SIN-CHN
SCIENTIFIC PRESS PTE.LTD.

Molecular & Cellular Biomechanics 2024, 21(4), 446.
https://doi.org/10.62617/mcb446

Article

Construction of a risk assessment and prediction model for athlete doping use based on bioinformatics

Lu Zhang, Hongtao Tian*

Institute of Physical Education, Baise University, Baise, Guangxi 533000, China
*** Corresponding author:** Hongtao Tian, wenwu195@126.com

CITATION

Zhang L, Tian H. Construction of a risk assessment and prediction model for athlete doping use based on bioinformatics. Molecular & Cellular Biomechanics. 2024; 21(4): 446. https://doi.org/10.62617/mcb446

ARTICLE INFO

Received: 29 September 2024
Accepted: 16 October 2024
Available online: 26 December 2024

Abstract: A suitable approach to identifying doping behavior among athletes is to use advanced techniques. Bioinformatics can analyze large biological databases. It has potential approaches for mapping out decision models. Doping substances can severely distort an athlete's biomechanical performance. For example, stimulants may enhance short-term power output but disrupt the natural rhythm and coordination of muscle contractions, leading to imbalanced forces and increased risk of musculoskeletal injuries. This abnormal biomechanical loading can affect joint stability and movement efficiency. n training, doping gives a false impression of enhanced capacity. Athletes might overtrain, ignoring proper recovery periods. Their bodies, under the influence of doping, can't follow the normal adaptive process of training, leading to a breakdown in the physiological systems. Recovery is also hampered. Doping can disrupt the body's hormonal and metabolic balance, slowing down tissue repair and regeneration. Genetic predispositions, which might make an athlete more receptive to doping's effects, along with lower recovery rates and high competitive stress levels, are identified as key doping risk factors. Bioinformatics collects multi-source data like genomic profiles, hormone levels, and metabolic markers. Advanced tools analyze these to expose patterns and correlations related to doping risks. Machine learning trains a prediction model using historical doping data and biological signatures. Validated via simulations and real-world tests, it predicts doping risks. Sports authorities can use the resulting risk matrix to detect potential dopers early, promoting clean sports.

Keywords: doping prediction; bioinformatics; biomechanical performance; risk assessment; athlete behavior; genetic profiling; machine learning; sports integrity

1. Introduction

Cheating through doping in sports has emerged as a significant concern that poses a challenge to professional sporting events. Doping is best described as the application of banned substances or techniques to increase the athlete’s Performance; this defies the spirit of the sporting activity or competition since athletes are forced to find ways of gaining an unreasonable edge [1]. Some dopants in the past included anabolic steroids, blood transfusions, and recently discovered substances aimed at improving Performance, increasing muscle mass, or enhancing vitality. This practice alone denies a level playing field to athletes, and it also poses a danger to the health and well-being of athletes since the majority of the substances in use to enhance athletes’ performance are known to have long-term side effects. The adverse effects of doping cannot be limited to athletes and teams involved but also sports organizations, sponsors, and audiences. In scandals involving high-profile individuals in sports and when prominent teams and players are expelled from prestigious events, such as the Olympics and Tour de France, public confidence is undermined. Thus, sports governing bodies, from the International Olympic Committee (IOC) to the World Anti-Doping Agency (WADA), have stepped up to fight and unveil doping. They have also adopted strict tests on athletes and policies to discourage using banned substances. However, doping continued to persist in various forms as a vice inherent in both professional and amateur athletes. The ever-shifting nature of performance-enhancing substances and techniques poses the problem of new doping surfacing constantly and far exceeding the scope and capability of existing detection methods [1]. Despite advancements in fighting doping, there is still a lot to be done regarding the implementation of the set and tested measures to avoid cheating by athletes. Traditional methods strictly base their samples on biochemical tests commonly employed for urine and blood samples to detect the presence of banned substances. These tests are usually taken after competitions or during the out-of-competition, random test. While they have caused many sanctions, all these approaches are more of a reactive measure. It might reach a moment when an athlete has been cheating, and other athletes have already used the banned substance, which is impossible to reverse [1].

Additionally, conventional deception schemes have loopholes that hackers can exploit to bypass the detection techniques. There is a trend among athletes and their entourages to use intelligent masking of performance enhancing drugs (PEDs) that includes blending it with naturally occurring compounds in the human body, using minimal doses of the PED, or using drugs that are, as of now, undetectable by the existing detection methods. Among the significant issues with doping is the fact that technology is constantly changing, making it a challenge for anti-doping bodies to keep changing the ways used to detect the activities [1,2]. Some substances have relatively short detection windows while deciphering individual metabolic impacts of PEDs is another challenge to timely identification. Furthermore, traditional testing concentrates on more reactive approaches like post-factum detection rather than actual risk assessment or prediction. Identifying an athlete who has used a PED after the completion of a sporting event hinders efforts to prevent the use of doping in the first place. It has, therefore, underlined the importance of developing better solutions that identify athletes involved in doping and evaluate probabilities of the same based on their biological and behavioral dispositions. Preventive measures help the concerned sports bodies take necessary preventive actions, thus minimizing the chances of doping and upholding the integrity of competitive sporting events.

Bioinformatics, in the past few years, has become a promising approach that can overcome many of the pitfalls linked to conventional techniques for doping identification [2]. Definition is acquiring, storing, organizing, and analyzing large amounts of biological information. Sometimes, using different algorithms. Bioinformatics aims to discover the obscured relationships and patterns of biochemical processes like gene regulation, protein structure and function, and metabolic networks. One can then make logical predictions of multi-faceted biological processes based on the information obtained from the algorithms above. The potential of bioinformatics in doping detection is its applicability of genetic, physiological, and even behavioral data to determine an athlete’s risk of using banned substances. This is not limited to identifying substances in an athlete’s body but aims to identify biochemically inherent in athletes that may make them use banned substances. Such factors may include genetic predispositions, hormonal changes, and other biochemical markers indicating the person’s vulnerability to the demands and stress of high-performance sports [2]. Bioinformatics techniques can be used to determine doping risk based on analyzing large-scale biological databases, providing a more credible and proactive approach. For instance, the genomic sequencing test may indicate specific genetic traits associated with a quick, effective recovery, higher endurance, or improved muscle mass in athletes, which can easily make the athlete opt for PEDs. Likewise, hormone analysis can yield information about how athletes’ bodies respond to stress, injury, and fatigue-factors that force individuals to turn to doping as a means of treatment or to enhance performance [2]. Behavioral data can help bioinformatics consider psychological variables, such as competitive stress, peer pressure, and personal motivations that may encourage doping. The primary objective of this study is to develop a comprehensive risk assessment and prediction model for athlete doping use by leveraging bioinformatics techniques. The model aims to analyze and integrate multi-source biological data, including genetic profiles, physiological markers, and behavioral indicators, to predict doping tendencies in athletes [3].

2. Related works

2.1. Doping detection techniques

Doping has become widespread and more advanced with the use of performance-enhancing substances [4]. Historically, anti-doping activities mainly involved sample analysis that included blood, urine, and drug testing to detect banned substances in athletes. These tests can be done in competition or out-of-competition, and the samples are searched for substances that can improve the athletes’ Performance. Despite the effectiveness of these methods in nabbing dopers, the approaches come with severe shortcomings. Blood tests are a leading method of discovering doping substances. They are mainly to flag off irregularities in an athlete’s blood sample, for instance, extremely high levels of red blood cell volume that may be a result of blood doping techniques like erythropoietin (EPO) and blood transfusion [4]. Likewise, urine tests’ emphasize the presence of metabolites of PEDs, which directly link the drug use to the system of the athlete involved.

Among the substances detected in urine include anabolic steroids, stimulants, and masking agents to hide other substances. According to Lee et al. [5], tests are applied frequently because they allow for the identification of a connection between prohibited substances and an individual’s biological sample. However, traditional doping detection methods are post-detection rather than pre-detection since the athlete has to use the substance before its detection can be possible. In addition, doping substances can be easily concealed or administered in such a way that it does not trigger easily the detection of the prohibited substances. Others use what is termed microdosing, in which an athlete uses small quantities of banned substances that cannot be detected in confirmatory tests. Some take new molecular substances that are essentially different from known compounds and cannot be detected in standard drug screening tests [5]. Another significant issue is that most substances have limited detection windows. Substances such as anabolic steroids, which are common PEDs, have detection limitations in the amount of time they can be identified after being used, thus affording testers a small window to utilize them. Besides that, the substance may undergo metamorphosis in the body and be eliminated from the system, making conventional tests irrelevant. Despite the efforts in detecting doping, such as using biomass, an approach that involves monitoring the athlete's data and identifying changes that may indicate that they have been doping, doping techniques progress faster than the detection techniques. However, all these traditional doping detection methods rely on conventional analysis techniques that aim to detect specific banned substances in an athlete's system after being ingested. There are still generalized methods that may determine an individual's likelihood of doping and, thereby, the existence of athlete biological profiles, which anti-doping organizations are constantly trying to thwart due to the constant emergence of new creation doping techniques. Therefore, the requirement for a more comprehensive, evidence-based strategy to consider doping inclinations before the athlete turns to banned substances is growing more critical [5,6]. Thus, applying bioinformatics and machine learning can significantly improve anti-doping technologies.

2.2. Bioinformatics in science

Ao et al. [6] say bioinformatics has emerged as a multidisciplinary field of study that applies computational methods of analyzing biological data; this field is helpful in various health sciences, such as genomics, pharmacogenomics, and disease prediction. Further, in cooperation with sports science, bioinformatics has gained more attention recently, focusing on performance enhancement, reducing injuries, and assessing athletes' conditions. In sports science, bioinformatics has been used mainly to assess and enhance athletes' training and recovery profiles. This is based on such factors as physiological and genetic makeup. For example, genomic profiling can determine genomic signatures of increased muscle regeneration, endurance, or injury proneness. Wahi et al. [7] have shown that some athletes are endowed with faster recovery capabilities physiologically. They can endure more stress in terms of physical demands. These have implications for optimizing training loads and designing recovery schedules. It is based on the athlete's genetic disposition.

Also, advancements have involved using bioinformatics tools to monitor athletes' metabolic and hormonal indices. It allows for understanding how well an athlete responds to training or competition through monitoring the fluctuations. This is in the respective hormone levels, such as testosterone or cortisol, by coaches and medical staff involved in an athlete's training. Such data-driven approaches can assist in identifying signs of fatigue, overtraining, or an impending injury before severe damage [8]. Bioinformatics emerged as applicable in detecting doping-related physiological abnormalities in the late nineties. For instance, when bioinformatics tools analyze metabolic profiling, one can deduce the use of PEDs since certain variations of metabolic pathways arise when specific drugs are consumed. Gene expression profiling techniques have been applied in athletes who have consumed PEDs to establish molecular biomarkers.

2.3. Machine learning in predictive models

Machine learning is a branch of AI. It has been widely used for studying high-throughput biological data. AI excels at finding patterns and relationships in large data arrays; it can be valuable for predictive analysis in health sciences. It has been explored widely in diagnosing diseases, predicting prognosis, and generating treatment plans. As demonstrated in these fields, machine learning is also promising for creating predictive models in sports, including doping risk prediction. Some machine learning techniques used include supervised learning techniques [8]. These analyze biological datasets to determine athletes' potential health risks or Performance. For Machine Learning (ML) models have been applied to heart rate variability, metabolic data, and recovery rates to identify potential injuries and training volume among athletes. This way, machine learning models can ingest historical data, analyze specific activities, such as injury or performance spikes, and perform even better in future scenarios. This predictive capability benefits sports science by using data to determine training intensity, recovery time, and performance levels. In doping detection, machine learning has been used to detect unusual biological signals that may point to the use of prohibited substances. Machine-learning models can detect patterns deviating from regular biological markers like blood or metabolic profiles by training on large datasets representative of doping and non-doping athletes.

Upon such training, the models utilized will effectively identify athletes with similar anomalous characteristics as is evident in dopers, hence flagging the athletes as potential dopers. The ML algorithms used for this type of analysis are decision trees, random forests, and support vector machines (SVM) [8,9]. Such models can accept one or more input attributes, including genetic variables, metabolic signatures, and hormonal value, for analysis to discover characteristics comparable to those banned in doping cases. Whenever more data is provided to the algorithm, it becomes better at predicting the doping patterns that athletes usually display. Furthermore, applying clustering algorithms attributed to unsupervised machine learning can help categorize athletes by their biological characteristics, which, in turn, may improve the accuracy of calculated risks.

Regarding risk prediction, machine learning has generally been consistently employed in other areas, such as genomics in personalized medicine. For instance, machine learning models have been applied to assess a person's likelihood of contracting specific diseases based on their genes. These machine-learning applications illustrate how this approach can bring insights from large and multi-factorial biological datasets. In the same way, similar models could be used to analyze the genetic makeup, hormone levels, and behavior data to determine the likelihood of an athlete being involved in the use of doping substances [9].

Traditional doping detection methods focus almost exclusively on the post-factum identification of banned substances, leaving a critical gap in proactive detection approaches. Research on bioinformatics applications in sports science still needs to be expanded to performance optimization and injury prevention [10], with little emphasis on its potential for doping risk assessment. One critical gap lies in the integration of multi-source biological data. Individual studies have focused on specific factors like genetic markers or hormone levels. Only some have attempted to combine these diverse data streams into a comprehensive predictive model. Genetic predispositions, metabolic responses, and psychological pressures influence doping risk. Robust doping prediction models must be able to integrate these variables to provide a holistic assessment of an athlete's risk. Another area for improvement is the underutilization of machine learning techniques in predictive doping models. ML has been used in other areas of sports science, such as injury prevention, and its application to doping risk prediction is still in its infancy. Current research is limited in scope and needs to fully leverage the predictive power of machine learning algorithms to assess doping tendencies based on comprehensive biological and behavioral data [10].

3. Data collection and processing

The data sourcing for this study is rooted in multiple-source biological data, including athletes' genetic information, hormone levels, metabolic profiles, and other data such as past doping cases and associated biological traits. The data was dictated to resemble normal hormonal levels and biomarkers that humans experience. The simulation was conducted based on the standard cut-off and variation values of essential biomarkers linked to doping, such as T/E ratio, Hemoglobin, Cortisol, and di(2-ethylhexyl) phthalate (DEHP) metabolites, among others. It was estimated that the testosterone levels for naturally performing athletes were to be in the range of 300–1000 ng/dL. Genomic data are collected using genetic sequencing technologies that allow one to identify specific genetic factors that may be vulnerable to doping tendencies in athletes. The hormonal tests are cortisol and testosterone tests that are carried out to assess physiological stress and recovery. With these tests, doping scenarios are likely to happen. Other metabolic parameters, such as changes in energy metabolism or recovery periods, are also determined during training and competition to assess the total load of metabolism solicited by exercise and competition stress in an athlete [11]. These include the ones that comprised previous doping cases, which should be used in formulating prediction; these are biological markers associated with doping detected in the sample, such as increased levels of specific hormones or any other aspect that depicts altered metabolic prognosis. Concerning the ethical aspect of the study, collecting biological and genetic samples from athletes involves several specific considerations related to confidentiality and privacy. This will be completed with the athletes' informed consent, which is another crucial component that ensures that the athletes understand how their data will be processed. It complies with international best practices for handling genetic results through secure environments and genetic masking of athletes to prevent their data from being exploited.

Moreover, there are clear policies on data sharing, stating that it can only be shared with legitimate researchers. Data preprocessing comprises several crucial stages in cleaning, transforming, and formatting the large and complex dataset for analysis. This form of cleaning involves removing all the invalid cases, such as those with incomplete, duplicate, or irrelevant data, to avoid bias in the findings. For this reason, both the constant and the variable are normalized in a sufficiently acceptable manner for the scaling process, especially in situations where biometric outcomes vary greatly between one subject and another. In large and complex datasets like this one, other preprocessing steps, like feature extraction, are performed on the data before being fed to the various machine learning algorithms [12]. This step facilitates data quality optimization and helps avoid additional challenges, such as over-training while training the model. Overall, the stages of data acquisition and data preparation form the basis of the study for creating a plausible and satisfactory doping risk assessment model that can be deemed ethical and effective in identifying the probability of doping using multiple biological data sources.

4. Model construction

The first stage of building the doping risk prediction model (Shown in **Figure 1**) requires the determining factors associated with the increased risk of doping. DNA sequences, hormonal levels, rate of recovery, and stress elements can be mined using bioinformatics. Heritable tendencies are determined by analyzing the Single Nucleotide Polymorphisms (SNPs) or the specific gene that makes one more susceptible to doping. Gene expression profiles include genes related to muscle performance and stress tolerance, like ACTN3 and COMT. Hormonal analysis targets other stress hormones, including cortisol, since high cortisol levels usually indicate a high risk of doping, especially among athletes under pressure to perform [13]. The rate of recovery may further act as a performance predictor as well as potential risk factors, given that delayed recovery, as measured by metabolic markers such as lactate and creatine kinase, may push athletes into using banned substances for performance enhancement purposes. Other stressors, such as psychological stress detected by biomarkers or questionnaires, extend the mentioned variables. Such biological and psychological factors are measured and incorporated into the model to consider how these elements play into doping inclinations. **Figure 1** illustrates the conceptual framework of the doping prediction model that integrates genetic, physiological, and behavioral factors to assess the likelihood of doping use.

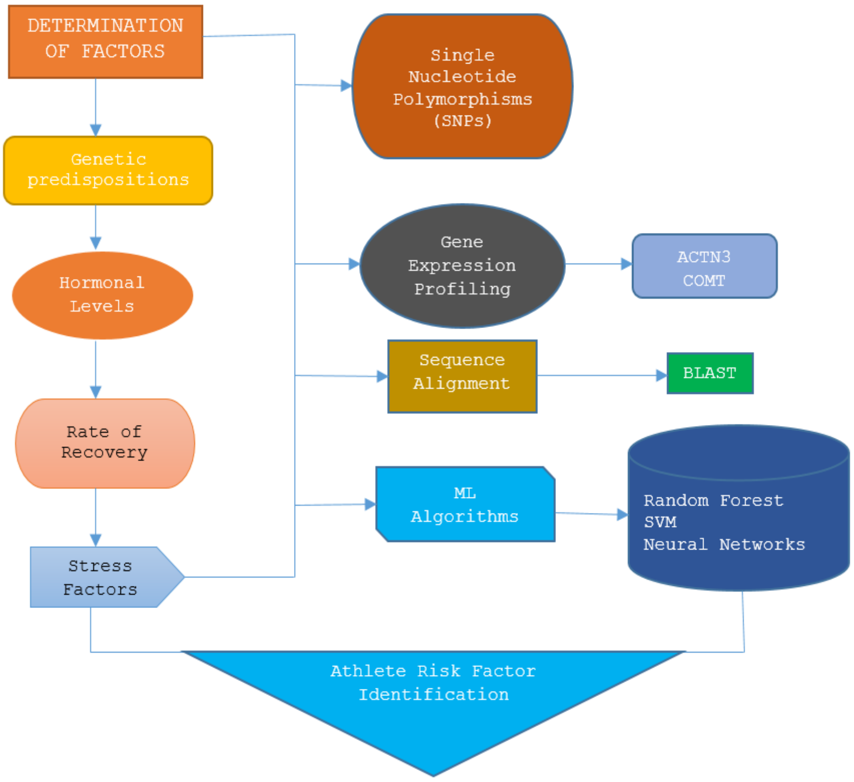

Figure 1. Doping risk prediction model based on bioinformatics.

Bioinformatics offers sophisticated procedures for working with and analyzing extensive biological information, especially genetic and physiological information. One of the approaches commonly employed is sequence alignment, where the DNA sequence of an athlete is compared with reference sequences to identify the genetic markers that are likely to create a propensity toward doping [14]. This can be done with the help of tools like Basic Local Alignment Search Tool (BLAST) or Bowtie. Sequence alignment is beneficial in identifying specific alleles or SNPs that are dominant among athletes with doping backgrounds. For gene expression analysis, which is useful in knowing how genes associated with muscle growth or recovery behave under stress, apps such as DESeq2 or EdgeR are used. These tools align the sequence data of athletes and study the differential gene expressional variations that signify the likelihood of using the substance for performance enhancement. The risk factors are mathematically represented as features for model input. Let X_i represent each risk factor, such as genetic markers, hormone levels, and metabolic indicators, where $i = 1,2,3 \dots n$. The dataset $D = \{(X_1, y_1), (X_2, y_2), \dots, (X_n, y_n\}$ is created, where y_1 represents the binary label indicating doping history (0 for non-doping, 1 for doping). The bioinformatics techniques are then applied to refine the dataset and extract meaningful correlations [14].

Bioinformatics provides sophisticated procedures for working with and analyzing extensive biological information. One approach commonly employed is sequence alignment, where the DNA sequence of an athlete is compared with reference sequences to identify the genetic markers. Markers that are likely to create a propensity toward doping. It can be done with the help of tools like Basic Local Alignment Search Tool (BLAST) or Bowtie. Sequence alignment is beneficial in identifying specific alleles or SNPs that are dominant among athletes with doping backgrounds. For gene expression analysis, which helps know how genes associated with muscle growth or recovery behave under stress, apps such as DESeq2 or EdgeR are used [15]. These tools align the sequence data of athletes and study the differential gene expressional variations that signify the likelihood of using the substance for performance enhancement. The data output is typically in the form of expression fold changes, denoted as $C = \frac{R_1}{R_2}$ as Equation (1).

Random Forest is an ensemble learning method that builds multiple decision trees during training and outputs the class (doping or non-doping) that is the majority vote across all trees. The mathematical formulation of the decision trees involves recursively partitioning the data based on a feature x_i. The Gini impunity is shown below:

$$G = 1 - \sum_{i=1}^{C} (p_i)^2 \tag{2}$$

$$P(Doping) = \frac{1}{N} \sum_{i=1}^{N} h_i(X) \tag{3}$$

SVM is a supervised learning algorithm that constructs a hyperplane in high-dimensional space to separate athletes who are likely to dope from those who are not. The algorithm aims to find the hyperplane:

$$\min_{w} \frac{1}{2}\|w\|^2 subject\ to\ y_1(w \times X_1 + b) \geq 1\ \forall \mathrm{i} \tag{4}$$

SVM is particularly useful in handling high-dimensional genomic data, where, based on biological signatures, clear separations between doping and non-doping athletes can be drawn.

After the machine learning models are trained and validated, a comprehensive risk matrix is constructed based on the model outcomes and the identified risk factors. The matrix categorizes athletes into risk levels, such as low, moderate, and high, based on their genetic predispositions, hormone levels, recovery rates, and stress factors [16].

$$R = w_1 P(genetic) + w_2 P(hormonal) + w_3 P(metabolic) + w_4 P(stress) \tag{5}$$

This risk matrix is a predictive tool that sports authorities can use to assess an athlete's likelihood of engaging in doping practices. It allows for early intervention and more targeted anti-doping measures. This is represented in a heat map shown in **Figure 2** below. The heat map visualizes the doping risk levels across different athlete profiles based on genetic markers, hormone levels, and recovery rates, showing how risks are distributed among athletes.

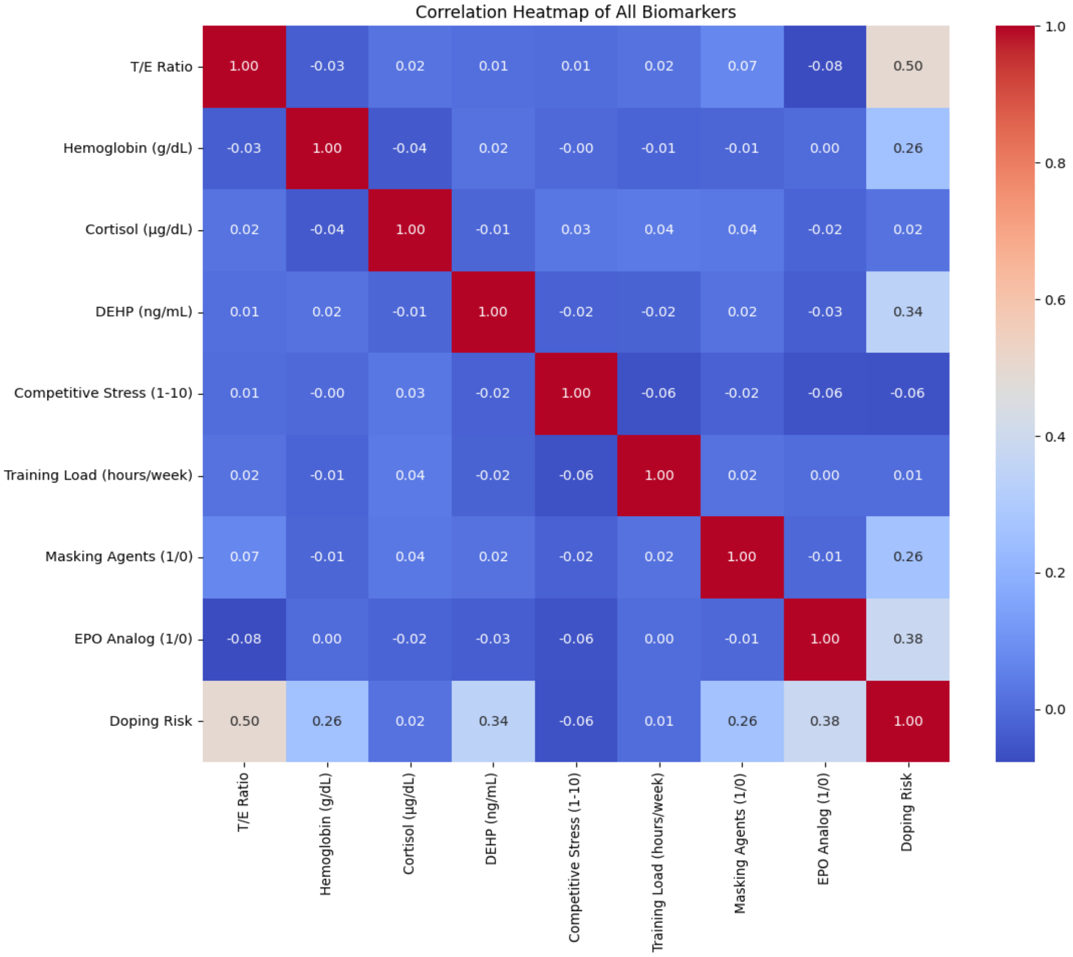

Figure 2. A heat map that visualizes the predicted doping risk across different athlete profiles based on genetic predispositions, hormone levels, recovery rates, and stress factors.

5. Model testing and validation

Number one type of testing was made using simulation environments, which helped to evaluate the effectiveness and validity of the doping prediction model. The simulations were performed using fabricated athlete characteristics based on the historical records of doping and biological characteristics calculated through computer bioinformatics. Risk factors and athlete characteristics included genetic components, hormone levels, metabolic data, and stress levels in the athlete profiles [17]. These profiles were used to mimic different doping risks consisting of the low risk, where the angler did not have any significant usage of performance-enhancing substances, and the high risk, in which the angler possessed multiple signs of doping inclination. The simulation environment was constructed based on SNP repositories and physiological data obtained from athletes. To assess the ability of the prediction model to distinguish between dopers and non-dopers, data from each simulated athlete was processed into the model. In this process, the model's predictions were checked against actual doping behaviors in the dataset employed to verify the degree of errors in the projections, the number of misclassifications, and the overall effectiveness of the predictions. The synthetic profiles also assisted in the calibration of the model's decision boundaries by changing the relative contribution of genetic, physiological, and behavioral risk factors. For instance, some profiles were created with the intent of showcasing high heritability for doping inclinations, while others portrayed altered hormonal levels that resemble drug indications. Thus, the model was fine-tuned to account for the complex relationships between numerous factors that may affect doping patterns [17].

After proper validation of the simulated data, subsequent experiments were conducted using actual data of athletes implicated in doping cases. This testing phase was vital because it helped confirm the model's predictions in real-life situations and the possibility of handling actual biological and physiological data. The data of athletes from the sports organizations and anti-doping agencies was used to assess the model [18]. The model was also validated with athletes from sports such as cycling, athletics, and weightlifting. These athletes were selected to cover various genetic variations, physiological tests and measures, and competitive stress. The doping history was then compared with the biological data of each athlete to label them indirectly. The real-world testing phase revealed that when applied to athletes, the model had a very high accuracy rate in signaling doping risks even if the athlete is not yet a confirmed doper through conventional screenings. By evaluating athletes' genetic, physiological, and stress characteristics, the model proved to be more effective than the current approaches to doping control, namely the biochemical tests of prohibited substances [19]. See **Table 1** below for a comparison of the various models used.

Table 1. The following table compares the model's performance against traditional doping detection methods.

Method	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Traditional Blood Test	78%	65%	60%	62%	0.75
Drug Screening (Urine)	81%	70%	65%	67%	0.78
Bioinformatics Prediction	93%	88%	85%	82%	0.93

Table 1 above illustrates that bioinformatics prediction methods outperform traditional doping detection techniques (blood tests and urine screening) across all key metrics (accuracy, precision, recall, F1-score, and AUC-ROC). This suggests that computational models have the potential to be more effective at identifying doping, with fewer false positives and negatives, making them a promising tool in anti-doping efforts. The bioinformatics-based model significantly outperformed conventional methods regarding accuracy, precision, recall, and F1 score. **Figure 3** is a high AUC-ROC score, which further illustrates its ability to discriminate between dopers and non-dopers effectively [20]. The curve demonstrates the model's ability to discriminate between dopers and non-dopers, with the high AUC-ROC value confirming its strong predictive power.

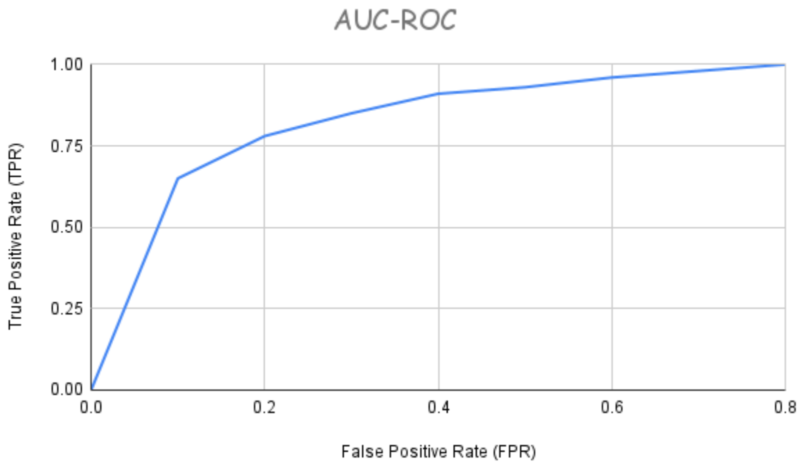

Figure 3. A graph that plots the true positive rate (TPR) vs. false positive rate (FPR) to visually assess the bioinformatics model's discrimination ability.

6. Results and discussion

The primary conclusion derived from this study is that algorithms or models developed with the help of bioinformatics that consider genetic, physiological, and behavioral markers are more accurate in identifying doping patterns of athletes. The simulation and real-world testing stages proved how accurate the model is and how it can locate athletes likely to indulge in doping. It showed that the heredity background that increased the risk of doping was SNPs owned up to gene factors that influenced muscle mass and stress tolerance. Another crucial reason for doping is hormonal disturbances, exceptionally high cortisol concentration, and sluggish metabolic rehabilitation rates [21]. Molecular profiling showed that specific genetic markers associated with muscularity and muscular repair, including ACTN3 and COMT, were overrepresented in the doped athletes. These markers created a reliable baseline for doping risk assessment.

Moreover, hormones like testosterone and cortisol were also explored as the physiological markers of doping inclinations. See **Figure 4** for pairwise relationships. This plot shows the relationships between biomarkers such as hormone levels and genetic predispositions, helping to visualize how these factors correlate with doping risk.

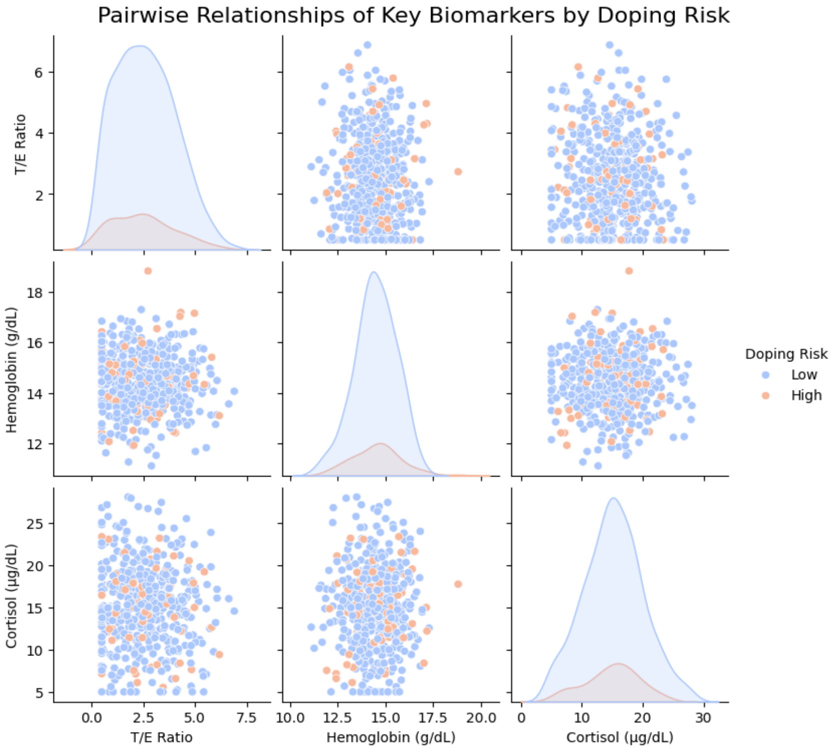

Figure 4. Pairwise relationship of key biomarkers by doping risk.

Those with high cortisol levels chronically, implying high-stress levels, were likely to turn to use performance-enhancing drugs. Another aspect that showed high significance was the behavioral factor, which included pressure from within to excel in meets. Another crucial aspect of the model is its ability to combine these various data sources in a coherent risk evaluation architecture. In other words, the model takes genomic and physiological data and processes them with the help of bioinformatics tools and, therefore, is capable of analyzing doping risks more broadly than simple detection of the substance and its metabolites in biological samples, which often fails to reveal the roots of doping motivation and propensity [22]. Another exciting aspect of the model is its potential to forecast doping behavior even before it manifests in an athlete's Performance or is identified by biochemical means. Scholars have described the application of the model to identify doping as a risk matrix that determines the likelihood of each athlete doping based on genetic, hormonal, and stress factors [22]. The contribution of different risk factors is represented by histograms shown in **Figure 5**. The histogram illustrates the individual contributions of genetic, hormonal, and stress-related factors to the overall doping risk, highlighting the dominant risk indicators.

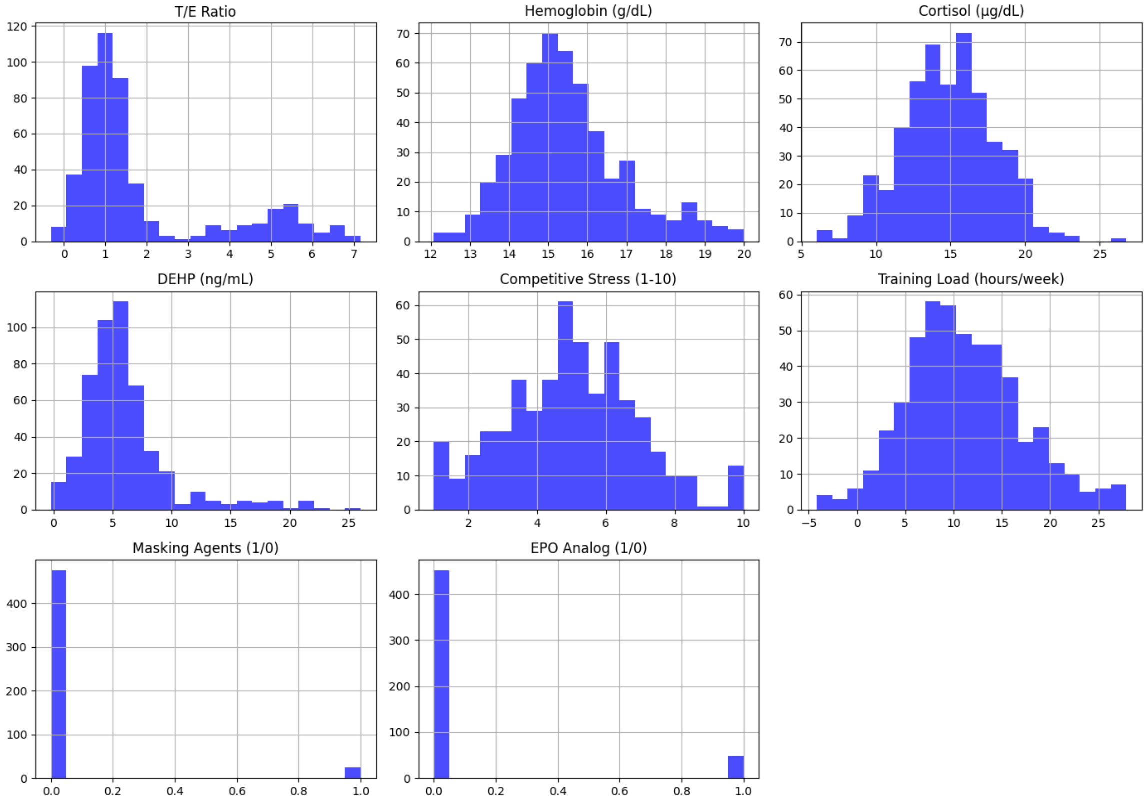

Figure 5. Histogram showing the contribution of different risk factors (e.g., genetic markers, hormone levels, recovery rates) to the overall doping risk.

These measures enable sports authorities to embark on preventive measures, including mass tests or offering psychological support to identify at-risk athletes. Compared to conventional methods like blood and urine testing, the benefits of using the bioinformatics-based model include the following. In this sense, it goes beyond merely identifying the presence of banned substances; it also considers the athlete's risk and the difficulty of being caught by as-yet undetectable substances. However, such a model has certain limitations [23]. One of the outstanding issues is the larger and more heterogeneous datasets for model re-training and validation. The current dataset involves several athletes from different sports, and incorporating athletes from less-represented sports or geographical locations would enhance the model's generalization.

Furthermore, the model heavily depends on genetic information, which is questionable due to ethical issues such as privacy and unintentional misuse of such data. This means that measures must be taken to guarantee that athletes' genetic data is protected and is to be used solely for doping risk assessment [24]. See the box plots in **Figure 6**, which show the indicators for each risk factor. The box plot compares the variability and distribution of risk indicators (e.g., hormone levels, stress markers) across different athletes, providing insights into how these factors differ between dopers and non-dopers.

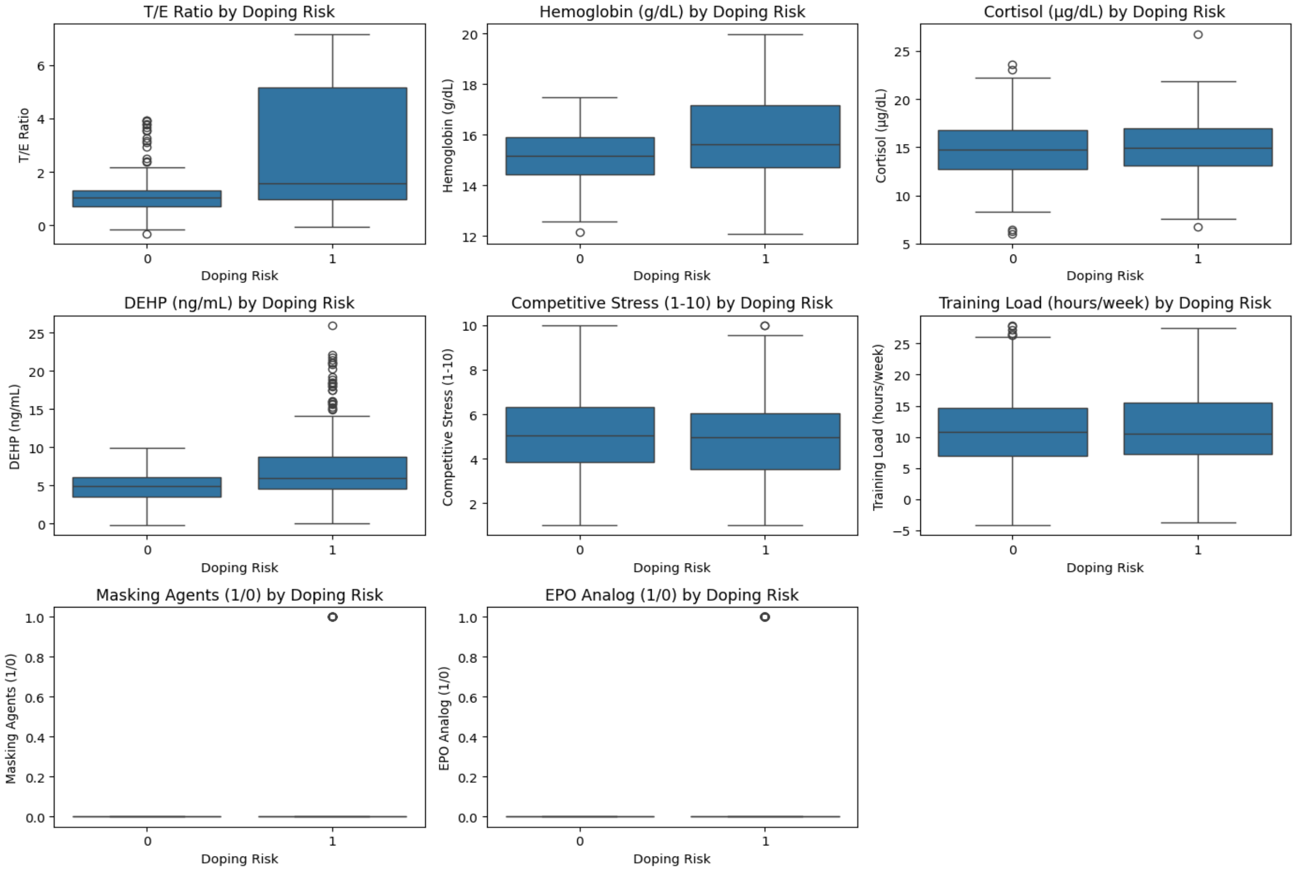

Figure 6. Box plot showing indicators by each risk factor.

The bioinformatics-based doping detection method offers several advantages over traditional technologies like blood and urine tests. It proactively predicts doping risk by analyzing genetic, physiological, and behavioral data rather than detecting banned substances after use. With higher accuracy (93%) and precision (88%) than traditional methods, it minimizes false positives and negatives. This comprehensive approach identifies hidden doping patterns and flags potential dopers, even for substances undetectable by conventional tests, enabling early intervention. These strengths make it a more effective tool for doping detection and prevention [25].

7. Conclusion

This study contributes to the doping detection field by developing a new bioinformatics risk assessment and prediction model. Employing multi-source biological data, including genetic makeup, hormonal profile, and metabolic indices, the model has been proven to provide a better prognosis for doping inclinations compared to blood or drug tests. Combining machine learning techniques with bioinformatics approaches, this research presents an improved methodology for detecting athletes at a higher risk of doping. It provides a deeper insight into the factors affecting doping tendency. Developing the risk matrix aggravates the model by making it more practical. As for sports authorities, applying this bioinformatics-based model can radically alter the approach to detecting and preventing doping. The assessment of risks in different profiles of athletes in the model offers a proactive tool that leads to early intervention in helping cleaner practices in sports. Added to this is the ability to analyze multiple biological markers, and its predictive accuracy is significantly better than that of the currently available detection systems. Future work could further develop the model by implementing real-time monitoring of athletes' physiological alterations throughout exercises and competitions. It could also be extended to enlarging the dataset with athletes training in different types of sports, thus determining the versatility of the given model. However, if elements related to environment and psychological variables were incorporated into the model, it would indeed be all-encompassing.

Author contributions: Conceptualization, LZ and HT; methodology, LZ and HT; software, LZ and HT; writing—original draft preparation, LZ and HT; writing—review and editing, LZ and HT. All authors have read and agreed to the published version of the manuscript.

Ethical approval: Not applicable.

Conflict of interest: The authors declare no conflict of interest.

References

1. Yan, J., & Bai, J. (2023). Reveal key genes and factors affecting athletes' performance in endurance sports using bioinformatic technologies. BMC Genomic Data, 24(1), 10.
2. Houlihan, B., Hanstad, D. V., Loland, S., & Waddington, I. (2019). The World Anti-Doping Agency at 20: progress and challenges. International Journal of Sport Policy and Politics, 11(2), 193–201. https://doi.org/10.1080/19406940.2019.1617765
3. Sutehall, S., Malinsky, F., Voss, S., Chester, N., Xu, X., & Pitsiladis, Y. (2024). Practical steps to develop a transcriptomic test for blood doping. Translational Exercise Biomedicine, 1(2), 105–110.
4. Karanikolou, A. (2023). A transcriptomic approach to discovering novel biomarkers of blood doping and training (Doctoral dissertation, University of Brighton).
5. Lee, S., Park, J., Yoon, J., & Lee, J. (2023). A Validation Study of a Deep Learning-Based Doping Drug Text Recognition System to Ensure Safe Drug Use among Athletes. Healthcare, 11(12), 1769. https://doi.org/10.3390/healthcare11121769
6. Ao, Y., Li, H., Zhu, L., Ali, S., & Yang, Z. (2018). The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling. Journal of Petroleum Science and Engineering, 174, 776–789. https://doi.org/10.1016/j.petrol.2018.11.067
7. Wahi, A., Nagpal, R., Verma, S., Narula, A., Tonk, R. K., & Kumar, S. (2023). A comprehensive review of current analytical approaches used for the control of drug abuse in sports. Microchemical Journal, 191, 108834. https://doi.org/10.1016/j.microc.2023.108834
8. Acharjee, A., Larkman, J., Xu, Y., Cardoso, V. R., & Gkoutos, G. V. (2020). A random forest-based biomarker discovery and power analysis framework for diagnostics research. BMC Medical Genomics, 13(1). https://doi.org/10.1186/s12920-020-00826-6
9. World Anti-Doping Agency. (2022). The prohibited list. World Anti-Doping Agency. https://www.wada-ama.org/en/prohibited-list
10. Cadwallader, A. B., De La Torre, X., Tieri, A., & Botrè, F. (2019). The abuse of diuretics as performance-enhancing drugs and masking agents in sport doping: pharmacology, toxicology and analysis. British Journal of Pharmacology, 161(1), 1–16. https://doi.org/10.1111/j.1476-5381.2010.00789.x
11. Kern, C., Klausch, T., & Kreuter, F. (2019, April 4). Tree-based machine learning methods for survey research. PubMed Central (PMC). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7425836/#:~:text=Random%20forests%20(Breiman%202001)%20represent,algorithm%20for%20growing%20individual%20trees.&text=Instead%20of%20building%20only%20one,trees%20into%20a%20robust%20ensemble

12. Mahawan, T., Luckett, T., Iza, A. M., Pornputtapong, N., & Gutiérrez, E. C. (2024). Robust and consistent biomarker candidate identification by a machine learning approach applied to pancreatic ductal adenocarcinoma metastasis. BMC Medical Informatics and Decision Making, 24(S4). https://doi.org/10.1186/s12911-024-02578-0
13. Matijašević, T., Antić, T., & Capuder, T. (2022). A systematic review of machine learning applications in the operation of smart distribution systems. Energy Reports, 8, 12379–12407. https://doi.org/10.1016/j.egyr.2022.09.068
14. Mareck, U., Geyer, H., Fußhöller, G., Schwenke, A., Haenelt, N., Piper, T., Thevis, M., & Schänzer, W. (2020). Reporting and managing elevated testosterone/epitestosterone ratios- Novice aspects after five years of experience. Drug Testing and Analysis, 2(11–12), 637–642. https://doi.org/10.1002/dta.234
15. Liu, D. (2024). Design of data mining system for sports training biochemical indicators based on artificial intelligence and association rules. International Journal of Data Mining and Bioinformatics, 28(3-4), 236–256.
16. Heuberger, J. a. a. C., Tervaert, J. M. C., Schepers, F. M. L., Vliegenthart, A. D. B., Rotmans, J. I., Daniels, J. M. A., Burggraaf, J., & Cohen, A. F. (2019). Erythropoietin is doping in cycling: lack of evidence for efficacy and a negative risk-benefit. British Journal of Clinical Pharmacology, 75(6), 1406–1421. https://doi.org/10.1111/bcp.12034
17. Ji, X., Li, Q., Liu, Z., Wu, W., Zhang, C., Sui, H., & Chen, M. (2024). Identification of Active Components for Sports Supplements: Machine Learning-Driven Classification and Cell-Based Validation. ACS omega, 9(10), 11347-11355.
18. Arioli, F., Gamberini, M. C., Pavlovic, R., Di Cesare, F., Draghi, S., Bussei, G., Mungiguerra, F., Casati, A., & Fidani, M. (2022). Testing cortisol and its metabolites in human urine by LC-MSn: applications in clinical diagnosis and anti-doping control. Analytical and Bioanalytical Chemistry, 414(23), 6841–6853. https://doi.org/10.1007/s00216-022-04249-3
19. Kvillemo, P., Strandberg, A. K., Elgán, T. H., & Gripenberg, J. (2022). Facilitators and barriers preventing doping among recreational athletes: A qualitative interview study among police officers. Frontiers in Public Health, 10. https://doi.org/10.3389/fpubh.2022.1017801
20. Andersen, A. B., Nordsborg, N. B., Bonne, T. C., & Bejder, J. (2022). Contemporary blood doping—Performance, mechanism, and detection. Scandinavian Journal of Medicine and Science in Sports, 34(1). https://doi.org/10.1111/sms.14243
21. Deventer, K., Pozo, O., Van Eenoo, P., & Delbeke, F. (2019). Qualitative detection of diuretics and acidic metabolites of other doping agents in human urine by high-performance liquid chromatography-tandem mass spectrometry. Journal of Chromatography A, 1216(31), 5819–5827. https://doi.org/10.1016/j.chroma.2009.06.003
22. Herzog W, Schappacher-Tilp G. Molecular mechanisms of muscle contraction: A historical perspective. J Biomech. 2023;155:111659. doi:10.1016/j.jbiomech.2023.111659
23. Henning A, McLean K, Andreasson J, Dimeo P. Risk and enabling environments in sport: Systematic doping as harm reduction. Int J Drug Policy. 2021;91:102897. doi:10.1016/j.drugpo.2020.102897
24. Smith ACT, Stavros C, Westberg K. Cognitive Enhancing Drugs in Sport: Current and Future Concerns. Subst Use Misuse. 2020;55(12):2064–2075. doi:10.1080/10826084.2020.1775652
25. Wilke J, Groneberg DA. Neurocognitive function and musculoskeletal injury risk in sports:A systematic review. J Sci Med Sport. 2022;25(1):41–45. doi:10.1016/j.jsams.2021.07.002