Detection of Erythropoietin in Blood to Uncover Doping in Sports using Machine Learning

Maxx Richard Rahman
Saarland University
Germany
m.rahman@iss.uni-saarland.de

Jacob Bejder
University of Copenhagen
Denmark
jbr@nexs.ku.dk

Thomas Christian Bonne
University of Copenhagen
Denmark
tbonne@nexs.ku.dk

Andreas Breenfeldt Andersen
University of Copenhagen
Denmark
anan@nexs.ku.dk

Jesús Rodríguez Huertas
University of Granada
Spain
jhuertas@ugr.es

Reid Aikin
World Anti-Doping Agency
Canada
reid.aikin@wada-ama.org

Nikolai Baastrup Nordsborg
University of Copenhagen
Denmark
nbn@nexs.ku.dk

Wolfgang Maass
Saarland University
Germany
wolfgang.maass@iss.uni-saarland.de

2022 IEEE International Conference on Digital Health (ICDH) | 978-1-6654-8149-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICDH55609.2022.00038

***Abstract*—Sports officials around the world are facing challenges due to the unfair nature of doping practices used by unscrupulous athletes to improve their performance. This practice includes blood transfusion, intake of anabolic steroids or even hormone-based drugs like erythropoietin to increase their strength, endurance, and ultimately their performance. While direct detection and identification of erythropoietin in blood samples of athletes have proven an effective means to uncover doping, not all the cases are easily detectable, and some analyses are too costly to be carried out on every sample. This leads to a need to develop an indirect method for detecting erythropoietin in blood samples based on different blood biomarkers. In this paper, we presented a comparison of different machine learning algorithms combined with statistical analysis approaches to identify the presence of erythropoietin drug in blood samples collected at both sea level and moderate altitude. The results presented indicate that ensemble methods like random forest and XGboost algorithms may provide an effective tool to aid anti-doping organisations in most effectively distributing scarce resources. Implementation of these methods on the samples from elite athletes may both enhance the deterrence effect of anti-doping as well as increases the likelihood of catching doped athletes.**

***Index Terms*—Erythropoietin, Blood Doping, Machine Learning, Drug Abuse, rhEPO, Sports**

I. Introduction

Artificial Intelligence (AI) has shown potential improvement in the sports industry, whether to identify players' unique talents, detect previous injuries, or even assist decision-making. Automated Sports Journalism is a good example where AI is being used in guiding sports journalism [Galily 2018]. However, the applications are not only limited to the development of sports but can also be used to ensure fairness in sports by athletes.

Athletes have a desire to increase their physical performance to obtain better results which leads some of them to seek alternative or even prohibited methods. Therefore, doping practices in sports have been around for several decades. Blood doping can be performed by mainly three methods: intake of erythropoietin, synthetic oxygen carriers to enhance the oxygen transport capacity and blood transfusion [Jelkmann 2016]. Erythropoietin (EPO) is a peptide hormone naturally secreted by the kidney to stimulate the production of red blood cells in the blood. It increases the blood capacity to transport oxygen which results in increasing of body endurance [Jelkmann 2016]. One way to naturally increase the production of EPO is through altitude training. The body compensates for the reduced oxygen concentration at high altitude by releasing EPO. However, several synthetically produced substances can stimulate endogenous EPO production, like recombinant human erythropoietin (rhEPO). In collaboration with its stakeholders, the World Anti-Doping Agency (WADA) oversees the list of substances that are prohibited in sport [WADA 2021]. One such substance is rhEPO, a recombinant protein that stimulates erythropoiesis which increases the oxygen-carrying capacity of the blood [John et al. 2012]. While the laboratory-based method exists for the detection of rhEPO, it is too expensive and time-consuming to be applied to all blood samples [Martin at al. 2021]. Moreover, the sensitivity and specificity of the method are also a concern in some regards. These limitations led us to investigate whether an AI-based approach like machine learning algorithms could be applied to laboratory results already being generated in anti-doping in order to better direct rhEPO analysis. Since rhEPO intake produces characteristic changes in haematological parameters, it is possible to authenticate athletes based on indirect indicators of blood doping.

In this paper, we start by reviewing the literature on indirect detection methods with an emphasis on statistical methods. Then, we present the procedure of the clinical experiment we conducted to collect the data and analyse it by using different statistical methods. We found the potential biomarkers of rhEPO, which were used to perform machine learning analysis to identify the presence of rhEPO in blood samples. Finally, we show the performance of the trained algorithms and discuss the results with possible future research.

978-1-6654-8149-6/22/$31.00 ©2022 IEEE
DOI 10.1109/ICDH55609.2022.00038

Authorized licensed use limited to: Saarl Universitaets. Downloaded on October 13,2022 at 09:34:35 UTC from IEEE Xplore. Restrictions apply.

II. Related Work

While the earliest attempts of indirect methods of detecting blood doping included so-called "no start rules", where athletes with blood parameters outside of population-based limits were prohibited from competing, these were prone to an unacceptable number of false positives due to athletes with naturally elevated blood parameters. Therefore, approaches were soon taken to personalise decision limits to the individual's own biomarker values [Sharpe et al. 2006], [Malcovati et al. 2003].

[Manfredini et al. 2011] proposed the statistical-based score parameter calculated from different blood parameters by considering the shift of the value from their baseline values. [Sharpe et al. 2006] developed a statistical approach to estimate athlete's baseline values from just one prior sample. [Parisotto et al. 2001] showed in their work that how different statistical models (ON and OFF-model) based on the potential parameters behaved in their study.

There are also some studies performed using machine learning algorithms in anti-doping analysis. [Kelly et al. 2019] compared the performance of different machine learning algorithms to identify the risk of doping among 791 UFC athletes based on their performance data and reported their best results as a sensitivity of 44%. [Sottas et al. 2006] introduced the Abnormal Blood Profile Score (ABPS), which is an indirect test based on the statistical classification of indirect biomarkers. The calculation of ABPS is based on Support Vector Machine and Naive Bayes algorithm that achieved a sensitivity of 45% at 100% specificity. Therefore, this is the current state-of-the-art (SOTA) method which is used as a baseline in this paper.

Since the direct detection methods like IEF-PAGE analysis [Martin at al. 2021] are expensive and time consuming to perform and therefore, there is a necessity for an indirect method. The existing work on indirect methods is limited to some applications in anti-doping analysis (except blood doping) using statistical analysis and some machine learning algorithms. This shows why there is a need to explore data-driven approaches in blood doping analysis. In this paper, we performed a study to compare different machine learning algorithms combined with statistical analysis approaches to identify the presence of rhEPO in blood samples to uncover blood doping.

III. Clinical Experiment

A. Goal Definition

The goal of this study is to develop an indirect method that can detect the doping practices performed by athletes. In other words, a model that can detect the presence of doping substances rhEPO in the blood sample given by the athlete i.e., the model should be able to differentiate between a clean and suspicious blood sample and triggers in the case of a suspicious sample. So, we performed a data-driven approach, i.e., conducted clinical experiment, performed exploratory analysis, applied different machine learning algorithms and compared their performance to develop such a model. In addition, a comprehensive analysis of blood samples was conducted to understand the underlying principles of different biomarkers.

B. Data collection

In reality, it is not easy to gather such health-related data of elite athletes because of data availability, privacy and other issues. Therefore, we set up such an experiment to mimic the real-life situation and analysed the collected data from the participants in this study. We performed a clinical experiment for 12 weeks with two arms: "sea-level" (34 participants) and "altitude" (39 participants). The unequal experimental design between the sea-level and altitude arms was a cost-benefit decision based on economical reasons. The time span of 12 weeks was divided into three periods: baseline (week 1-4), intervention (week 5-8) and follow-up (week 8-12). In the experiment, the baseline and follow-up periods of both the arms were performed at sea-level, whereas the four weeks of intervention period were performed at either sea-level or a moderate altitude of 2300m. The choice to have a four-week intervention period is based on current practice in the athletic population, where altitude training camps are rarely longer than four weeks.

None of the participants was given any doping substituent in the baseline and follow-up periods, whereas in the intervention period, 11 injections were given to all the participants after every other day. 25 participants were given rhEPO injections, and 9 participants were given placebo injections in the sea-level arm. In the altitude arm, 12 participants were injected with rhEPO, and 27 were given placebo injections. Every participant was monitored regularly, and the blood sample of each participant was collected every week. So, we collected 864 blood samples in total, and the detail of data statistics is summarised in Table I.

TABLE I
Number of blood samples collected at sea-level and altitude

Blood samples	Sea-level	Altitude (=2300m)
Controlled samples (Placebo)	609	107
rhEPO samples	100	48
Total samples	709	155

For each blood sample, the haematological profile is quantified. The haematological profile consists of a set of haematological parameters that show significant changes in their values due to rhEPO intake. These haematological parameters are haemoglobin concentration (HB), haematocrit (HCT), reticulocytes percentage (RET%), reticulocytes count (RET#), reticulocytes haemoglobin (RET-HB), mean corpuscular volume (MCV), mean corpuscular haemoglobin mass (MCH), mean corpuscular haemoglobin concentration (MCHC), red blood cell count (RBC), red blood cell distribution width - standard deviation (RDW-SD), red blood cell distribution width - coefficient of variation (RDW-CV), white blood cell count (WBC), immature reticulocyte fraction (IRF), low fluorescence reticulocyte fraction (LFR), medium fluorescence reticulocyte

Authorized licensed use limited to: Saarl Universitaets. Downloaded on October 13,2022 at 09:34:35 UTC from IEEE Xplore. Restrictions apply.

fraction (MFR), high fluorescence reticulocyte fraction (HFR) [Parisotto et al. 2001], [Zorzoli 2011]. In addition, we have the OFF-HR score (OFF-HR) which tells us the relationship of reticulocytes to haemoglobin and can be calculated using the below expression [Gore et al. 2003]:

$$OFF - HR = HB(g/L) - 60 * \sqrt{RET\%}$$

The importance of OFF-HR parameter can be understood by an example. Let us consider a scenario where an athlete is doping with small doses of rhEPO. This might not result in a significant increase in haemoglobin, but reticulocytes will likely react significantly, causing OFF-score to be affected. Similarly, an athlete could also take large doses of rhEPO. If this athlete successfully keeps haemoglobin constant through a plasma increase, it would go undetected. However, infusing a blood bag will certainly decrease reticulocytes to the extent that would trigger the OFF-HR. So, OFF-HR can indicate the acceleration or deceleration of erythropoiesis process, and therefore, it is an important parameter to consider.

C. Data preprocessing

We found that the data contained missing values for some haematological parameters in a few samples. This could be mainly due to the measurement error when collecting and analysing the blood samples in the laboratory. Since we have less number of samples in the data, we decided to impute the values instead of discarding that sample. We used a median imputation strategy where these missing values for the parameter were imputed by the median value of that parameter. The median value of that parameter is calculated by taking all the samples of that participant from the respective arm. This helps to avoid any kind of bias caused due to the values of other participants.

IV. Exploratory Analysis

A. Multivariate Analysis

We performed the multivariate analysis to analyse different haematological parameters and the relationship between them. For that, all the parameters were plotted against each other. In some distributions, we observed certain regions where the majority of the controlled samples lie and form a cluster. Fig. 1 shows three such distributions of RET# *vs* RET%, HFR *vs* RET% and RET# *vs* RDW-SD and regions are marked with a black box. These regions show the baseline distribution of these haematological parameters of the normal human population.

B. Cut-based Method

Based on the multivariate analysis, we formed certain threshold cuts on some parameters to develop a cut-based strategy to distinguish controlled samples from rhEPO samples. Fig. 2 shows the developed thresholds for the samples collected at both sea-level and altitude, and the corresponding plots showing the proportion of samples satisfying these thresholds. It can be observed from the plots that the proportion of samples that do not satisfy the developed thresholds

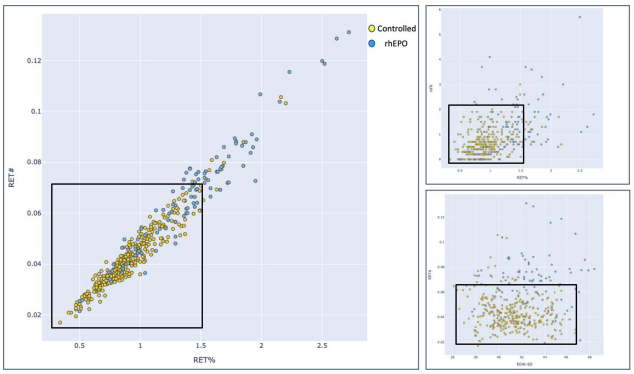

Fig. 1. Distributions of controlled and rhEPO samples collected at sea-level showing the relationship between RET# *vs* RET%, HFR *vs* RET% and RET# *vs* RDW-SD parameters.

contain 93% controlled samples for sea-level and 94% for altitude. Moreover, all the samples that satisfied the threshold cuts for altitude are the rhEPO samples. This shows the efficiency and the significance of these developed cuts on the parameters.

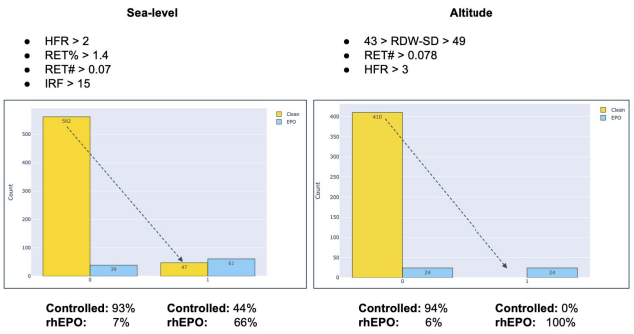

Fig. 2. Developed threshold cuts on haematological parameters for both sea-level and altitude arms and their corresponding plots for the proportion of samples that satisfy these cuts.

C. Statistical Analysis

In addition to exploring each parameter and examining their distribution, we observed some potential strong indicators. Some parameters are susceptible to rhEPO intake and show a significant change in their values, whereas the difference is negligible in other parameters. To quantify, we performed 2 sample Kolmogorov-Smirnov test (K-S test) to identify the best haematological parameters, which show a significant change in their values on rhEPO intake.

The K-S test [Dimitrova et al. 2020] is a standard statistical test for deciding whether a dataset is consistent with another dataset. The maximum difference between the cumulative distribution function of the two population distributions (controlled $F_a(x)$ and rhEPO $F_b(x)$) is calculated by using::

$$D_{a,b} = sup|F_a(x) - F_b(x)|$$

Authorized licensed use limited to: Saarl Universitaets. Downloaded on October 13,2022 at 09:34:35 UTC from IEEE Xplore. Restrictions apply.

The null distribution of this statistic is calculated under the null hypothesis that the two distributions are drawn from the same parent distribution. The null hypothesis is rejected at level α if:

$$D_{a,b} > \sqrt{-\ln\frac{\alpha}{2}.\frac{1+\frac{b}{a}}{2b}}$$

where a and b are the number of controlled and rhEPO samples respectively. We chose α = 0.001 to determine the best indicators at confidence level of 99.99%. Fig. 3 shows the distribution of WBC and RET# for controlled (blue) and rhEPO (red) samples at sea-level. WBC has a p-value of 0.56, which shows that there is no significant change observed in the values of this parameter because of rhEPO intake. On the other hand, the p-value of RET# (0.0001) indicates that it is potentially a strong indicator to observe the effect of rhEPO.

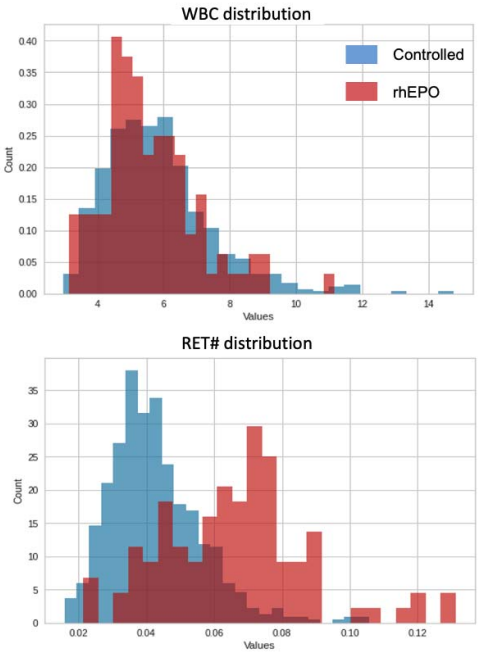

Fig. 3. Distribution of WBC and RET# parameters for controlled (blue) and rhEPO (red) samples.

The skewed distributions of the haematological parameters are statistically described by means of the median, first (IQ1 = 0.25), and third (IQ3 = 0.75) quartile based upon the data. The inter-quartile range between the first and third quartile includes 50% of the samples. Table II and Table III show the detailed descriptive statistics of the haematological parameters of controlled and rhEPO samples, respectively.

D. Variable Selection

Based on the K-S test, we selected a set of potential biomarkers, i.e., haematological parameters that shows a significant change in their values due to the rhEPO intake. The potential biomarkers are RET%, RET#, IFR, LFR, MFR, HFR, RDW-SD and OFF-HR for sea-level and RET#, RET%, RDW-SD, RDW-CV, MCHC and HCT for altitude. Fig. 4 shows the distinguishing power (independent of the model) of all the potential biomarkers for sea-level and altitude. Since machine learning algorithms are sensitive to the choice of the parameters used to train the model, this step is needed. These potential biomarkers are used to perform the machine learning analysis.

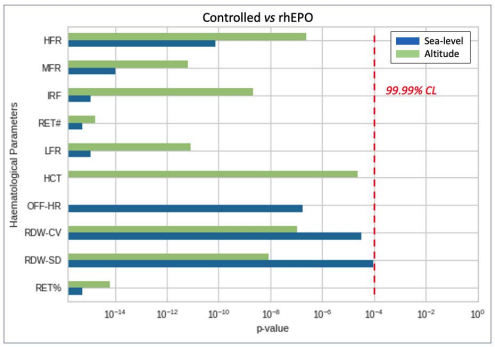

Fig. 4. Potential biomarkers (at 99.99% CL) from the K-S test at sea-level and altitude.

V. Machine Learning Studies

A. Model Selection

In this section, we describe the machine learning studies conducted to perform the classification task. In this analysis, we trained different machine learning algorithms and evaluated their performance on data. The considered algorithms are:

- Logistic Regression (LR) [Cox 1958]
- Naive Bayes (NB) [Zhang 2004]
- Support Vector Machine (SVM) [Hearst et al. 1998]
- K-Nearest Neighbour (KNN) [Mucherino at al. 2009]
- Decision Tree (DT) [Wu at al. 2008]
- Random Forest (RF) [Preiman 2001]
- eXtreme Gradient Boosting - XGBoost (XGB) [Chen et al. 2016]

Table IV shows the different hyperparameters values selected to train each model. These values are considered after performing the optimisation step.

B. Training

We randomly partitioned the data samples such that 80% of the data was used for training and 20% for testing the algorithm. The data contains the potential biomarkers which are selected after performing the K-S test. However, prior to training, parameters were rescaled and normalised to ensure each parameter had a mean of 0 and a standard deviation of 1 using the SCIKIT-LEARN package [Pedregosa et al. 2011].

We implemented all the algorithms using the SCIKIT-LEARN package [Pedregosa et al. 2011], except the XGBoost algorithm, which was implemented using XGBOOST package [Chen et al. 2016]. Over-fitting is a major issue in training the model because it reduces the ability of the algorithm to predict new samples accurately. The best-fitting model for a

Authorized licensed use limited to: Saarl Universitaets. Downloaded on October 13,2022 at 09:34:35 UTC from IEEE Xplore. Restrictions apply.

TABLE II
DESCRIPTIVE STATISTICS OF HAEMATOLOGICAL PARAMETERS FOR RHEPO AND CONTROLLED SAMPLES AT SEA-LEVEL INCLUDING THE p-VALUE FROM THE K-S TEST.

	rhEPO (n=100)						**controlled samples (n=609)**						***p*-value**
Parameter	mean±std.	min	IQ1	median	IQ3	max	mean±std.	min	IQ1	median	IQ3	max	
HB	14.5±1.2	11.4	13.6	14.4	15.4	17.0	14.2±1.1	11.5	13.4	14.3	15.0	17.0	4.1e-02
HCT	42.3±3.3	32.8	40.2	42.1	44.5	51.2	41.2±3.0	33.2	39.3	41.2	43.2	50.1	7.1e-03
RET#	0.1±0.0	0.0	0.1	0.1	0.1	0.1	0.0±0.0	0.0	0.0	0.0	0.1	0.1	1.3e-15
RET%	1.4±0.4	0.5	1.1	1.4	1.7	2.7	0.9±0.3	0.3	0.7	0.9	1.1	2.2	1.3e-15
RET-HB	33.5±1.7	29.4	32.5	33.6	34.5	36.9	33.5±1.8	27.4	32.5	33.5	34.7	38.0	9.7e-01
MCV	89.5±2.9	84.2	87.5	88.9	91.8	96.0	88.9±3.2	81.0	86.7	88.6	91.0	98.1	9.2e-02
MCH	30.6±1.1	28.3	29.6	30.6	31.2	32.8	30.7±1.4	26.7	29.7	30.7	31.7	35.1	7.1e-01
MCHC	34.2±0.8	32.2	33.6	34.1	34.6	36.5	34.5±1.1	32.0	33.7	34.4	35.0	38.9	3.9e-03
RBC	4.7±0.4	3.5	4.5	4.7	4.9	5.7	4.6±0.4	3.6	4.4	4.7	4.9	5.7	9.7e-02
RDW-SD	42.5±2.6	37.7	40.4	42.4	44.4	48.4	41.7±2.6	35.2	40.0	41.5	43.1	54.1	9.2e-04
RDW-CV	12.9±0.6	11.6	12.4	12.9	13.2	14.4	12.7±0.8	11.6	12.1	12.6	13.0	17.0	5.6e-04
WBC	5.6±1.4	3.1	4.6	5.4	6.3	11.1	5.8±1.6	3.0	4.7	5.6	6.7	14.8	5.6e-01
IRF	9.9±3.8	1.1	7.4	9.9	12.3	22.7	6.2±2.7	0.0	4.3	6.0	7.9	15.0	1.3e-15
LFR	90.1±3.8	77.3	87.7	90.1	92.6	98.9	93.7±2.7	85.0	92.1	94.0	95.6	99.3	1.3e-15
MFR	8.6±3.0	1.1	6.7	8.7	10.6	19.0	5.7±2.3	0.7	4.0	5.5	7.1	13.4	1.3e-15
HFR	1.4±0.9	0.0	0.7	1.3	1.8	5.7	0.6±0.5	0.0	0.2	0.5	0.9	3.0	2.6e-12
OFF-HR	74.3±15.7	36.2	63.7	73.0	85.8	111.6	85.0±14.3	44.8	75.5	85.4	94.8	119.1	2.6e-09

TABLE III
DESCRIPTIVE STATISTICS OF HAEMATOLOGICAL PARAMETERS FOR RHEPO AND CONTROLLED SAMPLES AT ALTITUDE INCLUDING THE p-VALUE FROM THE K-S TEST.

	rhEPO (n=48)						**controlled samples (n=107)**						***p*-value**
Parameter	mean±std.	min	IQ1	median	IQ3	max	mean±std.	min	IQ1	median	IQ3	max	
HB	15.1±1.4	12.2	13.6	15.3	16.0	17.6	14.5±1.1	11.9	13.6	14.6	15.5	16.8	3.9e-02
HCT	43.6±3.9	35.5	40.2	44.1	46.5	51.5	41.7±2.8	34.7	39.5	41.8	43.9	47.5	7.3e-04
RET#	0.1±0.0	0.0	0.1	0.1	0.1	0.1	0.1±0.0	0.0	0.1	0.1	0.1	0.1	1.9e-06
RET%	1.7±0.5	0.8	1.4	1.7	2.1	2.8	1.3±0.4	0.5	1.0	1.3	1.6	2.9	6.1e-05
RET-HB	34.7±1.8	30.3	33.6	34.7	35.7	37.7	34.6±1.8	28.4	33.8	34.8	35.8	38.0	9.1e-01
MCV	89.9±2.6	85.4	88.3	89.6	92.0	95.9	87.6±3.3	81.3	85.4	87.3	90.0	96.6	3.2e-02
MCH	31.1±0.9	29.6	30.1	31.1	31.8	32.7	30.5±1.4	27.0	30.0	30.7	31.5	33.1	3.7e-02
MCHC	34.6±0.9	33.3	34.1	34.4	34.7	37.7	34.9±0.9	33.0	34.2	34.9	35.4	37.4	3.4e-04
RBC	4.8±0.4	4.3	4.5	4.9	5.2	5.6	4.8±0.3	3.7	4.5	4.8	5.0	5.4	3.7e-02
RDW-SD	44.3±2.5	39.1	42.5	44.5	45.9	48.8	42.5±3.7	36.6	40.5	41.5	42.7	55.3	1.4e-07
RDW-CV	13.4±0.6	11.8	13.0	13.4	13.9	14.7	13.2±1.2	11.8	12.5	12.9	13.5	17.8	1.5e-04
WBC	7.1±2.3	4.0	5.2	6.7	8.5	14.5	6.4±1.9	3.7	5.0	6.2	7.5	12.8	3.6e-01
IRF	10.3±3.2	4.3	8.4	10.3	11.6	22.8	9.3±3.4	0.0	6.9	9.2	11.5	21.4	1.3e-01
LFR	88.9±3.5	77.2	87.2	89.8	91.1	94.4	91.0±3.0	82.9	88.8	90.9	93.3	97.8	1.5e-02
MFR	9.4±2.3	4.9	8.0	9.2	10.5	15.3	7.9±2.5	2.2	5.9	8.0	9.6	15.5	7.3e-03
HFR	1.7±1.5	0.1	0.9	1.3	1.8	7.5	1.1±0.6	0.0	0.6	1.0	1.6	2.8	2.9e-01
OFF-HR	72.6±18.6	31.9	61.2	73.8	85.2	107.8	77.1±15.5	25.7	68.7	76.1	85.6	111.7	2.9e-02

Authorized licensed use limited to: Saarl Universitaets. Downloaded on October 13,2022 at 09:34:35 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
HYPERPARAMETERS VALUES OF DIFFERENT MACHINE LEARNING MODELS

Model	Parameter value
Logistic Regression (LR)	*penalty* = l2
	max iter = 100
Naive Bayes (NB)	*kernel* = gaussian
	var smoothing = 1e-9
Support Vector Machine (SVM)	*kernel* = rbf
	degree = 3
	max iter = -1
K-Nearest Neighbor (KNN)	*n neighbors* = 5
	power = euclidean distance
	leaf size = 30
Decision Tree (DT)	*criterion* = gini
	min samples split = 2
	max features = no. of features
Random Forest (RF)	*n estimators* = 100
	criterion = gini
	min samples split = 2
	bootstrap = True
XGBoost (XGB)	*objective* = binary logistic
	learning rate = 0.1
	max depth = 5
	alpha = 10
	n estimators = 10

single dataset is very likely to be a worse fit for future data. Since we have a comparatively small dataset for training, it is a major concern in our analysis. Therefore, we performed k-fold cross-validation method [Refaeilzadeh et al. 2009] to train the algorithms where we chose $k = 5$. It is a resampling procedure where we partitioned the training dataset into non-overlapping 5 folds. Each of the folds is used as a held-back validation set, whilst all other folds collectively are used as a training dataset. So, a total of 5 models were trained for each algorithm and evaluated on the 5 holdout validation datasets, and the mean performance is reported. A schematic of training and validation of the algorithm is shown in Fig. 5.

C. Optimisation

Each algorithm consists of a set of hyperparameters that can be tuned to improve the training of the algorithm. Therefore, we need to perform a coarse grid search for finding the best combination of these hyperparameters. We used a hyperparameter optimisation framework to automate hyperparameter search efficiently in large grid space and prune unpromising trials for faster optimisation. We implemented it using the OPTUNA package [Akiba et al. 2019] to improve the results of the trained model on the validation dataset.

After the optimisation of the model is performed, we applied the optimised trained model on the testing set to predict the new unseen samples and evaluated the model's performance by calculating different evaluation metrics.

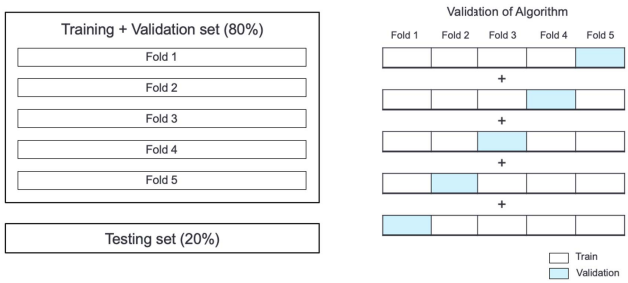

Fig. 5. Schematic of the data partition and the training of the algorithm including k-fold validation step.

D. Evaluation

To evaluate the predictive performance of the algorithms, a set of metrics is calculated by applying the trained model to a testing set and generating predictions. These predictions are based on the probabilities, i.e., a sample is classified as suspicious of blood doping case if the probability is greater than 0.5. Measures used to evaluate the performance of each model include accuracy, sensitivity, specificity, area under ROC curve (AUC).

Since we have a highly imbalanced dataset, it is important to assess both the classes (controlled and rhEPO) separately. Therefore, we evaluated sensitivity which tells us the proportion of correctly identified rhEPO samples, and specificity, which measures the proportion of correctly identified controlled samples.

$$Sensitivity = \frac{TP}{TP + FN} \quad Specificity = \frac{TN}{TN + FP}$$

where TP and TN denote the number of samples classified correctly by the algorithm as rhEPO and controlled, respectively, while FN and FP denote the number of misclassified rhEPO and controlled samples, respectively.

E. Results

We performed the evaluation of all the considered machine learning algorithms using the training and testing set. These algorithms use different approaches to predict the probability of a blood sample to be either a controlled sample or contained rhEPO. Since we performed the cross-validation method for training all the models, Fig. 6 and Fig. 7 show the comparison of training accuracy of all the models using the box plot for the sea-level and altitude arm, respectively.

Table V and Table VI show the performance comparison of all the models by calculating the evaluation metrics for sea-level and altitude arms, respectively. We used the results of the SOTA method [Sottas et al. 2006] as the baseline to compare the performance of our models. Overall, the random forest algorithm performs better performance compared to all the other algorithms for both sea-level and altitude arms. It

Authorized licensed use limited to: Saarl Universitaets. Downloaded on October 13,2022 at 09:34:35 UTC from IEEE Xplore. Restrictions apply.

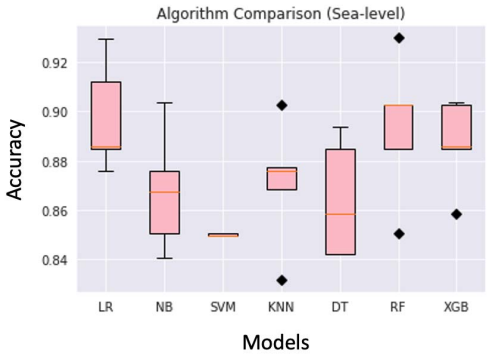

Fig. 6. Performance comparison of training accuracy of all the models for sea-level arm.

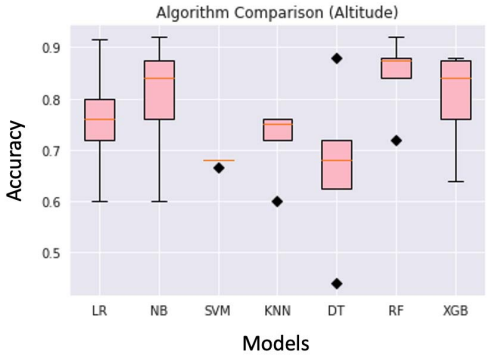

Fig. 7. Performance comparison of training accuracy of all the models for altitude arm.

achieves an accuracy of 94% and 97% on the testing set of sea-level and altitude arms, respectively and 100% specificity for both arms. Similarly, XGBoost also performed well with an accuracy of 92% and 84% on the testing set of sea-level and altitude arms, respectively and 100% specificity for sea-level arm and slightly less of 96% specificity for the altitude arm. Fig. 8 and Fig. 9 show the ROC curve of all the trained models for sea-level and altitude arm, respectively.

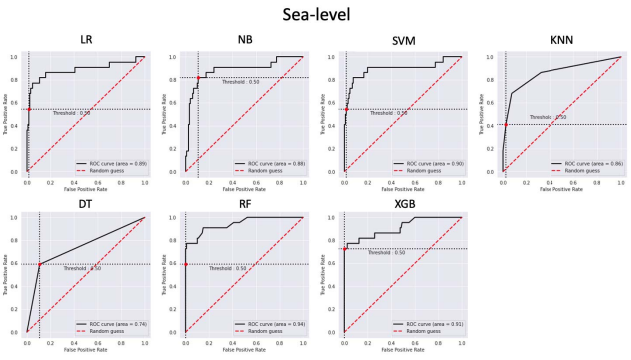

Fig. 8. ROC curves for all the trained models for sea-level arm.

It is important that the model achieve the high specificity

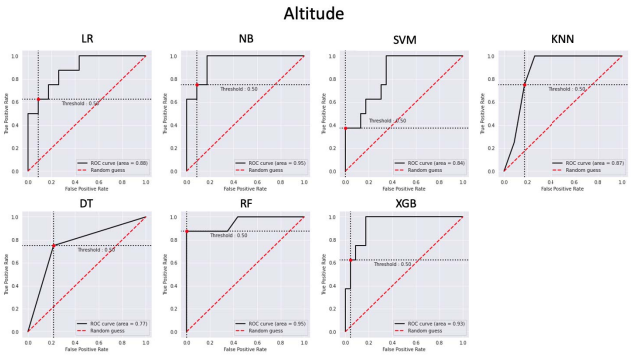

Fig. 9. ROC curves for all the trained models for altitude arm.

value of $\geq$ 99% in anti-doping analysis because any value of false-positive means the model misclassified the controlled samples as a suspicious sample. In reality, if the model triggers such kind of situation, it will end up with additional laboratory testing of the blood sample, which affects the cost and time resources of the authorities. Therefore, we look for the sensitivity of the model at roughly around 99% specificity. We observed that ensemble methods like random forest and XGBoost outperforms the SOTA method, whereas model like SVM could not achieve any value of sensitivity at high specificity.

In general, we expected both RF and XGBoost to show better performance than other algorithms, which is evident from our results. This is because both ensemble algorithms based on bagging and boosting consist of more than one learning algorithm for decision making. Our results show that the random forest algorithm could be used to improve the indirect detection of rhEPO in blood samples.

VI. Discussion

The objective of this analysis is to address a research question on how data-driven approach can help anti-doping analysis to improve the detection of blood doping in sports. In the recent past, several studies have discussed the possible application of machine learning, especially supervised learning algorithms in anti-doping analysis. These studies are usually conducted with the help of the data gathered in clinical experiments on individual populations. In our analysis, we conducted a study, which includes a step-by-step process from collecting data to finding the in-sights of the data.

In this paper, we presented an indirect method to detect the presence of rhEPO in blood samples. We conducted a clinical experiment where 864 blood samples (given placebo or rhEPO) were collected in two arms namely at sea-level and a moderate altitude of 2300m. We combined both statistical methods and machine learning algorithms to analyse the blood samples. At the 99.99% confidence level threshold, we found the potential biomarkers of rhEPO and used them for the machine learning analysis. We trained different machine learning algorithms on the blood samples and evaluated their

Authorized licensed use limited to: Saarl Universitaets. Downloaded on October 13,2022 at 09:34:35 UTC from IEEE Xplore. Restrictions apply.

TABLE V
PERFORMANCE COMPARISON OF ALGORITHMS WITH THE RESULTS FROM SOTA METHOD AT SEA-LEVEL. THE MEAN AND STANDARD DEVIATION VALUES FROM THE CROSS-VALIDATION ARE REPORTED FOR THE TRAINING SET.

Metric	**SOTA**	**LR**		**NB**		**SVM**		**KNN**		**DT**		**RF**		**XGB**	
	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
ACC	-	0.90±0.0	0.90	0.88±0.1	0.87	0.86±0.0	0.85	0.87±0.1	0.89	0.85±0.1	0.85	0.90±0.0	0.94	0.90±0.1	0.92
SN	0.45	0.47±0.2	0.45	0.81±0.2	0.77	0.21±0.3	0.00	0.55±0.1	0.41	0.65±0.1	0.59	0.75±0.2	0.59	0.62±0.2	0.50
SP	1.00	0.95±0.1	0.98	0.86±0.1	0.89	0.79±0.2	1.00	0.95±0.0	0.97	0.78±0.2	0.89	0.92±0.1	1.00	0.95±0.0	1.00
AUC	0.84	0.90±0.0	0.89	0.89±0.0	0.88	0.85±0.1	0.90	0.91±0.1	0.86	0.81±0.1	0.74	0.89±0.1	0.94	0.85±0.1	0.91

TABLE VI
PERFORMANCE COMPARISON OF ALGORITHMS WITH THE RESULTS FROM SOTA METHOD AT ALTITUDE. THE MEAN AND STANDARD DEVIATION VALUES FROM THE CROSS-VALIDATION ARE REPORTED FOR THE TRAINING SET.

Metric	**SOTA**	**LR**		**NB**		**SVM**		**KNN**		**DT**		**RF**		**XGB**	
	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
ACC	-	0.76±0.1	0.81	0.80±0.1	0.84	0.68±0.0	0.74	0.72±0.1	0.81	0.67±0.2	0.77	0.85±0.1	0.97	0.80±0.1	0.84
SN	0.45	0.37±0.3	0.50	0.83±0.2	0.75	0.11±0.2	0.00	0.67±0.1	0.75	0.63±0.2	0.75	0.82±0.2	0.88	0.71±0.2	0.50
SP	1.00	0.85±0.1	0.91	0.88±0.1	0.87	0.82±0.2	1.00	0.90±0.0	0.83	0.79±0.2	0.78	0.95±0.0	1.00	0.92±0.0	0.96
AUC	0.84	0.92±0.1	0.88	0.93±0.0	0.95	0.89±0.1	0.84	0.91±0.0	0.87	0.85±0.1	0.77	0.87±0.1	0.95	0.89±0.1	0.93

performance. Random forest and XGBoost algorithms showed better results and outperformed the SOTA method. Our results suggest that the ensemble methods are effective in mapping the effect of rhEPO in haematological parameters. However, the result is limited to the amount of data available for performing this study. Improving data availability and data quality are potential keys to further enhance the performance of the algorithms. It also opens up the use of more sophisticated non-linear algorithms like a neural network which often learns better with more data. In addition, there are data augmentation techniques like generative models that could possibly help to increase the data statistics. Another factor that could improve the result is by adding some domain knowledge of the haematological parameters in addition to the statistical results. Currently, the K-S test is used to select the potential biomarkers, which is biased towards the data distribution. The presence of outliers in the data could possibly impact the p-values.

In general, AI-based algorithms have the potential to improve the current indirect methods in sports by using the insights from the data for better decision making. In this paper, we showed how the application of a data-driven approach offers a promising result and can significantly improve the decision-making for the detection of drug-abused athletes in sports. Therefore, our work provides a possible solution to address the problem of blood doping in sports and contribute to developing an indirect method for the detection of prohibited substances.

A. Future Research

Deep learning algorithms like neural networks show potential for additional investigation for finding the possible application in anti-doping analysis. However, these algorithms are data-hungry and require a large number of samples for training. Gathering such an amount of data is very difficult because of the associated time restraints and cost factors. Therefore in future, using generative models to increase the data statistics and then apply deep learning algorithms to improve the results could be a potential approach.

REFERENCES

[Akiba et al. 2019] Akiba, T., Sano, S., Yanase, T., Ohta, T. and Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[Chen et al. 2016] Chen, T. and Guestrin, C. (2016). XGBoost: Scalable Tree Boosting System, arXiv:1603.02754.

[Cox 1958] Cox, D. R. (1958). The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological), 20(2), 215–232.

[Dimitrova et al. 2020] Dimitrova, D.S., Kaishev, V.K. and Tan, S. (2020). Computing the Kolmogorov–Smirnov Distribution when the Underlying cdf is Purely Discrete, Mixed or Continuous. Journal of Statistical Software, 95 (10): 1–42.

[Galily 2018] Galily, Y. (2018). Artificial intelligence and sports journalism: Is it a sweeping change?, Technology in Society, 54, 47-51.

[Gore et al. 2003] Gore, C.J., Parisotto, R., Ashenden, M.J. (2003). Second-generation blood tests to detect erythropoietin abuse by athletes. Haematologica 88(3), 333-344.

[Hearst et al. 1998] Hearst, M.A., Dumais, S.T., Osman, E., Platt, J. and Scholkopf, B. (1998). Support vector machines. IEEE Intelligent Systems and their Applications, 13, 18-28.

[John et al. 2012] John, M. J., Jaison, V., Jain, K., Kakkar, N., and Jacob, J. J. (2012). Erythropoietin use and abuse, Indian journal of endocrinology and metabolism, 16(2), 220–227.

[Jelkmann 2016] Jelkmann, W. (2016). Features of Blood Doping. Deutsche Zeitschrift für Sportmedizin, 67, 255-262.

[Kelly et al. 2019] Kelly, T., Beharry, A. and Fedoruk, M (2019). Applying Machine Learning Techniques to Advance Anti-Doping. European Journal of Sports and Exercise Science, 7:2.

[Malcovati et al. 2003] Malcovati, L., Pascutto, C., Cazzola, M. (2003). Hematologic passport for athletes competing in endurance sports: a feasibility study. Haematologica 88(5), 570-581.

Authorized licensed use limited to: Saarl Universitaets. Downloaded on October 13,2022 at 09:34:35 UTC from IEEE Xplore. Restrictions apply.

[Manfredini et al. 2011] Manfredini, F., Malagoni, A. M., Litmanen, H., Zhukovskaja, L., Jeannier, P., Follo, D., Felisatti, M., Besseberg, A., Geistlinger, M., Bayer, P. and Carrabre, J. (2011). Performance and blood monitoring in sports: The artificial intelligence evoking target testing in antidoping (AR.I.E.T.T.A.) project. J Sports Med Phys Fitness, 51(1):153-9.

[Martin at al. 2021] Martin, L, Martin, J-A, Collot, D, et al. (2021). Improved detection methods significantly increase the detection window for EPO microdoses. Drug Test Analysis, 13, 101–112.

[Mucherino at al. 2009] Mucherino, A., Papajorgji, P.J. and Pardalos, P.M. (2009). k-Nearest Neighbor Classification. In: Data Mining in Agriculture. Springer Optimization and Its Applications, vol 34. Springer, New York, NY.

[Parisotto et al. 2001] Parisotto, R., Wu, M., Ashenden, M.J., Emslie, K.R., Gore, C.J., Howe, C., Kazlauskas, R., Sharpe, K., Trout, G.J., Xie, M., Hahn, A.G. (2001). Detection of recombinant human erythropoietin abuse in athletes utilizing markers of altered erythropoiesis. Haematologica 86(2), 128-137.

[Pedregosa et al. 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12:2825-2830.

[Preiman 2001] Preiman, L. (2001), Random Forests. Machine Learning, 45, 5-32.

[Refaeilzadeh et al. 2009] Refaeilzadeh, P., Tang, L., Liu, H. (2009). Cross-Validation. In: Encyclopedia of Database Systems. Springer, USA.

[Sharpe et al. 2006] Sharpe, K., Ashenden, M.J., Schumacher, Y.O. (2006). A third generation approach to detect erythropoietin abuse in athletes. Haematologica 91(3), 356-363.

[Sottas et al. 2006] Sottas, P.-E., Robinson, N., Giraud, S, Taroni, F., Kamber, M., Mangin, P. and Saugy, M. (2006). Statistical Classification of Abnormal Blood Profiles in Athletes. The International Journal of Biostatistics, 2:1.

[WADA 2021] WADA (2021). World Anti-Doping Code 2021, World Anti-Doping Agency.

[Wu at al. 2008] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., et al. (2008). Top 10 algorithms in data mining. Knowledge and Information Systems, 14(1), 1–37.

[Zhang 2004] H. Zhang (2004). The optimality of Naive Bayes. Proceedings of FLAIRS.

[Zorzoli 2011] Zorzoli, M. (2011). Biological passport parameters, Journal of Human Sports and Exercise, 6:2.

Authorized licensed use limited to: Saarl Universitaets. Downloaded on October 13,2022 at 09:34:35 UTC from IEEE Xplore. Restrictions apply.