Available online at www.scholarsresearchlibrary.com

Scholars Research Library

European Journal of Sports & Exercise Science, 2019, 7 (2): 1-9
(http://www.scholarsresearchlibrary.com)

ISSN:2278–005X

Applying Machine Learning Techniques to Advance Anti-Doping

Tyler Kelly, Adam Beharry*, Matthew Fedoruk

United States Anti-Doping Agency, Colorado Springs, CO, USA

Corresponding Author: Dr. Adam Beharry, United States Anti-Doping Agency, 5555 Tech Center Drive, Suite 200, Colorado Springs, CO, 80920, USA, E-mail: abeharry@usada.org

ABSTRACT

Globally exists an ongoing battle between increasingly advanced doping methods and limited resources available to anti-doping organizations. Therefore, the developments of new tools for identifying athletes who may be doping are needed. Recognizing which athletes are at the highest risk of doping allows an anti-doping organization to distribute those limited resources in the most effective manner. Presented below is a comparison of multiple machines and statistical learning approaches, combined with resampling techniques, to identify which athletes are at the highest risk of doping. The results presented indicate that support vector classification and logistic regression, combined with oversampling, may provide an effective tool to aid anti-doping organizations in most effectively distributing scarce resources. Adoption and implementation of these methods may both enhance the deterrence effect of anti-doping, as well as increases the likelihood of catching athletes doping. Future research should be conducted to explore the feasibility of combining these performance-based measures with biological measures such as the Athlete Biological Passport to enhance anti-doping efforts.

Keywords: Anti-doping, Performance modeling, Machine learning, Mixed-martial-arts, Sports.

Abbreviations: ABP: Athlete Biological Passport; AUC: Area Under Curve; FN: False Negative; FP: False Positive; GNB: Gaussian Naïve Bayes; MNB: Multinomial Naïve Bayes; NPV: Negative Predictive Value; PED: Performance Enhancing Drug; PPV: Positive Predictive Value; RF: Random Forest; SVM: Support Vector Machine; TN: True Negative; TP: True Positive; USADA: United States Anti-Doping Agency

INTRODUCTION

In the world of sport, the use of banned Performance-Enhancing Drugs (PEDs) or methods is referred to as doping. Driven by the goal of obtaining an advantage over their competitor, some athletes look to artificially enhance their abilities. The estimated prevalence of doping in sport can be above 40% depending upon variables such as sport, country, and competition level [1-6]. However, only 1% to 2% of doping samples collected annually test positive for a banned substance (WADA 2012-2016). Thus, improved strategies that offer more robust, objective, and effective means of detection are needed.

Early PED testing methods relied heavily on the direct detection of the banned substance(s) in the biological sample collected from the athlete. However, current methods have shifted towards a more forensic approach with the implementation of the ABP [7,8]. The ABP measures multiple biomarkers that change in response to the use of a PED, which can indirectly reveal doping has occurred. Although the ABP has shown to be a successful deterrent and strategic testing tool, others have shown confounding factors (genetic polymorphisms and medications) or delivering PEDs in micro-doses can mask changes in biomarkers measured by the ABP [9-13]. Therefore, new tools for identifying athletes who may be doping are needed.

One emerging strategy, similar to that of the ABP, is to incorporate the tracking of athlete performance as another indirect marker of doping. Indeed, performance measures of many Olympic sports are well studied, with elite athlete's performance varying less than 2.0% (0.6% to 1.4%) in sprint and endurance sports like running and cycling [14]. Additionally, elite athletes in weightlifting and field events, which require more explosive power in a single effort, exhibit performance varying less than 4.0% (1.4% to 3.3%) [14]. Thus, the development of monitoring changes in athlete performance may indeed be another indirect marker of doping as recently demonstrated [15-17]. Furthermore, creating a mathematical performance model to measure critical power has been proposed as a sensitive method to detect performance modifying manipulations such as PED use [17]. The utilization of such modeling techniques has not been fully explored in the field of anti-doping but has been used effectively to better understand diseases. In fact, similar machine-learning algorithms have been shown to improve with the detection and diagnosis of many life-threatening diseases [18-22]. Thus, the implementation of machine-learning based classification models or neural networks may aid in the detection of PED use.

The goal of this paper was to develop and compare multiple classification models using career performance data from athletes who have been sanctioned for doping. For this purpose, five different supervised classification models were utilized: Support Vector Machine [23], Random Forest [24], Multinomial Naïve Bayes [25], Logistic Regression [26], and Gaussian Naïve Bayes [25]. Statistical analyses were also done to examine the performance of each model. In addition, sanctioned athlete's performances were evaluated to gain a better understanding of their fighting profile. Presented here is a novel method, using machine learning approaches, to better identify athletes who may have used PEDs.

MATERIALS AND METHODS

UFC athletes tested by the US Anti-Doping Agency (USADA) from July 2015 through May 2018 were included the dataset. For these athletes, career performance data was obtained for athletes who had competed in a minimum of one fight in the UFC from January 2015 to May 2018. Given the relatively low population, female athletes were not included in this analysis. Athletes were identified as having engaged in doping if they had publicly received a sanction for a doping-related offense from USADA.

All athlete data was obtained from the mixed martial arts data aggregation website Fightmetric.com. The performance data below was utilized in addition to fighting time, and athlete age. To determine if any differences existed between this data from non-sanctioned athletes and sanctioned athlete, unpaired t-test with Welch's correction was performed using GraphPad Prism version 7.03, GraphPad Software, La Jolla California USA.

Performance data

- Longest average fight time measured in seconds; minimum 5 UFC fights
- Significant striking accuracy; minimum 5 UFC fights and 350 significant strike attempts
- Significant strike defense (the % of opponent's strikes that did not land); minimum 5 UFC fights and 350 significant strike attempts
- Takedown accuracy; minimum 5 UFC fights and 20 takedown attempts
- Takedown defense (the % of opponents TD attempts that did not land); minimum 5 UFC fights and 20 takedown attempts by opponents
- Takedowns landed
- Knockdowns landed
- Submission average per 15 minutes; minimum 5 UFC fights
- Significant strikes landed per Minute; minimum 5 UFC fights
- Significant strikes absorbed per Minute; minimum 5 UFC fights

Machine-learning based classification models used include Support Vector Machine (SVM), Random Forest (RF), Multinomial Naïve Bayes (MNB), Logistic Regression, and Gaussian Naïve Bayes (GNB). All classifiers were trained using Python [27] version 3.6.5) and the scikit-learn (version 0.19.1) as previously described [28]. A parameter search was done to optimize each model to accurately predict sanctioned athletes [29,30]. More specifically, a coarse grid search using 5-fold cross-validation was performed via the GridSearchCV method of the scikit-learn library utilizing a 50/50 train/test split [28]. Parameter performance was evaluated based on the F1 score, where 1 is the best and 0 is the worst. The linear SVM model was trained with the scikit-learn SVM method using: the penalty parameter set at 0.1, a linear kernel, class weight inversely proportional to the frequency of occurrence,

and predictor variables not rescaled or normalized. The RF model was trained using the RandomForestClassifier method of the scikit-learn ensemble module while changing the size of the forest from 10 trees to 100. For the MNB model, predictor variables were first transformed from continuous variables to categorical variables utilizing the LabelEncoder method of the scikit-learn preprocessing module prior to training utilizing the MultinomialNB method of the scikit-learn naïve bayes module. The LR model was trained using the scikit-learn LogisticRegression method with L2 regularization, Newton Cost Gradient solver, fit with a y-intercept, and class weight adjusted to be inversely proportional to the frequency of occurrence. However, prior to training, predictor variables were rescaled and normalized to ensure each predictor variable had a mean of 0 and a standard deviation of 1 using the StandardScaler method of scikit-learn. For the GNB model, predictor variables were transformed utilizing the same method for the MNB model, and then rescaled and normalized using the method previously described for the Logistic Regression classifier. The GNB model was trained using the default values of the GaussianNB method in the scikit-learn naïve bayes module.

The presented data set was comprised primarily of non-sanctioned athlete performance data (majority class) and a small percentage of sanctioned athlete performance data (minority class) in a roughly 96/4 split. Therefore, the Tomek links undersampling [31] technique and Synthetic Minority Oversampling Technique (SMOTE) [32] were utilized to achieve better classifier performance. Briefly, the Tomek links technique removes cases of the majority class when those cases are highly similar to cases in the minority class, thus enhancing the boundary between classes. Tomek links were applied to training sets using the Tomek links method of the imblearn package [33]. Default parameters were used with the following two exceptions: ratio set to the majority and return indices set as true. Inversely, SMOTE is a method of addressing the class imbalance in classification problems by creating synthetic examples of the minority class. SMOTE was applied to training data sets via the SMOTE method of the imblearn package [33]. Default parameters were used with the following two exceptions: ratio set to minority and kind set to borderline1.

Measures used to evaluate the performance of each classification model include: Area under Curve (AUC), F1, Sensitivity, Specificity, Accuracy, True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN), Positive Predictive Value (PPV), and Negative Predictive Value (NPV) for each model permutation. The measures of accuracy, sensitivity, specificity, TP, TN, FP, FN, PPV, and NPV were generated using the confusion matrix as previously outlined [34-36]. Evaluation metrics averaged across 5-fold cross-validation for all models are presented in Table 1. The number of predicted doping athletes was found by predicting the probability of class membership for each example. Athletes were classified as doping if the probability of class membership was greater than 0.50. The cross-validated model of each permutation was used to calculate probability.

Table 1: Altering Sampling Size: Comparison of sanctioned athlete classification performance of four classification models utilizing either no sampling alteration, Tomek links undersampling, SMOTE oversampling, or both; Model performance metrics reported include Area under Curve (AUC), F1 score, sensitivity, specificity, accuracy, positive predictive value (PPV), negative predictive value (NPV), false positive rate (FP), false negative rate (FN), true positive rate (TP), and true negative rate (TN)

	Tomek	SMOTE	AUC	FI	Sensitivity	Accuracy	PPV	NPV	FP	FN	TP	TN
SVC	Yes	Yes	0.610	0.097	0.596	0.480	0.053	0.961	79.200	3.000	4.400	71.600
	Yes	No	0.560	0.000	0.000	0.953	0.000	0.953	0.000	7.400	0.000	150.800
	No	Yes	0.570	0.092	0.575	0.485	0.051	0.960	78.200	3.200	4.200	72.600
	No	No	0.530	0.000	0.000	0.953	0.000	0.953	0.000	7.400	0.000	150.800
Random forest	Yes	Yes	0.650	0.126	0.107	0.928	0.148	0.957	4.600	6.600	0.800	146.200
	Yes	No	0.660	0.178	0.107	0.953	1.000	0.958	0.000	6.600	0.800	150.800
	No	Yes	0.660	0.171	0.164	0.930	0.200	0.959	4.800	6.200	1.200	146.200
	No	No	0.660	0.183	0.111	0.953	1.000	0.958	0.000	6.600	0.800	150.800
Multinomial naive bayes	Yes	Yes	0.510	0.063	0.293	0.591	0.036	0.945	59.600	5.200	2.200	91.200
	Yes	No	0.480	0.000	0.000	0.948	0.000	0.953	0.800	7.400	0.000	150.000
	No	Yes	0.510	0.073	0.354	0.598	0.042	0.951	58.800	4.800	2.600	92.000

	No	No	0.490	0.000	0.000	0.946	0.000	0.953	1.200	7.400	0.000	149.600
Logistic regression	Yes	Yes	0.580	0.111	0.407	0.696	0.064	0.961	43.600	4.400	3.000	107.200
	Yes	No	0.620	0.104	0.468	0.633	0.059	0.961	54.000	4.000	3.400	96.800
	No	Yes	0.600	0.127	0.489	0.685	0.073	0.965	46.000	3.800	3.600	104.800
	No	No	0.600	0.105	0.482	0.625	0.061	0.962	55.600	3.800	3.600	95.200
Gaussian naive bayes	Yes	Yes	0.580	0.124	0.436	0.731	0.077	0.964	38.400	4.200	3.200	112.400
	Yes	No	0.670	0.000	0.000	0.951	0.000	0.953	0.400	7.400	0.000	150.400
	No	Yes	0.580	0.121	0.400	0.727	0.072	0.962	38.800	4.400	3.000	112.000
	No	No	0.660	0.000	0.000	0.953	0.000	0.943	0.000	7.400	0.000	122.000

RESULTS

As shown in Table 2, UFC athletes sanctioned for doping are significantly older (32.6 years old) compared to athletes who have not been sanctioned (30.9 years old). Additionally, of the nine performance measures, sanctioned athletes have a significantly greater takedown accuracy. Although takedown accuracy was the only performance measure to be significantly different, similar performance measures such as takedown defense and takedowns landed were higher in sanctioned athletes. To better understand if sanctioned athletes exhibit a similar "fighting strategy," the percent difference was calculated for each of the nine performance measures as well as fight time and age. Figure 1 illustrates sanctioned UFC athletes tend to excel in offensive and defensive takedown-based movements. Furthermore, sanctioned athletes tend to land more significant strikes per minute but knock their opponent down less and attempt fewer submissions. These data suggest sanctioned UFC athletes may employ a similar fighting strategy, best described as "ground-and-pound," to fighting an opponent.

Table 2: UFC performance markers, analysis of nine performance metrics, fight time, and age from non-sanctioned UFC athletes compared to UFC athletes sanctioned by USADA; data are presented as means ± SD; * significantly different ($p<0.05$)

UFC performance markers	Non-Sanctioned (n=754)	Sanctioned (n=37)	p-value
Age	30.9 ± 4.26	32.6 ± 4.23	0.021*
Longest average fight time	636.0 ± 200.7	642.1 ± 144.9	0.808
Significant strike accuracy	0.44 ± 0.09	0.44 ± 0.12	0.739
Significant strike defense	0.56 ± 0.09	0.59 ± 0.11	0.099
Takedown accuracy	0.37 ± 0.24	0.50 ± 0.29	0.014*
Takedown defense	0.58 ± 0.27	0.67 ± 0.31	0.107
Submission average	0.69 ± 1.2	0.63 ± 0.76	0.693
Takedowns landed	1.60 ± 1.5	I.79 ± 1.49	0.449
Knockdowns landed	0.50 ± 2.4	0.45 ± 0.45	0.656
Strikes landed per minute	3.40 ± 2.4	3.71 ± 1.7	0.309
Strikes absorbed per minute	3.34 ± 1.4	3.28 ± 2.1	0.88

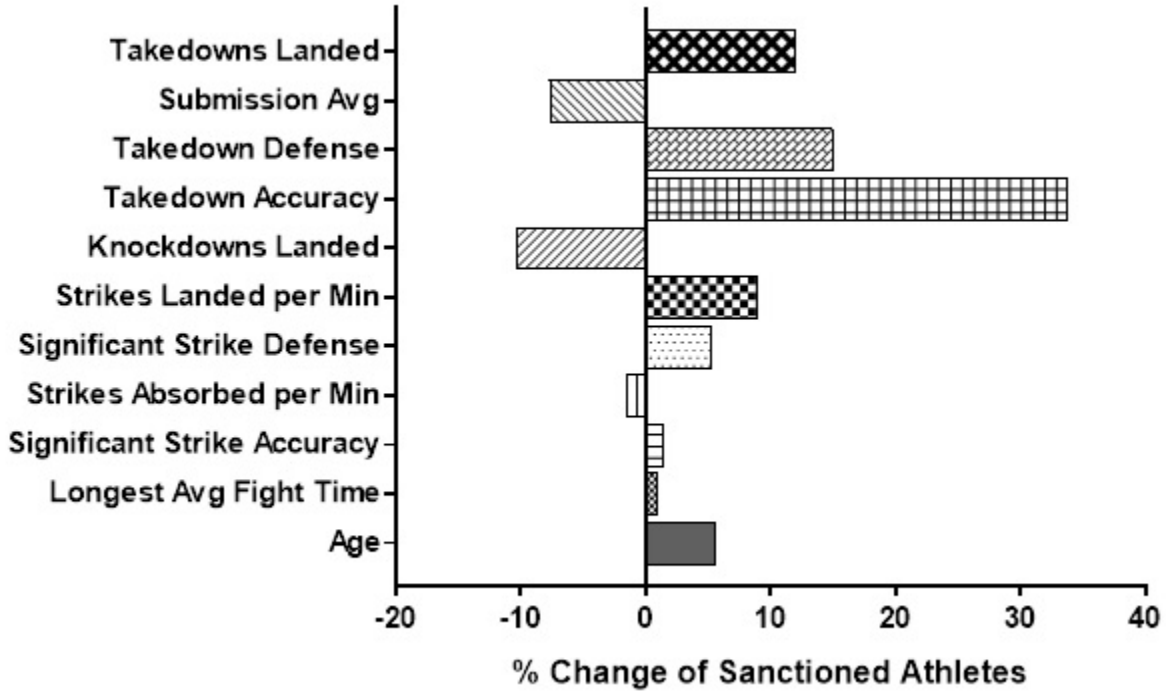

Figure 1: Sanctioned Athletes Performance Markers; The percent change in nine performance metrics, fight time, and age from non-sanctioned UFC athletes compared to UFC athletes sanctioned by USADA; Data are presented as the mean percent change of sanctioned athletes

The performance of each classification model with each permutation is shown in Table 1.

Many of the model permutations may not be useful anti-doping tools based on their inability to discriminate between sanctioned and non-sanctioned athletes. For example, any model that failed to identify a single TP can be discarded as this means they were not able to identify a sanctioned athlete. In addition, any model permutation which has accuracy greater than 90% may also not be useful for anti-doping purposes, as these models are likely simply classifying all examples as not sanctioned. Similar to accuracy, models with specificity greater than 95% may also be a misleading performance measure. This is primarily due to the imbalanced nature of the dataset, where a model which simply classifies all examples as the majority class (non-sanctioned) may still be highly accurate and specific despite having no discriminatory value. Models with an accuracy and specificity greater than 90% proved unable to classify more than 1 athlete as a TP. Therefore, it may be appropriate to eliminate any models with a low TP rate and high accuracy and specificity since they are not able to differentiate between classes. All RF models and SVC and Naïve Bayes models which did not utilize SMOTE over-sampling can be ruled out as suitable anti-doping tools based on their inability to distinguish between doping and non-doping athletes.

Due to class imbalance, some metrics may not be the most appropriate for anti-doping organizations to consider when evaluating model performance while utilizing specific model and sampling alteration combinations may be more appropriate. For each RF, SVC and Naïve Bayes classification models, over-sampling of the minority class resulted in improved sensitivity, PPV, FP rate, and TP rate when compared to either under-sampling of the majority class alone or no sampling alterations. Therefore, it could be concluded with this data set, classification models not implementing over-sampling of the minority class, via SMOTE, are unable to distinguish between classes. However, for each LR algorithm, this trend did not hold true as each model permutation performed similarly with only TN rate increasing with oversampling. SVC with under- and over-sampling or only over-sampling, any permutation of LR, and Naïve Bayes models with combined under- and over-sampling may be most appropriate for anti-doping purposes. As shown in Table 1, these models displayed sensitivity greater than 25%, specificity and accuracy greater than 45%, a TN rate less than 6, and were able to correctly classify more than 1 athlete in the minority class.

As an additional measure to evaluate each model’s performance, the number of athletes classified in the minority class (doping) was evaluated. As shown in Table 3, all models except RF were able to classify more than 25% of

athletes as doping with at least two model permutations. More specifically, RF models were only able to classify less than 5% of athletes as doping while only Naïve Bayes models utilizing SMOTE oversampling were able to classify more than 25% of athletes as doping. Multinomial Naïve Bayes with oversampling was able to classify roughly 10% more athletes as doping than GNB with oversampling. However, SVC and LR were able to classify 28% or more athletes as doping regardless if under- or over-sampling was applied. In fact, utilizing either SMOTE or Tomek did not appear beneficial for LR regarding classifying athletes as doping.

Table 3: Classification model performance: Evaluation of four classification models utilizing either no sampling alteration, Tomek links undersampling, SMOTE oversampling, or both to classify UFC athletes as doping; Data are presented as the number classified or a percentage of the total sampled population (n=791)

Models	Tomek	SMOTE	# of Athlete Classified as Doping	% of Athlete Population
SVC	Yes	Yes	318	40.20%
	Yes	No	287	36.30%
	No	Yes	306	38.70%
	No	No	304	38.40%
Random Forest	Yes	Yes	35	4.42%
	Yes	No	30	3.79%
	No	Yes	36	4.55%
	No	No	30	3.79%
Multinomial naive bayes	Yes	Yes	301	38.10%
	Yes	No	0	0.00%
	No	Yes	311	39.30%
	No	No	6	0.76%
Logistic regression	Yes	Yes	234	29.60%
	Yes	No	277	35.00%
	No	Yes	223	28.20%
	No	No	306	38.70%
Gaussian naive bays	Yes	Yes	221	27.90%
	Yes	No	1	0.13%
	No	Yes	212	26.80%
	No	No	0	0.00%

DISCUSSION AND CONCLUSION

Strategically testing athletes is critical for both maximizing the detection of PED use and increasing deterrence. Additionally, to maximize limited anti-doping resources, utilizing a tool which may help identify athletes using PEDs will enhance anti-doping efforts. Although a growing number of sources compile data on an athlete’s performance, leveraging this information for anti-doping purposes is still being explored. The concept of longitudinally monitoring an athlete’s performance for abnormalities has been proposed as another way to indirectly detect the use of PEDs [15-17,37]. Again, since the main objective of using a PED is to alter one’s performance, identifying atypical competition results may indeed be a valuable tool for anti-doping programs. Presented here is a machine-learning based approach that compared different classification models for identifying athletes who may have used PEDs based on career performance.

Regardless of the application, any classification model should be chosen based on its performance. However, for the purposes of anti-doping, the most important metrics on how to evaluate the performance of a classification model

may be those which most directly relate to the minority class (athletes sanctioned for doping). For example, accuracy may be an inappropriate evaluation metric given the large imbalance in the dataset presented here or that exists globally (WADA 2012-2016). This imbalance can result in the failure of a model to identify any athletes sanctioned for doping but still result in a misleadingly high accuracy value. Therefore, similar to accuracy, specificity may also be an unsuitable metric to measure. Specificity indicates the ratio of TNs to TNs and FPs combined, thus a low value for specificity indicates a higher degree of appropriateness of a model for anti-doping purposes. Although a large specificity value indicates a low number of FPs in comparison to the number of TNs, a relatively high FP rate may be "desirable" for anti-doping purposes, as athletes in this category may be at a higher risk for doping despite having not been sanctioned.

Three classification model performance metrics which may be valuable for anti-doping purposes are sensitivity, FP rate, and FN rate. The sensitivity indicates the ratio of TPs to TPs and FNs combined, thus the higher the sensitivity, the more likely this model will correctly identify athletes who have used a banned substance. As previously mentioned, a high FP rate, or type 1 error, may not be problematic in the anti-doping setting. In this context, FP rate in a classification task represents athletes who have not tested positive for a prohibited substance, but who have a high likelihood of testing positive based on similar performances to those athletes who have been sanctioned. Therefore, employing classification models that have a large number of false positives may be one way for anti-doping organizations to identify athletes in need of additional testing. However, this is in contrast with type 2 errors, or FN rate, which represents athletes who have been sanctioned for using a prohibited method or substance but are not identified as such. For anti-doping organizations, classification models with a high FN rate would have the consequence of incorrectly classifying an athlete using a prohibited substance as not. This poses a danger for a combat sport like the UFC as it could allow a doped athlete into the Octagon and lead to serious injury to the opposing fighter.

Another possible performance metric to evaluate each classification model would be the number of athletes classified as doping from each model permutation. Previous work using various surveys has estimated doping prevalence in sport to be anywhere from 4% to more than 45% depending on the sport [1-6]. Similarly, at least one model permutation from each different classification model was able to classify between 3% and 40% of the athlete population as doping. Even though these are very different methods to assess doping in sport, the agreements suggest a current methodology for selecting athletes for biological testing may need improvement. Therefore, developing a classification model which incorporates an athlete's performance and biological markers may be one method to optimally employ anti-doping resources.

Resources, expertise, and access to data on an athlete's performance may be some of the limiting factors when trying to investigate if the presented concept could be applied to other sports. One limitation in this study was the use of career performance across all fights and not individual performances each fight. By using career performances, the assumption is being made the athlete has used a PED for a large portion of their career which may not be the case. Future studies should explore if individual performances improve classification models ability to identify doped athletes. An additional limitation of this study not explored is the role athletic equipment and apparel plays in an athlete's performance. For example, a difference of ~ 1 m in the shot put performance may be suggestive of doping [37]. However, a runner's shoe choice could alter their peak vertical ground reaction force, step frequency, and ground contact time, all translating into an improved running velocity by ~ 3.4%. Thus, whatever methodology used to monitor an athlete's performance for anti-doping purposes may need to be sport and discipline-specific to account for many of the confounding factors.

In summary, multiple machine learning based classification models were investigated to determine how athlete performance could be used to identify the use of banned performance-enhancing drugs. To address the imbalance found in the dataset, each model examined if sampling alterations could improve each model's ability to differentiate sanctioned and non-sanctioned athletes. Under-sampling of the majority class via Tomek links did not improve model performance on test sets, while over-sampling of the minority class via SMOTE improved performance in all algorithms compared to no sampling alterations [38-40]. Based on these findings, as well as others, the creation of an athlete's performance passport may indeed advance anti-doping efforts. The integration of performance measures with existing biological markers (ABP), investigating more sophisticated machine learning methods, and considering the implementation of more sophisticated over-sampling techniques are all ways to better identify athletes who may have used a banned PEDs.

REFERENCES

[1] Dietz, P., et al., Associations between physical and cognitive doping-a cross-sectional study in 2.997 triathletes. *PLoS One*, **2013**. 8(11).

[2] Pitsch, W., and Emrich, E., The frequency of doping in elite sport: Results of a replication study. *International Review for the Sociology of Sport*, **2012**. 47(5): p. 559-580.

[3] Plessner, H., and Musch, J., Wie verbreitet ist Doping im Leistungssport? Eine www-Umfrage mit Hilfe der Randomized-Response-Technik. *Expertise in Sport*, **2002**: p. 78-79.

[4] Schröter, H., et al., A comparison of the cheater detection and the unrelated question models: A randomized response survey on physical and cognitive doping in recreational triathletes. *PloS One*, **2016**. 11(5).

[5] Striegel, H., Ulrich, R., and Simon, P., Randomized response estimates for doping and illicit drug use in elite athletes. *Drug and Alcohol Dependence*, **2010**. 106(2-3): p. 230-232.

[6] Ulrich, R., et al., Doping in two elite athletics competitions assessed by randomized-response surveys. *Sports Medicine*, **2010**. 48(1): p. 211-219.

[7] Robinson, N., et al., The athlete biological passport: an effective tool in the fight against doping. *Clinical Chemistry*, **2010**. 57: p. 830-832.

[8] Sottas, P.E., et al., The athlete biological passport. *Clinical Chemistry*, **2011**. 57(7): p. 969-976.

[9] Ashenden, M., et al., Current markers of the athlete blood passport do not flag microdose EPO doping. *European Journal of Applied Physiology*, **2011**. 111(9): p. 2307-2314.

[10] Mullen, J., et al., Pregnancy greatly affects the steroidal module of the athlete biological passport. *Drug Testing and Analysis*, **2018**. 10(7): p. 1070-1075.

[11] Mullen, J.E., et al., Urinary steroid profile in females‐the impact of menstrual cycle and emergency contraceptives. *Drug Testing and Analysis*, **2017**. 9(7): p. 1034-1042.

[12] Schulze, J.J., et al., The impact of genetics and hormonal contraceptives on the steroid profile in female athletes. *Frontiers in Endocrinology*, **2014**. 5: p. 50.

[13] Zorzoli, M., and Rossi, F., Implementation of the biological passport: the experience of the International Cycling Union. *Drug Testing and Analysis*, **2010**. 2(11‐12): p. 542-547.

[14] Malcata, R.M., and Hopkins, W.G., Variability of competitive performance of elite athletes: A systematic review. *Sports Medicine*, **2014**. 44(12): p. 1763-1774.

[15] Hopker, J., et al., Athlete performance monitoring in anti-doping. *Frontiers in Physiology*, **2018**. 9: p. 232.

[16] Iljukov, S., Bermon, S., and Schumacher, Y.O., Application of the athlete's performance passport for doping control: a case report. *Frontiers in Physiology*, **2018**. 9: p. 280.

[17] Puchowicz, M.J., et al., The critical power model as a potential tool for anti-doping. *Frontiers in Physiology*, **2018**. 9: p. 643.

[18] Cuingnet, R., et al., Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. *Neuroimage*, **2011**. 56(2): p. 766-781.

[19] Esmaeily, H., et al., Comparing three data mining algorithms for identifying the associated risk factors of type 2 diabetes. *Iranian Biomedical Journal*, **2018**. 22(5): p. 303.

[20] Kadah, Y.M., et al., Classification algorithms for quantitative tissue characterization of diffuse liver disease from ultrasound images. *IEEE Transactions on Medical Imaging*, **1996**. 15(4): p. 466-478.

[21] Ning, K., et al., Classifying Alzheimer's disease with brain imaging and genetic data using a neural network framework. *Neurobiology of Aging*, **2018**. 68: p. 151-158.

[22] Radhakrishnan, A., et al., Machine learning for nuclear mechano-morphometric biomarkers in cancer diagnosis. *Scientific Reports*, **2017**. 7(1): p. 17946.

[23] Cortes, C., and Vapnik, V., Support-vector networks. *Machine Learning*, **1995**. 20(3): p. 273-297.

[24] Breiman, L., Random forests. *Machine Learning*, **2001**. 45(1): p. 5-32.

[25] Lowd, D., and Domingos, P., Naive Bayes models for probability estimation. In Proceedings of the 22nd international conference on Machine learning, **2005**. p. 529-536.

[26] Hosmer, D.W., Lemeshow, S., and Sturdivant, R.X., Applied logistic regression. New York, NY, USA: Wiley, **1939**.

[27] Van Rossum, G., and Drake, F.L., Python Reference Manual, PythonLabs, Virginia, USA, **2001**.

[28] Pedregosa, F., et al., Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **2011**. 12: p. 2825-2830.

[29] Chang, C.C., LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, **2011**. 2(3): p. 27.

[30] Hsu, C.W., Chang, C.C., and Lin, C.J., A practical guide to support vector classification. **2003**.

[31] Elhassan, T., and Aljurf, M., Classification of imbalance data using tomek link (T-Link) Combined with random under-sampling (RUS) as a data reduction method. *Journal of Informatics and Data Mining*, **2016**. 1(2): p. 1-12.

[32] Chawla, N.V., et al., SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, **2002**. 16: p. 321-357.

[33] Lemaître, G., Nogueira, F., and Aridas, C.K., Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research*, **2017**. 18(1): p. 559-563.

[34] Provost, F., and Kohavi, R., Glossary of terms. *Journal of Machine Learning*, **1998**. 30(2-3): p. 271-274.

[35] Tang, Y., et al., SVMs modeling for highly imbalanced classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, **2009**. 39(1): p. 281-288.

[36] Powers, D.M., Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*. **2011**. 2(1): p. 37-63.

[37] Montagna, S., and Hopker, J., A Bayesian approach for the use of athlete performance data within anti-doping. *Frontiers in Physiology*, **2018**. 9: p. 884

[38] Batista, G.E., Prati, R.C., and Monard, M.C., A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, **2004**. 6(1): p. 20-29.

[39] Bowers, L.D., and Bigard, X., Achievements and challenges in anti-doping research. *In Acute Topics in Anti-Doping*, **2017**. 62: p. 77-90

[40] Awaisu, A., et al., Instructional design and assessment of an elective course on the use of drugs in sport. *Currents in Pharmacy Teaching and Learning*, **2018**. 10(8): p. 1124-1131.