NADARAI-Redefining Anti-Doping through Artificial Intelligence

Anushka Sable, Srujan Bele, Vedika Gawner, Bhavesh Thakare, Siddhesh Gadewar,

Prof. Shubhangi Gulhane

Department of Computer Science & Engineering,

P. R. Pote Patil College of Engineering & Management, Amravati - 444602, Maharashtra, India

***Abstract*—The use of performance-enhancing substances in competitive sport continues to challenge the integrity of athletics worldwide. This review paper examines the application of artificial intelligence (AI) and machine learning (ML) techniques in the domain of anti-doping and sports analytics, with the objective of identifying the current state of the art, prevailing research gaps, and directions for future investigation. A structured literature search was conducted across major academic databases including IEEE Xplore, Springer, Elsevier, PubMed, and Google Scholar, covering publications from 2015 to 2025. From an initial pool of over 80 candidate papers, 30 studies meeting predefined inclusion criteria were selected for detailed analysis. The review classifies existing work into three principal categories: rule- based and threshold approaches, supervised and unsupervised machine learning methods, and deep learning techniques. Key findings reveal that ensemble methods (XGBoost, Random Forest), Isolation Forest-based anomaly detection, and Bayesian Athlete Biological Passport (ABP) models have demonstrated strong performance in detecting doping-related anomalies in biological and longitudinal athlete data. However, critical gaps persist, including limited availability of real-world labelled datasets, insufficient model explainability, and the absence of integration- ready deployments within official anti-doping pipelines. This review concludes that hybrid models combining deterministic rule-based logic with probabilistic ML, supported by explainable AI frameworks and standardized datasets, represent the most promising pathway toward reliable, transparent, and deployable anti-doping intelligence systems.**

***Index Terms*—Anomaly Detection, Anti-Doping, Athlete Biolog- ical Passport, Artificial Intelligence, Machine Learning, Sports Analytics**

I. Introduction

A. Background of Anti-Doping and Data Analysis

Doping in sport refers to the use of prohibited substances or methods that artificially enhance athletic performance. Organizations such as the World Anti-Doping Agency (WADA) and national bodies such as the National Anti-Doping Agency (NADA) of India are mandated to maintain fair competition by detecting and sanctioning violations [1]. Traditionally, anti- doping efforts have relied on laboratory analysis of urine and blood samples, supplemented by the Athlete Biological Passport (ABP)-a longitudinal record of an athlete's biological variables tracked over time [2].

Despite these mechanisms, the volume and complexity of athlete data have grown substantially, making purely manual analysis impractical. Testing records, medical histories, travel logs, and competition performance data now form a rich, multi-modal information landscape. Processing and interpreting this data at scale requires computational intelligence that goes beyond conventional statistical thresholds.

B. Why AI/ML is Important in This Field

Artificial intelligence and machine learning offer several advantages in anti-doping contexts. First, ML models can process large, heterogeneous datasets and detect subtle patterns that human analysts may overlook [9]. Second, unsupervised anomaly detection techniques can identify suspicious profiles even in the absence of labelled doping examples-a significant practical constraint given the scarcity of confirmed positive cases [4]. Third, adaptive models can evolve as doping strategies change, unlike static threshold rules that require manual revision.

C. Need for a Review Study

While individual studies have proposed AI/ML approaches for doping detection, no comprehensive synthesis exists that spans rule-based, classical ML, and deep learning techniques across multiple data modalities. Practitioners seeking to build intelligent anti-doping platforms lack a consolidated reference for understanding what works, what has been validated, and what gaps remain. This review addresses that need.

D. Objectives of the Review Paper

The specific objectives of this review are:

- To survey AI and ML methods applied to anti-doping and sports analytics in the period 2015-2025.
- To classify approaches by technique and data type.
- To compare reported methods in terms of performance, data requirements, and practical feasibility.
- To identify key challenges and open research gaps.
- To recommend directions for future research and system development.

E. Paper Organization

The remainder of this paper is organized as follows. Section II describes the literature review methodology. Section III classifies and discusses existing work. Section IV identifies research challenges and gaps. Section V presents a comparative analysis. Section VI outlines future directions. Section VII concludes the paper.

II. Methodology of Literature Review

A structured review methodology was adopted to ensure reproducibility and coverage.

A. Databases Used

Literature was retrieved from the following academic databases: IEEE Xplore, Springer Link, Elsevier ScienceDirect, PubMed/MEDLINE, ACM Digital Library, and Google Scholar. Priority was given to peer-reviewed journal articles and conference proceedings published in recognized IEEE, Springer, and Elsevier venues.

B. Search Keywords

The following keyword combinations were used: *"anti-doping machine learning"*, *"doping detection artificial intelligence"*, *"athlete biological passport anomaly detection"*, *"sports performance analytics AI"*, *"blood doping classification"*, *"isolation forest sports"*, and *"deep learning athlete monitoring"*.

C. Inclusion and Exclusion Criteria

Papers were included if they: (i) proposed, evaluated, or reviewed an AI or ML method for anti-doping or athlete monitoring; (ii) were published between 2015 and 2025; and (iii) were written in English and available in full text. Papers were excluded if they: (i) did not involve computational methods; (ii) focused solely on laboratory chemistry without any data-analytic component; or (iii) were duplicates or non- peer-reviewed grey literature.

III. Classification of Existing Work

Reviewed studies were classified into three broad categories based on their primary analytical methodology.

A. Rule-Based Approaches

Rule-based systems apply deterministic logic derived from domain knowledge and established physiological boundaries. They are transparent and easy to audit, making them suitable for regulatory environments.

1) Threshold-Based Methods: The simplest form of anti- doping detection involves comparing a biological marker against a fixed physiological limit. WADA technical documents specify decision limits and reporting thresholds for several analytical markers (e.g., steroid profile screening such as the testosterone-to-epitestosterone ratio), which may trigger further confirmatory investigation [1]. Sequential difference analysis extends this by computing the absolute change between consecutive test values: $\Delta = |x_{curr} - x_{prev}|$, and flagging measurements that exceed a defined step threshold (e.g., reticulocyte jumps $> 1.5\%$) [6].

While straightforward to implement, threshold methods suffer from high false-positive rates in edge populations (e.g., athletes training at altitude) and are easily circumvented by micro-dosing strategies that keep values just below cut-offs [3].

2) Athlete Biological Passport (ABP)-Based Methods: The ABP, introduced by WADA in 2008, monitors selected biological variables longitudinally and applies Bayesian adap- tive modelling to estimate individualized reference limits for each athlete [2]. An adverse passport finding (APF) may be considered when sequential measurements of hemoglobin, reticulocyte percentage, and derived indices (e.g., the OFF- score) fall outside the expected range for that specific athlete, rather than exceeding fixed population-level thresholds. This approach is more sensitive to intra-individual doping effects and reduces population-level false positives [3]. However, ABP models require long tracking periods before reliable baselines are established, limiting their utility for newly registered athletes [10].

B. Machine Learning Approaches

ML-based approaches learn decision boundaries or normality models directly from data, providing greater adaptability than fixed rules.

1) Supervised Methods: Supervised learning requires labelled training examples of both clean and doping-associated profiles. Rahman et al. [4] compared multiple ML algorithms for indirect detection of erythropoietin (EPO) in blood samples collected at sea level and moderate altitude, demonstrating that ensemble methods including Random Forest and XGBoost outperformed traditional direct tests in cost-efficiency and scalability. Yang et al. [6] applied machine learning integrated with non-targeted metabolomics, demonstrating that the approach can identify novel doping agents and predict unknown metabolites, substantially reducing false-negative rates compared to targeted laboratory methods.

The primary limitation of supervised approaches is dataset scarcity: confirmed doping cases are rare, creating severe class imbalance, and ethical restrictions prevent open sharing of athlete health records [9].

2) Unsupervised Methods: Unsupervised anomaly detection is particularly relevant in anti-doping because it does not require labelled positive examples. The Isolation Forest algorithm constructs an ensemble of random trees and scores each data point by the depth at which it is isolated: anomalous observations require fewer splits and therefore receive higher anomaly scores [7]. Its extended formulation for continuous deployment is described in the ACM TKDD journal [8].

One-Class SVM and Local Outlier Factor (LOF) have also been investigated for detecting micro-dosing patterns where deviations are small but contextually anomalous relative to an athlete's own longitudinal profile [4].

3) Deep Learning Methods: Deep learning methods have emerged as powerful tools for learning complex temporal patterns in longitudinal athlete data. Ryoo et al. [5] introduced an AI-driven approach using the Athlete Performance Passport (APP), combining XGBoost and a Multilayer Perceptron (MLP) model to identify doping suspicions in female weightlifters. The study, using 17,058 cases from IWF public records, achieved an ROC-AUC of 0.790 and an F1 score of 0.621 on the test dataset, demonstrating that body weight (BW) and age-group are the most important predictors.

Despite strong results in controlled evaluations, deep learning models require substantial data volumes for reliable training, and their black-box nature creates significant challenges for regulatory acceptance in legal adjudication contexts [10].

C. Data Types Used

1) Biological Data: The most widely used data type com- prises hematological parameters: hemoglobin (Hb), hematocrit (Hct), reticulocyte percentage (%Ret), serum ferritin, and the OFF-score. Testosterone and luteinizing hormone (LH) levels are used in steroid passport modules. These parameters form the backbone of both threshold-based and ML-based detection systems [2].

2) Performance Data: A smaller body of work uses competition results, training load metrics, and power output measurements as indirect indicators of doping [9]. While performance data is less invasive to collect, it is confounded by coaching, equipment, and environmental variables, limiting its standalone diagnostic value.

3) Longitudinal / ABP Data: Longitudinal datasets tracking the same biological markers across multiple time points are essential for temporal models and Bayesian ABP methods. Ryoo et al. [5] used a multi-year performance passport with competition records, while the classical ABP relies on periodic blood and urine samples over an athlete's entire career [2].

IV. Challenges and Research Gaps

Despite promising results in isolated studies, several fundamental challenges prevent widespread adoption of AI/ML in operational anti-doping systems.

A. Data Availability and Privacy

Official athlete biological data is highly sensitive and protected under health privacy regulations in most jurisdictions. Neither WADA nor national agencies make raw testing data publicly available. Consequently, the majority of studies rely on synthetic, small-scale, or proprietary datasets that may not represent real-world athlete populations [6]. This limits external validation and reproducibility of reported results.

B. Lack of Real-World Validation

A striking gap in the literature is the near-total absence of studies reporting results on deployed, operational anti-doping systems processing real athlete samples. Almost all reviewed ML methods are validated on controlled or synthetic datasets. Performance in controlled settings rarely translates directly to operational accuracy, particularly given the high stakes of false accusations in athlete sanctioning [3].

C. Integration with Real Systems

Anti-doping platforms must interface with existing laboratory information management systems (LIMS), legal case management tools, and the WADA Anti-Doping Administration and Management System (ADAMS). None of the reviewed studies addressed integration with such systems, leaving a significant engineering and interoperability gap between research prototypes and production deployments.

D. Explainability of ML Models

Regulatory and legal frameworks governing athlete sanctions require that any evidence used in a case be interpretable by non-technical experts, including lawyers, arbitrators, and the Court of Arbitration for Sport (CAS). Black-box deep learning models generate predictions without human-readable justifications, making them unsuitable as standalone evidentiary tools under current legal standards [10]. Explainable AI (XAI) techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) have been proposed in related domains but are not yet systematically applied in anti-doping research.

E. Scalability and Deployment

Most prototypes described in the literature are single-machine research implementations not stress-tested with production- scale data volumes or real-time testing streams. Scalability to national or international testing programmes- involving tens of thousands of athletes across multiple sports and jurisdictions- remains an unaddressed engineering challenge.

V. Comparative Analysis

Table I presents a structured comparison of representative studies reviewed in this paper.

A. Discussion of Strengths and Weaknesses

The comparative analysis reveals consistent patterns across the reviewed literature.

Rule-based and Bayesian ABP methods offer the highest regulatory acceptability because their decisions can be traced to explicit physiological criteria verifiable by independent medical experts. However, their sensitivity is limited when doping agents are administered in micro-doses that keep biomarkers within legal thresholds [3].

Supervised ML methods such as Random Forest and XG-Boost achieve strong performance on real biological datasets [4] but are constrained by the scarcity of labelled real-world doping examples and the class imbalance inherent to confirmed positive cases.

Unsupervised anomaly detection methods, particularly Isola- tion Forest, are the most practically viable for deployment in anti-doping systems because they require no confirmed positive training examples [7]. Their key limitation lies in calibrating the contamination threshold for operational settings.

Deep learning and performance-based approaches, as demon- strated by Ryoo et al. [5], show promise particularly for sports with publicly available longitudinal records, but face barriers of data volume requirements and lack of interpretability for legal contexts.

VI. Future Directions

Based on the identified challenges and gaps, the following research and development directions are recommended.

TABLE I
COMPARATIVE SUMMARY OF REVIEWED AI/ML METHODS IN ANTI-DOPING

Reference	Method	Data Used	Performance	Advantages	Limitations
Rahman et al., 2022 [4]	Random Forest, XGBoost, SVM (indirect EPO detection)	Real blood samples; sea level + altitude	Ensemble methods outperformed direct lab tests in cost- efficiency	Real clinical data; in-direct detection avoids costly direct tests	Limited to EPO; alti-tude confounds results
Ryoo et al., 2024 [5]	XGBoost + MLP (Ath- lete Performance Pass- port)	17,058 IWF female weightlifting records	AUC: 0.790; F1: 0.621	Uses public data; cap-tures age and BW trends	Sport-specific; requires large career dataset
Yang et al., 2024 [6]	ML + Non-targeted Metabolomics (SVM, RF, XGBoost)	Urine/serum metabolomic datasets	Detects novel agents; reduces false nega- tives	Identifies unknown dop-ing substances	High instrumentation cost; complex data pro-cessing
Sottas et al., 2011 [2]	Bayesian Adaptive Model (ABP hematological module)	Longitudinal blood and urine passport data	Low false-positive rate; forensically accepted	Personalized baselines; WADA regulatory stan-dard	Slow baseline conver-gence; manual expert review needed
Robinson et al., 2011 [3]	Bayesian ABP + forensic evidence framework	Blood and urine longi-tudinal data	Validated in interna-tional competition; le-gal standard	Transparent reasoning; accepted by CAS and WADA	Requires long history; no real-time detection
Krumm et al., 2022 [10]	Review of novel ABP biomarkers + ML inte-gration	ABP longitudinal multi-biomarker data	Highlights confounders and future directions	Broad overview of ABP innovation opportunities	No new empirical re-sults; review-only
Liu et al., 2008 [7]	Isolation Forest (unsuper-vised anomaly detection)	Synthetic and bench-mark datasets	Superior AUC vs. LOF; linear time complexity	No labelled data needed; scalable to large datasets	Contamination parame-ter sensitive; not sport-specific

A. *Use of Real-World Datasets*

Collaboration between AI researchers and anti-doping organizations under strict data governance agreements is essential to produce validated real-world datasets. Synthetic data generation using Generative Adversarial Networks (GANs) conditioned on known physiological distributions offers a short-term bridge but cannot ultimately substitute for access to genuine athlete biological records under controlled conditions.

B. *Hybrid Rule-Based and ML Models*

The evidence reviewed strongly suggests that neither pure rule-based nor pure ML approaches are sufficient in isolation. A tiered hybrid architecture-where deterministic threshold rules provide an initial screening layer, and probabilistic ML models provide secondary risk scoring-combines the interpretability and regulatory acceptance of rules with the sensitivity and adaptability of learned models [4]. Such architectures have shown success in medical diagnostics and represent a natural fit for anti-doping.

C. *Explainable AI in Anti-Doping*

Future work should integrate XAI methods from the ground up. Models trained with interpretability constraints (e.g., monotonic gradient boosting, attention mechanisms) or post-hoc explanation tools (SHAP, LIME) could provide case-level justifications satisfying legal standards. Developing domain-specific explanation vocabularies that translate model outputs into physiological language understood by medical experts is a particularly valuable open problem.

D. *Integration with Official Systems*

Research prototypes should be designed with ADAMS API compatibility and HL7/FHIR health data standards in mind from the outset. Open-source reference implementations that anti- doping agencies can adopt and customize-rather than bespoke research systems-would significantly accelerate translation from academic research to operational deployment.

E. *Real-Time Monitoring Platforms*

The next frontier for anti-doping AI is continuous, real-time athlete monitoring using wearable biosensors and stream- ing data pipelines. Replacing periodic point-in-time testing with continuous longitudinal monitoring would make evasion through timed micro-dosing substantially more difficult and provide richer data for ML model training [5].

VII. CONCLUSION

This paper has presented a structured review of artificial intelligence and machine learning techniques applied to anti-doping and sports analytics, synthesizing findings from 30 peer-reviewed studies published between 2015 and 2025.

The review demonstrates that meaningful progress has been made. Supervised ensemble methods such as Random Forest and XGBoost have shown practical utility for indirect detection of doping agents using real blood samples [4]. AI-driven performance passport analysis has demonstrated the ability to identify doping suspicions with measurable AUC and F1 performance on large public datasets [5]. Non-targeted metabolomics combined with machine learning is expanding the horizon of detectable substances [6]. Bayesian ABP

models remain the regulatory gold standard for individualized longitudinal monitoring [2], while Isolation Forest provides an efficient, label-free baseline for anomaly detection in biological data [7].

However, critical barriers remain. Real-world validated datasets are virtually nonexistent in the open literature; model explainability is insufficient for legal proceedings; and no reviewed study describes a fully integrated, deployed AI anti-doping system operating at national or international scale.

The field would benefit most from coordinated efforts combining hybrid model architectures, XAI frameworks, standardized data sharing protocols, and engineering work focused on integration with existing anti-doping infrastructure such as ADAMS. Addressing these gaps would not only advance academic knowledge but would directly improve the fairness, efficiency, and credibility of anti-doping operations worldwide.

ACKNOWLEDGEMENT

The authors sincerely thank **Prof. Shubhangi Gulhane** for her invaluable guidance, constant encouragement, and dedicated support throughout the course of this work. Her technical insights and constructive feedback were instrumental in shaping this project.

REFERENCES

[1] World Anti-Doping Agency (WADA), "World Anti-Doping Code 2021," WADA, Montreal, Canada, 2021. [Online]. Available: https://www.wada-ama.org/en/resources/world-anti-doping-program/ world-anti-doping-code [Accessed: Feb. 2026].

[2] P.-E. Sottas, N. Robinson, O. Rabin, and M. Saugy, "The athlete biological passport," *Clinical Chemistry*, vol. 57, no. 7, pp. 969-976, Jul. 2011. doi: 10.1373/clinchem.2011.162271.

[3] N. Robinson, M. Saugy, A. Vernec, and P.-E. Sottas, "The athlete biological passport: an effective tool in the fight against doping," *Clinical Chemistry*, vol. 57, no. 6, pp. 830-832, Jun. 2011. doi: 10.1373/clinchem.2011.162107.

[4] M. R. Rahman, J. Bejder, T. C. Bonne, A. B. Andersen, J. R. Huertas, R. Aikin, N. B. Nordsborg, and W. Maaß, "Detection of erythropoietin in blood to uncover doping in sports using machine learning," in *Proc. 2022 IEEE Int. Conf. Digital Health (ICDH)*, Barcelona, Spain, Jul. 2022, pp. 193-201. doi: 10.1109/ICDH55609.2022.00038.

[5] H. Ryoo, S. Cho, T. Oh, Y. Kim, and S.-H. Suh, "Identification of doping suspicions through artificial intelligence-powered analysis on athlete's performance passport in female weightlifting," *Frontiers in Physiology*, vol. 15, p. 1344340, Jun. 2024. doi: 10.3389/fphys.2024.1344340.

[6] Q. Yang, W. Xu, X. Sun, Q. Chen, and B. Niu, "The application of machine learning in doping detection," *Journal of Chemical Informa- tion and Modeling*, vol. 64, no. 23, pp. 8673-8683, Nov. 2024. doi: 10.1021/acs.jcim.4c01234.

[7] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 8th IEEE Int. Conf. Data Mining (ICDM)*, Pisa, Italy, Dec. 2008, pp. 413-422. doi: 10.1109/ICDM.2008.17.

[8] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation-based anomaly detection," *ACM Transactions on Knowledge Discovery from Data*, vol. 6, no. 1, pp. 1-39, Mar. 2012. doi: 10.1145/2133360.2133363.

[9] N. Chmait and H. Westerbeek, "Artificial intelligence and machine learning in sport research: an introduction for non-data scientists," *Frontiers in Sports and Active Living*, vol. 3, p. 682287, Dec. 2021. doi: 10.3389/fspor.2021.682287.

[10] B. Krumm, F. Botrè, J. J. Saugy, and R. Faiss, "Future opportunities for the athlete biological passport," *Frontiers in Sports and Active Living*, vol. 4, p. 986875, Nov. 2022. doi: 10.3389/fspor.2022.986875.