Journal of Ubiquitous Computing and Communication Technologies (ISSN: 2582-337X)
www.irojournals.com/jucct/

Augmentation for Blood Doping Discovery in Sports using Random Forest Ensembles with LightGBM

D. Sasikala[1], K. Venkatesh Sharma[2]

[1]Professor, Department of CSE, JB Institute of Engineering & Technology, Moinabad, Hyderabad, Telangana, India
[2]Professor, Department of CSE, CVR College of Engineering, Vastu Nagar, Mangalpally, Hyderabad, Telangana, India

E-mail: [1]godnnature@gmail.com, [2]venkateshsharma.cse@gmail.com

Abstract

Athletics bureaucrats round the globe are tackling implausible encounters owing to the partial methods of customs executed by the athletes to progress their enactment in their sports. It embraces the intake of hormonal centred remedies or transfusion of blood to upsurge their power and the effect of their coaching. On the other hand, the up-to-date direct test of discovery of these circumstances embraces the laboratory-centred technique viz restricted for the reason that of the cost factors, handiness of medical experts, etc. This ends us to pursue for indirect assessments. By the emergent curiosity of Artificial Intelligence (AI) in healthcare, it is vital to put forward a process built on blood factors to advance decision making. In this research script, a statistical and machine learning (ML) centred tactic was suggested to ascertain the concern of doping constituent rhEPO in blood units.

Keywords: Blood doping, artificial intelligence, drug abuse, rhEPO, wada, sports

1. Introduction

Artificial Intelligence (AI) has revealed latent upgrading in the sports business, if it is to categorize participants' distinctive aptitudes, spot earlier injuries, or just support decision-making. Computerized Sports Press is a virtuous instance where AI is castoff in sports journalism directorial over computerization. Computerized Insights, an AI-motivated stage that deciphers data from sports confederations into stories by natural language, upsurges the media's reportage proficiency [Galily 2018]. Not merely the custom is illimitable to the progress of sports, then it will likewise be castoff to guarantee justice by athletes in sports. Athletes ensure an aspiration to upsurge their physical enactment to attain superior outcomes

DOI: https://doi.org/10.36548/jucct.2022.2.006
Received: 28.05.2022, received in revised form: 30.06.2022, accepted: 16.07.2022, published: 26.07.2022

that indicates a number of of them to search for the alternative ways. And so, blood doping carry outs in sports have existed about for quite a lot of eras.

The foremost purpose for the countless practices of blood doping is that they are comfortable to implement, and the discovery is challenging. Blood doping is executed by a number of ways and means such as, over and done with blood transfusion besides using the ingestion of erythropoietic stimulating substance or hormones including recombinant human erythropoietin (rhEPO), the transfusion of haemoglobin-established alternates to escalate the quantity of red blood cells (RBC) or augment oxygen transmission to the muscles over an amendment of athlete's oxygen carrying capability [Jelkmann 2016]. The 'World Anti-Doping Program' have been instigated by World Anti-Doping Agency (WADA), and the law enforcement authorities, to define the forbidden constituents wholly that is castoff as doping remedies amid athletes and their outcome on the carry out of the athletes [WADA 2021].

One and only constituent is rhEPO, a recombinant-centred cure that upsurges erythropoiesis that has the capability to amplify oxygenation in the blood. Owing to the circumstance that rhEPO is not simply differentiable commencing subsequently from the logically ensuing erythropoietin, it is universally abused by athletes in their coaching [John et al. 2012]. WADA is in charge for ascertaining the athletes who procure rhEPO cure. The vital dissimilarity is prepared amid direct and indirect systems of the exposure of rhEPO in blood. The straightforward approaches encounter the medico-legal necessities of WADA in which the existence of rhEPO is reviewed in the blood or urine samples via laboratory-centred procedures. Specific systems are the mass spectrometry, biomarker test, RNA testing, liquid chromatography, immunoassays, etc. Yet, these approaches wholly entail professionals to have the suitable domain awareness to amass and examine the blood units rendering to the protocols. In addition, proficient examination of blood units embraces time and cost factors that too must be taken into account in the global decision-making practice. These confines indicate us to examine the indirect approaches of discovery. Meanwhile rhEPO consumption creates distinctive deviations in haematological factors; it is feasible to validate athletes established on indirect pointers for blood doping.

2. Review of Literature

The interrelated workings are concisely reviewed in terms of indirect approaches for the discovery of rhEPO in blood. Studying the effect of rhEPO in blood by statistical approaches is not new in the doping community.

A.R.I.E.T.T.A.: [Manfredini et al. 2011] projected the statistical-centred software named A.R.I.E.T.T.A. that examined the haematological profiles of athletes to spot anomalous configurations. A menace score is reckoned established on diverse factors in view of the transferal on the present-day value from the locus values centred on the quantity of standard deviations.

GASepo: Another software GASepo proposed by [Bajla et al. 2009] intents at the computable scrutiny of imageries acquired by electrofocusing and double-blotting procedures. It is established on the image segmentation of specific ensembles that are castoff to spot rhEPO in doping controls.

Machine Learning Algorithms: [Kelly et al. 2019] relates diverse ML algorithms including, Support Vector Classification, Naive Bayes, Logistic Regression by dissimilar resampling procedures such as, TOMEK and SMOTE systems to categorize the doping actions of the athletes. They studied the blood units of 791 UFC troops. The fallouts gotten advocated that support vector classification and logistic regression united with oversampling can be an efficient technique for ascertaining doping instances. Furthermore, they achieved a sensitivity of 44% with an accuracy of 73% in their results.

Abnormal Blood Profile Score (ABPS): [Sottas et al. 2006] hosted the Abnormal Blood Profile Score (ABPS that is an indirect and global test built on the arithmetical categorization of incidental biomarkers of reformed erythropoiesis. The reckoning of ABPS is built on Support Vector Machine (SVM) and Naive Bayes algorithms mutually using cross-validation practices to map labelled locus profiles to goal efficiencies. As a result, they achieved a sensitivity of 45% at 100% specificity, which is used to comply with the medical and legal standards required by WADA. So, this is the recent state-of-the-art (SOTA) technique that is castoff as a reference line for this research script. As an extension, [Sch¨utz et al. 2018] in their work developed R-based package ABPS to calculate the Abnormal Blood Profile Score. The existing works on indirect methods is limited to the statistical analysis and classical machine learning algorithms with one classifier. In this work, a unified ensemble and boosted algorithm is implemented, where more than one classifier is trained after pooling, and the final prediction is made by the collective decision of the classifiers.

3. Applications

Applications of the AI-based Process for Blood Doping Discovery in Sports include:

1. 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.
2. Offers a promising result and can contribute in a significant manner in improving the decision making for the detection of drug-abused athletes in sports.

4. Prevailing Methodologies

The aim of this study was to detail the three machine learning algorithms that were trained to identify doping cases and also describe the evaluation of our algorithms using Support Vector Machine, Random Forest and XGBoost Process.

4.1 Experiment

4.1.1 Data Collection

The haematological profile of these blood samples that was castoff to study the effect of rhEPO and labelled accordingly was reflected from the prevailing studies.

4.2 Statistical Analysis

4.2.1 Data preprocessing and Finding best indicators

The haematological profile of each collected blood sample is measured that consists of 17 haematological parameters was also pondered from the prevalent studies of blood doping discovery. On the contrary, the best 8 factors are selected based on the previous exploration.

4.3 Machine Learning Studies

Three ML SVC, RF (tree-based ensemble) and XGBoost (tree-based boosting) processes that were trained and evaluated using the enactment metrics.

4.3.1 Evaluation Metrics

For binary categorization, the appraisal of the process in coaching is well-defined centered on the confusion matrix. Additional estimation metrics such as, Accuracy, F1-score, Sensitivity and Specificity are operated [Hossin et al. 2015].

4.4 Results

As per previous investigations the best haematological factors for both sealevel and high altitude are presented with their unique power independent feature of the model. It is observed that HCT is a good factor to analyse the effect of rhEPO only at high altitudes, whereas OFF-HR shows the distinctive power only at sea-level (low altitude). In the next step, SVC, RF and XGBoost algorithms are trained, and the performance metrics are evaluated, as shown in Table I.

These algorithms use different approaches to predict the probability of a blood sample to be either a controlled sample or contained rhEPO. XGBoost achieves an accuracy of 89% that is higher than SVC (68%) and RF(76%). Results are compared in terms of sensitivity, XGBoost outperforms the by 65%, whereas underperforms by 5% in specificity. Over-all, RF and XGBoost together disclosed enhanced enactment than SVC. This is for the reason that mutually boosting and ensemble processes entail additional classifier for decision making.

Finally, the performance of the algorithms is shown with the discussion of the results with possible future research. Erythropoietin, [Jelkmann et al. 2011], Haematological profile, and [Jelkmann 2016] were considered for experimentation based on these base probes accomplished.

Disadvantages

- In prior work, without developing a test for autologous transfusions. Health risks: Like the other forms of blood doping, transfusions can have serious medical consequences. A different individual's blood can comprise a virus that is accidentally distributed on in the transfusion. An athlete who customs his or her own blood be capable of putting themselves at major fitness hazards if the system is not completed accurately or if the blood is not dealt with or put in storage in a appropriate manner. Besides, abnormally high red blood cell points upsurge the menace of stroke, heart attack, and cerebral embolism or pulmonary.

- Amid the complications confronting testers are the prices and troubles of spotting drugs.

- Assessments to spot enactment-augmenting medications in athletes are doing well, and then still the latest ensure restrictions, examining professionals approximately, ensuing the new-fangled reports of pervasive doping in athletics.

- Maximum doping examinations are fabricated to categorize well-known constituents. Then again athletes are continually on the sentry for constituents that are not on testers' position finder. Progressively, these are inherently-fixed up commodities that simulates the body's individual hormones and proteins.

- One such drug popular among athletes is Epogen. Made by Amgen Inc AMGN.O it is a recombinant version of erythropoietin (EPO), a substance produced in the kidney and used by the body to form oxygen-carrying red blood cells. It is castoff to cure anaemia related to chronic kidney malady.

5. Recommended Methodology

RF ensemble with LightGBM algorithm is trained, and their metrics are evaluated, as shown in Table 1. This achieves an accuracy of 95% that is higher than others. Results are compared in terms of sensitivity, outperforms these by 72%, whereas underperforms by 4% in specificity. In general, RF ensemble with LightGBM showed better performance than existing works. This is because of the given reasons: Light GBM is a fast, distributed, high-performance gradient boosting frame. It splits the tree leaf wise with the best fit reducing more loss.

Advantages

- Endure in equipping anti-doping experts using the facts and logical proficiencies required to spot enactment augmenting medications. Maximizing these proficiencies will assist as a restrictive to minimalize doping, uphold health in sport, and sustain a veneer of justice.

- Benefit to mark sporting races virtuous and nondiscriminatory in the future. Using self-learning computer systems make it faster and simpler to uncover doping violations. By feeding these systems with data from doping tests, the systems become increasingly efficient at detecting sporting fraud.

6. Implementation

6.1 Sample Coding: Light GBM

```
reg = lgb.LGBMClassifier(boosting_type="rf",
                         num_leaves=165,
```

```
colsample_bytree=.5,
n_estimators=400,
min_child_weight=5,
min_child_samples=10,
subsample=.632,
subsample_freq=1,
min_split_gain=0,
reg_alpha=0,
reg_lambda=5, # L2 regularization
n_jobs=3)
```

6.2 Comparison of Results

Table 1. Appraisal of Assessment Outcomes of the Machine Learning Processes

ML Algorithms	SVC	RF	XGBoost	RF Ensemble with LightGBM
Accuracy	0.71	0.8	0.88	0.95
F1-Score	0.79	0.89	0.91	0.94
Sensitivity	0.41	0.31	0.62	0.64
Specificity	0.82	0.95	0.95	0.98
AUC	0.72	0.8	0.88	0.93

7. Conclusion

The direct laboratory-based methods are expensive, time consuming process that needs the continuous monitoring of haematological factors of each individual and require domain experts to analyse the blood samples. So, there is a prerequisite for indirect approaches of discovery.

In this paper, an indirect method was presented to detect the presence of rhEPO in blood. A medical investigation was steered where the blood units of 39 individuals (provided placebo or rhEPO) were gathered. Both statistical methods and machine learning algorithms were combined to analyse the blood samples. The K-S examination exposed that the outcome of rhEPO has further impact on RET#. A threshold of 99.99% confidence level was set where the 8 best indicators are selected for training the machine learning algorithms discarding the factors that showed high correlations.

SVC, RF and XGBoost algorithms were trained on the blood samples and evaluated the algorithms by calculating the performance metrics. RF Ensemble with LightGBM algorithm showed better results and outperformed the methods in all the metrics (except specificity). Our results suggest that the ensemble and boosting algorithms are effective for mapping the effect of rhEPO in haematological parameters. However, the result is limited to the amount of data available for performing this study.

8. Future Scope

Refining data handiness and data excellence are latent bases to promote augmenting the enactment of the process. It also opens up the use of more sophisticated non-linear models like a neural network which often learns better with more statistics. There are data escalation systems such as, procreative prototypes that can perhaps benefit to upsurge the data statistics.

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. Yet, it is constrained by the handiness and concealment together of the athletes' data. In this paper, it is revealed how the purpose of AI compromises a hopeful outcome and will subsidize in a substantial way in refining the decision making for the discovery of drug-abused athletes in sports. In future, it is intended to use generative models to increase the data statistics and train deep learning algorithms to improve the results.

References

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

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

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

[4] 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.

[5] 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.

[6] Bajla, I., Holl¨ander, I., Czedik-Heiss, D. and Granec, R. (2009). Classification of image objects in Epo doping control using fuzzy decision tree. Pattern Analysis and Applications, 12(3):285-300.

[7] 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.

[8] 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.

[9] Sch¨utz, F. and Zollinger, A. (2018). ABPS: An R Package for Calculating the Abnormal Blood Profile Score. Frontiers in Physiology, 9.

[10] Jelkmann, W. (2016). Erythropoietin. Front Horm Res, 47:115-27.

[11] Jelkmann, W. and Lundby, C. (2011). Blood doping and its detection, Blood, 118:9.

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

[13] 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.

[14] 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.

[15] Re, M. and Valentini, G. (2012). Ensemble methods: A review, Chapman and Hall, 563-594.

[16] B¨uhlmann, Peter (2012). Bagging, Boosting and Ensemble Methods, Handbook of Computational Statistics.

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

[18] Freund, Y. and Schapire, R.E. (1996), Experiments with a new boosting algorithm. ICML, 148-156.

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

[20] 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.

[21] 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.

[22] Hossin, M. and Sulaiman, M.N. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. International Journal of Data Mining and Knowledge Management Process, 5:01-11.