols do not recreate the physiological responses of championship combat. *International journal of sports medicine*, *34*(07), 573-581. https://doi.org/10.1055/s-0032-1327578

Bridge, C. A., Jones, M. A., & Drust, B. (2009). Physiological Responses and Perceived Exertion During International Taekwondo Competition. *International Journal of Sports Physiology and Performance*, *4*(4), 485-493. https://doi.org/10.1123/ijspp.4.4.485

Bunker, R. P., & Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied Computing and Informatics*, *15*(1), 27-33. https://doi.org/10.1016/j.aci.2017.09.005

Cao, C. (2012). *Sports data mining technology used in basketball outcome prediction* [Master, Technological University Dublin]. Dublin, Ireland. https://arrow.tudublin.ie/scschcomdis/39/

Casolino, E., Cortis, C., Lupo, C., Chiodo, S., Minganti, C., & Capranica, L. (2012). Physiological Versus Psychological Evaluation in Taekwondo Elite Athletes. *International Journal of Sports Physiology and Performance*, *7*(4), 322-331. https://doi.org/10.1123/ijspp.7.4.322

Chiodo, S., Tessitore, A., Cortis, C., Cibelli, G., Lupo, C., Ammendolia, A., De Rosas, M., & Capranica, L. (2011). Stress-related hormonal and psychological changes to official youth Taekwondo competitions. *Scandinavian Journal of Medicine & Science in Sports*, *21*(1), 111-119. https://doi.org/10.1111/j.1600-0838.2009.01046.x

Croft, H., Lamb, P., & Middlemas, S. (2015). The application of self-organising maps to performance analysis data in rugby union. *International Journal of Performance Analysis in Sport*, *15*(3), 1037-1046. https://doi.org/10.1080/24748668.2015.11868849

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis* (5 ed.). John Wiley & Sons. https://www.wiley.com/en-jp/Cluster+Analysis%2C+5th+Edition-p-9780470749913

Fachrezzy, F., Maslikah, U., Safadilla, E., Reginald, R., & Hendarto, S. (2021). Physical fitness of the poomsae taekwondo athletes in terms of agility, balance and endurance. *Kinestetik: Jurnal Ilmiah Pendidikan Jasmani*, *5*(1), 111-119. https://doi.org/10.33369/jk.v5i1.14364

Formalioni, A., Antunez, B., Del Vecchio, F., Cabistany, L., Coswig, V., Letieri, R., & Fukuda, D. (2020). Anthropometric characteristics and physical performance of taekwondo athletes. *Revista Brasileira de Cineantropometria e Desempenho Humano*, *22*(3), e55697. https://doi.org/10.1590/1980-0037.2020v22e55697

Gopi, A. P., Jyothi, R. N. S., Narayana, V. L., & Sandeep, K. S. (2023). Classification of tweets data based on polarity using improved RBF kernel of SVM. *International Journal of Information Technology*, *15*(2), 965-980. https://doi.org/10.1007/s41870-019-00409-4

Gu, W., Foster, K., Shang, J., & Wei, L. (2019). A game-predicting expert system using big data and machine learning. *Expert Systems with Applications*, *130*(3), 293-305. https://doi.org/10.1016/j.eswa.2019.04.025

Izadyar, M., Memari, Z., & Mousavi, M.-H. (2016). Pricing Equation for Iranian Premier League Football Players. *Journal of Economic Research (Tahghighat- E- Eghtesadi)*, *51*(1), 25-40. https://doi.org/10.22059/jte.2016.57595

Kazemi, M., Waalen, J., Morgan, C., & White, A. R. (2006). A profile of olympic taekwondo competitors. *Journal of Sports Science and Medicine*, *5*, 114-121. https://www.jssm.org/jssm-05-CSSI1-114.xml%3Eabst

Kohara, K., & Enomoto, S. (2018). Clustering Professional Baseball Players with SOM and Deciding Team Reinforcement Strategy with AHP. In P. Perner (Ed.), *Advances in Data Mining. Applications and Theoretical Aspects* (pp. 135-147). Springer International Publishing. https://doi.org/10.1007/978-3-319-95786-9_10

Kostakis, O., Tatti, N., & Gionis, A. (2017). Discovering recurring activity in temporal networks. *Data Mining and Knowledge Discovery*, *31*(6), 1840-1871. https://doi.org/10.1007/s10618-017-0515-0

Markovic, G., Jukic, I., Milanovic, D., & Metikos, D. (2007). Effects of Sprint and Plyometric Training on Muscle Function and Athletic Performance. *The Journal of Strength & Conditioning Research*, *21*(2), 543-549. https://journals.lww.com/nsca-jscr/fulltext/2007/05000/effects_of_sprint_and_plyometric_training_on.44.aspx

Memari, z., Hoda, K., & Safaie, A. (2020). The Valuation of Football Players with Data Mining Technique (Case Study: Esteghlal Club). *Sport Management Journal*, *12*(3), 735-757. https://doi.org/10.22059/jsm.2019.262922.2128

Mendes-Neves, T., Meireles, L., & Mendes-Moreira, J. (2024). Towards a foundation large events model for soccer. *Machine Learning*, *113*(11), 8687-8709. https://doi.org/10.1007/s10994-024-06606-y

Mirmohammadi, S. (2017). Comparison of Selected physiological and Physical Fitness characteristics of Professional Women taekwondo athletes in Kiurogi and Poomsae Styles. *Journal of Sport and Exercise Physiology*, *10*(2), 47-58. https://doi.org/10.48308/joeppa.2017.98882

Musa, R., Abdul Majeed, A. P. P., Taha, Z., Siow-Wee, C., Nasir, A. F. A., & Abdullah, M. (2019). A machine learning approach of predicting high potential archers by means of physical fitness indicators. *PLOS ONE*, *14*(1), e0209638. https://doi.org/10.1371/journal.pone.0209638

O'Donoghue, P. (2010). *Research Methods for Sports Performance Analysis*. Routledge. https://books.google.com/books/about/Research_Methods_for_Sports_Performance.html?id=6sxH9b7NmEEC&source=kp_book_description

Park, Y. H., Park, Y. H., & Gerrard, J. (2009). *Tae Kwon Do: The Ultimate Reference Guide to the World's Most Popular Martial Art*. Infobase Publishing. https://books.google.com/books?id=gONCfB_XxyAC

Pelechrinis, K., & Papalexakis, E. (2018, February 5-9). *Athlytics: Winning in Sports with Data* [Conference session]. Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, Marina Del Rey, CA, USA. https://doi.org/10.1145/3159652.3162005

Pieter, W., & Falcó, C. (2011). Skinfold patterning in elite Spanish and American junior taekwondo-in. *Journal of Martial Arts Anthropology*, *11*(1), 47-51. https://www.researchgate.net/publication/237101466_Skinfold_Patterning_in_Elite_Spanish_and_American_Junior_Taekwondo-in

Sarlis, V., & Tjortjis, C. (2020). Sports analytics — Evaluation of basketball players and team performance. *Information Systems*, *93*, 101562. https://doi.org/10.1016/j.is.2020.101562

Taha, Z., Musa, R. M., Abdul Majeed, A. P. P., Abdullah, M. R., Zakaria, M. A., Alim, M. M., Jizat, J. A. M., & Ibrahim, M. F. (2018, December 7-8). *The identification of high potential archers based on relative psychological coping skills variables: A Support Vector Machine approach* [Conference session]. The 4th Asia Pacific Conference on Manufacturing Systems and the 3rd International Manufacturing Engineering Conference Yogyakarta, Indonesia. https://dx.doi.org/10.1088/1757-899X/319/1/012027

Takemura, Y., Oda, K., & Ono, M. (2018). Analysis of Team Relationship using Self-Organizing Map for University Volleyball Players. *Journal of Robotics, Networking and Artificial Life*, *5*(3), 199-203. https://doi.org/10.2991/jrnal.2018.5.3.12

Takemura, Y., Yokoyama, M., Omori, S., & Shimosak, R. (2014, October 27). *Development of SOM algorithm for relationship between roles and individual's role adaptation in rugby* [Conference session]. 2014 World Automation Congress (WAC), Waikoloa, HI, USA. https://doi.org/10.1109/WAC.2014.6935638

Tichy, W. (2016). Changing the Game: Dr. Dave Schrader on sports analytics. *Ubiquity*, *2016*(May), 1-10. https://doi.org/10.1145/2933230

Yeung, C., Bunker, R., Umemoto, R., & Fujii, K. (2024). Evaluating soccer match prediction models: a deep learning approach and feature optimization for gradient-boosted trees. *Machine Learning*, *113*(10), 7541-7564. https://doi.org/10.1007/s10994-024-06608-w

Zahradník, D., & Korvas, P. (2018). *The Introduction into Sports Training*. Masaryk University. https://ecuni.publi.cz/en/book/52-the-introduction-into-sports-training

Zheng, H., Hatachi, T., Masuda, M., Aoki, K., & Kato, C. (2020). Psychological Analysis of Rugby Players Based on Self-Organizing Map (SOM) and Ward's Method. *Advances in Physical Education*, *10*(4), 410-420. https://doi.org/10.4236/ape.2020.104033

Ziv, G., & Lidor, R. (2010). Vertical jump in female and male basketball players—A review of observational and experimental studies. *Journal of Science and Medicine in Sport*, *13*(3), 332-339. https://doi.org/10.1016/j.jsams.2009.02.009