scientific reports

Check for updates

OPEN

Detection of sample swapping in anti-doping investigations using machine learning

Maxx Richard Rahman[1,2]✉, Thomas Piper[3], Mario Thevis[3] & Wolfgang Maass[1,2]

The substitution of a urine sample that may result in an adverse analytical finding with a previously collected, clean sample is strictly prohibited under the World Anti-Doping Agency (WADA) regulations and is referred to as sample swapping. When an athlete reuses their own clean sample, detection becomes particularly difficult through conventional analytical methods. In this paper, we propose a similarity detection framework that explicitly accounts for pattern complexity in the analysis of urinary steroid profiles. The framework is based on a convolutional network to capture more complex and subtle variations in profile pairs. Using a dataset of 67,651 steroid profiles collected between 2021 and 2023, the framework was evaluated on both synthetic and laboratory-confirmed similar samples, reflecting realistic variability in doping control processes. The results show that the proposed framework outperforms several baseline models, achieving higher accuracy compared to different baselines. These findings demonstrate the potential of machine learning to improve anti-doping workflows by enabling the automated detection of reused or identical urine samples within large-scale sample collection managed by the Athlete Biological Passport.

High-level athletes are regularly subjected to doping control tests and are required to provide urine or blood samples for analysis[1]. Urine remains the preferred biological matrix for doping control due to its high sensitivity and extended detection window for a broad range of prohibited substances[2]. As testing procedures and analytical methods have improved over time, athletes seeking to evade detection have turned to increasingly sophisticated forms of sample manipulation, as demonstrated exemplarily during the Sochi Olympic Games 2014[3]. One such method involves substituting a potentially positive urine sample with another sample that is free from prohibited substances. This form of sample swapping can be carried out by submitting a clean urine sample from another athlete or by reusing a previously collected sample from the same athlete, provided it was collected prior to the administration of a prohibited substance. The former scenario can be detected due to the unique and stable nature of an athlete's steroid profile over time. Such profiles are monitored through the steroidal module of the Athlete Biological Passport (ABP)[4], which enables longitudinal tracking of an athlete's endogenous steroid concentrations and their associated ratios. Any sudden deviation from the established baseline or individual thresholds is likely to trigger suspicion and prompt follow-up investigation. The ABP, developed by the World Anti-Doping Agency (WADA), leverages statistical models like the Bayesian approach to flag such atypical profiles[5]. In addition, different machine learning approaches have recently been proposed to detect such cases of sample swapping[6,7].

However, when the same athlete reuses their own clean urine sample (collected before doping activity) as a substitute during a doping control test, the deception becomes much more difficult to detect. Since the substituted sample originates from the same athlete, its steroid profile is inherently consistent with the athlete's longitudinal passport. Detection becomes even more elusive when such a sample is analyzed at a different time or in a different WADA-accredited laboratory, as there is no inherent mechanism to flag the identical nature of the specimen. Previous cases of this form of sample reuse have only been discovered by chance, typically because both samples were processed within the same laboratory within a short time frame (i.e., in close temporal proximity).[8,9]. An example of sample reuse is the case of Russian para cross-country and biathlon athlete Nikolay Polukhin during the Sochi 2014 Paralympic Winter Games[10]. Investigations revealed that 'dirty' urine samples were replaced with 'clean' ones provided by the same athlete, to conceal the use of prohibited substances. Forensic analysis of Polukhin's sample bottle revealed evidence of tampering, and DNA analysis confirmed that the urine matched the athlete's profile, indicating that he had reused his own urine sample. This led to the disqualification of his results, including one gold and two silver medals[10].

[1]German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany. [2]Saarland University, Saarbrücken, Germany. [3]German Sport University Cologne (DSHS), Cologne, Germany. ✉email: maxx_richard.rahman@dfki.de

Despite advances in steroid profiling and longitudinal monitoring, identifying reused samples from the same athlete remains a technically unresolved problem. Firstly, the ABP contains thousands of samples collected from different athletes over many years and processed by different WADA-accredited laboratorie[11]. There is currently no automated mechanism to compare a newly collected sample against this vast archive for potential identity or reuse. Performing such comparisons manually is infeasible due to the volume of data and the high-dimensional nature of steroid profiles. Next, while small deviations may exist due to biological or technical variability, reused samples from the same athlete typically exhibit near-identical values across all steroid concentrations and ratios[12]. Furthermore, existing machine learning methods in anti-doping have focused predominantly on detecting atypical values indicative of doping, rather than systematically identifying identical profiles across time[5–7,13–15]. Thus, the detection of sample reuse remains an unresolved problem in anti-doping analytics.

In this study, we address these gaps to detect this more subtle form of sample manipulation. We propose a similarity detection framework based on machine learning algorithm to detect reused or identical urine samples with high specificity. The goal is to identify suspicious sample pairings submitted by the same athlete across time based solely on the steroid profile. The underlying hypothesis is that while minor biological variability may introduce noise into longitudinal data, two identical samples will exhibit near-perfect agreement across all steroid markers and their derived ratios. Using this information, we develop our model to classify pairwise steroid profiles as similar or dissimilar. The results demonstrate the feasibility and effectiveness of data-driven techniques in improving current anti-doping frameworks and advancing the forensic intelligence capabilities of sports integrity programs.

Methodology

Data collection

The dataset contains 67,651 steroid profiles of urine samples analyzed by a WADA-accredited doping control laboratory from 2021–23. In 2021, there were 17,100 male samples and 7,410 female samples, totalling 24,510. The following year, 2022, saw a slight decrease with 15,779 male samples and 7,641 female samples, adding up to 23,420. In 2023, the sample counts decreased further to 13,261 for males and 6,460 for females, resulting in a total of 19,721 samples. This dataset illustrates a year-over-year decline in sample counts for both genders, while the percentage of female samples was relatively stable at 32%, as shown in Table 1.

Each steroid profile comprises of a set of urinary concentrations of key endogenous steroids, including testosterone (T), epitestosterone (E), androsterone (A), etiocholanolone (Etio), 5α-androstane-3α,17β-diol (5αAdiol), and 5β-androstane-3α,17β-diol (5βAdiol). These are complemented by ratios such as T/E, A/T, A/Etio, 5αAdiol/5βAdiol, and 5αAdiol/E, which together form the steroidal passport of an athlete[4,16]. The measurement uncertainty (μ_c) associated with these profiles is a critical factor in evaluating the reliability and interpretability of the steroid profile. It quantifies the degree of doubt associated with a measurement result and captures the cumulative effects of imprecision introduced during sample preparation, mass spectrometry-based quantification, and other analytical processes. According to the WADA's technical guidelines[16], measurement uncertainty during the initial testing procedure is permitted to range between 15% and 30%, depending on the specific steroidal marker. In contrast, during the confirmation procedure, a stricter upper limit of 15% is enforced to ensure analytical robustness. Within this study, and in collaboration with domain experts from a WADA-accredited laboratory, a uniform measurement uncertainty of 15% was adopted across all steroid parameters. While this dataset provides a comprehensive representation of urinary steroid profiles collected under routine anti-doping testing, it is not sufficient on its own for training similarity detection models. In particular, confirmed cases of sample reuse are rare, and the available dataset does not contain a large number of ground-truth positive examples of identical samples. To overcome this limitation, we complemented the real-world data with synthetically generated profile pairs, where controlled perturbations were applied to simulate measurement uncertainty and laboratory variability. This approach ensured a balanced dataset of similar and dissimilar pairs, enabling the model to learn robust similarity patterns that generalize beyond the limited confirmed cases, as described in Table 2.

Confirmation samples The steroidal module of the ABP evaluates the longitudinal steroid profiles obtained on an individual athlete over time[4]. These longitudinal profiles represent data collected during the initial testing procedure routinely applied by all doping control laboratories. If a steroid profile is identified with values close to or above the individually defined thresholds in a passport, the values (i.e. the complete steroid profile of this individual sample) are confirmed upon request by the Athlete Passport Management Unit (APMU)[17] in charge of evaluating the passport. During this confirmation procedure, laboratories employ a slightly different sample preparation approach, resulting in slightly improved measurement uncertainties. Between 2021 and 2023, more than 800 urine samples were confirmed regarding their steroid profiles by one of the WADA-accredited laboratories. In this confirmation stage, the same urine samples are analyzed again, but using a slightly modified preparation and analytical procedure compared to the initial test. This repreparation step introduces small variations due to differences in handling, preparation, or instrument settings, while measurement uncertainty is

Samples	2021	2022	2023	Confirmed
Male athletes	17,100	15,779	13,261	350
Female athletes	7,410	7,641	6,460	193
Total	24,510	23,420	19,721	553

Table 1. Total number of samples collected in different years and confirmation samples used in this study.

Class	Description
Class 1 (Similar Pairs)	These are generated by introducing synthetic noise into the reference (testing) sample. Specifically, the second sample in the pair is created by applying a small pointwise perturbation to each steroid parameter of the testing sample $\mathrm{x}_T \in \mathbb{R}^p$: $\mathrm{x}_T = \mathrm{x}_T \pm \eta \cdot \mathrm{x}_T, \quad \eta \sim \mathcal{U}(0, 0.15)$ This procedure simulates the type of variability that naturally occurs due to laboratory conditions, measurement uncertainty, or slight differences during sample preparation (e.g., re-analysis in confirmation testing). The resulting profile pair $\mathrm{X}_i = \{\mathrm{x}_T, \mathrm{x}_T\}$ considered highly similar and assigned a label $y_i = 1$.
Class 0 (Dissimilar Pairs)	The second sample in the pair is randomly selected from the dataset, ensuring that it has no biological relationship to the reference sample. The resulting pair is labeled $y_i = 0$.

Table 2. Description of how training pairs are generated for the binary classification task. Class 1 (similar pairs) are created by adding controlled synthetic noise to the reference sample to mimic measurement uncertainty and laboratory variability. Class 0 (dissimilar pairs) are formed by pairing the reference sample with a randomly selected profile from another athlete.

simultaneously reduced to a maximum of 15% as required by WADA guidelines. Therefore, such confirmation samples provide a realistic proxy for repeated testing conditions: they reflect both biological consistency (since the urine originates from the same athlete) and procedural variability. For this reason, we used this subset to evaluate our methods for detecting identical samples, as they closely resemble the practical scenarios encountered in real doping control workflows.

Principal component analysis

To facilitate interpretable analysis, we apply dimensional reduction to project these high-dimensional profiles into a low-dimensional space. Within this space, we compute similarity distance between the test sample and all other samples to explore potentially identical samples based on a learned similarity threshold.

Dimensional reduction

Let $X \in \mathbb{R}^{n \times p}$ denote the original data matrix, where n is the number of samples and p is the number of features (i.e., steroid parameters). Each row vector $\mathrm{x}_i \in \mathbb{R}^p$ corresponds to a sample. We performed principal component analysis (PCA)[18] for dimensionality reduction, which transforms the original correlated feature space into an orthogonal basis, where the axes (principal components) are ranked by the amount of variance they capture in the data. In this study, we reduce the data to three dimensions by retaining the top three eigenvectors and construct the matrix $V_3 = [\mathrm{v}_1, \mathrm{v}_2, \mathrm{v}_3] \in \mathbb{R}^{p \times 3}$. Each sample is then projected into this 3D principal component space:

$$\mathrm{y}_i = \mathrm{z}_i V_3, \quad \text{for } i = 1, \ldots, n$$

where $\mathrm{z}_i \in \mathbb{R}^p$ is the standardized vector for sample i, and $\mathrm{y}_i \in \mathbb{R}^3$ is its 3D embedding. This projection retains the directions of highest variability in the data, while significantly reducing dimensionality.

Similarity distance computation

With all samples embedded in the same 3D space, we compute the L1 distance between the testing sample and every other sample:

$$D_i(\mathrm{y}_i, \mathrm{y}_\mathrm{T}) = \sum_{k=1}^{3} |y_{ik} - y_{\mathrm{T},k}|$$

This metric is robust to small deviations and less sensitive to outliers compared to L2 (Euclidean) distance, making it particularly suited for detecting sample-level perturbations due to measurement uncertainty or biological noise. After computing $D(\mathrm{y}_i, \mathrm{y}_\mathrm{T})$ for all $i = 1, \ldots, n$, we compare each distance to a predefined threshold D_T. This threshold is determined empirically by evaluating model performance across different cut-off values and selecting the one that best balances sensitivity and specificity, as shown in Sect. 3.2. The samples satisfy:

$$D_i(\mathrm{y}_i, \mathrm{y}_\mathrm{T}) < D_T$$

are considered similar, and their distances serve as a proxy for the degree of similarity. The output is a ranked list of candidate samples, ordered by ascending L1 distance. This procedure provides a non-parametric and interpretable filtering mechanism, where the threshold D_T effectively controls the sensitivity of the detection. A smaller threshold increases precision (fewer false positives) but may miss slightly altered true matches, while a larger threshold improves recall at the risk of including unrelated samples.

Similarity detection framework

To capture non-linear dependencies and complex variations in steroid profiles that may arise due to biological variability, or deliberate manipulation, our proposed framework is based on convolutional neural networks (CNN)[19]. CNNs are particularly well-suited for this task because they can efficiently model local dependencies between steroid parameters (e.g., ratios such as T/E or A/Etio), while remaining computationally lightweight and scalable to large datasets. Unlike transformers, which typically require long sequential context and substantial computational resources, the input structure of steroid profiles is relatively short and fixed in dimensionality,

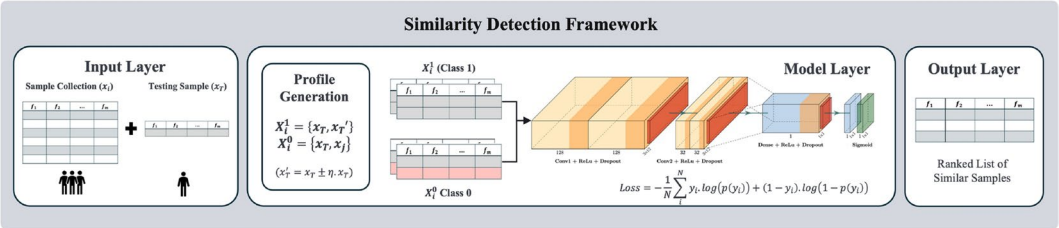

Fig. 1. Overview of the proposed similarity detection framework for detecting identical or highly similar steroid profile. The input data includes the collection of urine samples and a testing sample which is feed into convolutional neural network, where synthetic similar (class 1) and dissimilar (class 0) sample pairs are generated to train a model for similarity classification.

Component	Description
Input	Each input pair is reshaped into a shape of (1, 2, p), i.e., *(channel, height, width)*.
Convolutional Block 1	Applies 128 filters of kernel size (2×1), producing 128 intermediate feature maps. This is followed by Rectified Linear Unit (ReLU) activation and *dropout* (rate = 0.3) to prevent overfitting.
Convolutional Block 2	Applies 32 filters of size (2×1), again followed by ReLU and dropout. These layers capture local patterns across adjacent steroid features and are particularly useful for modeling interaction terms (e.g., how the ratio of two steroids varies between samples).
Dense layer	The output from convolutional layers is flattened and passed through a fully connected layer with 64 neurons and a ReLU activation. This serves as a high-capacity integrator of learned features.
Output Layer	A single neuron with sigmoid activation computes the probability of similarity: $p(y_i) = \sigma(z_i) = \frac{1}{1+e^{-z_i}}, \quad z_i = \mathrm{w}^T h_i + b$ where h_i is the output of the dense layer, w and b are trainable weights, and σ is the sigmoid function.

Table 3. Architecture of the CNN model and component-wise description for similarity prediction. This architecture was selected after hyperparameter optimization, where different filter sizes, number of layers, dropout rates, and learning rates were evaluated. The final configuration provided the best balance between evaluation metrics, making it well-suited for anti-doping datasets.

making transformer-based models unnecessarily complex in this setting. In contrast, classical machine learning approaches rely on predefined similarity measures and often fail to capture higher-order feature interactions in biosample data. CNNs overcome this limitation by directly learning latent representations from raw profile pairs, enabling the detection of subtle but meaningful similarities that are critical for identifying identical or highly similar profiles in anti-doping analysis. Figure 1, the framework takes as input a steroid profile (testing sample) and compares it against a historical collection of different athlete profiles. It classifies profile pairs based on learned similarity patterns and produces ranked outputs of candidate matches, which can be cross-referenced or independently validated in forensic workflows.

Profile generation

We frame the similarity detection task as a binary classification problem, i.e., given a pair of steroid profiles, the model predicts whether both profiles originate from the same athlete. To generate such training pairs, we use a profile generator that constructs input pairs by combining a testing sample (reference) with another sample from the dataset. We consider two types of profile pairs as described in Table 2. Each input pair is structured as a 2D tensor of shape $(2 \times p)$, where p denotes the number of steroid parameters. In this format, the two rows correspond to the two samples being compared. By processing these paired profiles jointly, the CNN can learn complex relationships and subtle differences between the two sets of steroid parameters.

To generate training data, we combined all available male and female profiles, excluding the confirmation samples to avoid bias. This resulted in a dataset of 63,179 profiles (43,118 from male and 20,061 from female athletes). To create a balanced test set, we randomly selected 20% of the female profiles (4,013 samples) and matched this number with an equal number of male profiles. The remaining 55,153 profiles were used for training. From these training profiles, we constructed 110,306 profile pairs: half labeled as similar (55,153 pairs) and half as dissimilar (55,153 pairs). Similar pairs were generated by applying controlled perturbations to simulate measurement uncertainty, while dissimilar pairs were created by randomly combining samples from different athletes. This balanced strategy ensured that the model learned to distinguish genuine biological similarity from variability introduced by laboratory procedures, which is a distinction important in the context of anti-doping analysis.

Model architecture

The model consists of two convolutional blocks followed by a fully connected dense layer and a sigmoid output unit, as shown in Table 3. The use of dropout and small filter sizes ensures both regularization and generalization.

Training and loss optimization

We optimize the model using the binary cross-entropy (BCE) loss:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p(y_i)) + (1-y_i)\log(1-p(y_i))]$$

Here, $y_i \in \{0,1\}$ is the true class label for the ith profile pair, and $p(y_i)$ is the predicted similarity score from the model. This loss penalizes incorrect predictions while pushing the model to output confident probability estimates. We use the Adam optimizer[20] with an initial learning rate of 10^{-4} and apply a mini-batch training strategy with batch size 64. Early stopping is used to terminate model training when validation loss ceases to improve, reducing the risk of overfitting. Furthermore, dropout regularization (rate = 0.3) is applied after both convolutional and dense layers to maintain robustness.

Model inference

Once trained, the CNN is used as a similarity scoring function. Given a test sample x_T, we generate candidate profiles $\{\mathrm{x}_T, \mathrm{x}_j\}$ by pairing it with every sample $\mathrm{x}_j \in \mathcal{D}$ in the dataset. Each profile is passed through the trained model, which outputs a probability score:

$$s_j = f_\theta(\mathrm{x}_T, \mathrm{x}_j)$$

where f_θ denotes the trained CNN with parameters θ, and $s_j \in [0,1]$ represents the predicted similarity between x_T and x_j. To determine whether a candidate sample is "similar" to the test sample, we compare the score s_j against a decision threshold τ. This threshold is chosen based on validation performance:

$$\hat{y}_j = \begin{cases} 1, & \text{if } s_j \geq \tau \\ 0, & \text{otherwise} \end{cases}$$

A higher threshold increases precision (fewer false positives), while a lower threshold improves recall (fewer false negatives). The final output is a ranked list of samples with similarity scores, providing a scalable tool for scanning large datasets for screening duplicates or swapped samples.

Rationale and advantages

The CNN-based approach offers several distinct advantages over classical methods. First, it does not rely on assumptions about linear separability or fixed variance structure. Instead, it learns directly from data, adjusting to both inter-individual variation and common noise sources. Second, the use of convolutional layers enables the model to capture local dependencies and feature interactions, such as co-fluctuations between related steroid ratios (e.g., T/E or A/Etio). This makes the model particularly well-suited to detecting similarity patterns that may not be captured by distance-based metrics. Furthermore, the CNN can be integrated into an automated screening pipeline, where each new test sample is processed in batch against the historical database, resulting real-time similarity estimates.

Ethical considerations

All methods and analyses in this study were carried out in accordance with the World Anti-Doping Code[21] and the International Standard for Laboratories[22] currently in force, as well as all relevant guidelines and declarations governing anti-doping research. The study involved the secondary use of existing samples and data collected under the regulatory framework of these standards, which constitute the relevant licensing and oversight authority for such analyses. Data re-evaluation was done using anonymized datasets, containing previously determined urinary concentrations and concentration ratios of steroidal analytes, re-used in accordance with the regulations of the International Standard for Laboratories[22] (Chapters 5.3.6 and 5.3.12.2) of the World Anti-Doping Agency. Due to the fact that the conditions required for research on routine doping control samples, as outlined in the World Anti-Doping Code[21], were fulfilled, the World Anti-Doping Agency waived the need to obtain further informed consent from sample donors.

Due to the fact that the conditions required for research on routine doping control samples were fulfilled, as outlined in the World Anti-Doping Code[21] and the International Standard for Laboratories[22], the requirement for ethical approval was waived by the World Anti-Doping Agency.

Results and discussion

Statistical analysis

Before selecting or training any machine learning model for similarity detection, it is important to understand the underlying structure, variability, and distributional properties of the steroid biomarker data. These in-depth investigations serve as a prerequisite for informed model choice, feature transformation, and threshold selection, especially given the physiological complexity and inter-individual variation in the athlete population.

We begin with a comprehensive statistical analysis of urinary steroid parameters from athletes tested in 2021–23. The descriptive statistics (see Tables 4, 5, 6) include the mean ± standard deviation, minimum, quartiles (IQ1, IQ3), and maximum values for each steroid marker and ratio, enabling a thorough understanding of both central tendency and dispersion. The found urinary concentrations corroborate earlier finding published over the years[23–25]. A consistent trend across all three years is the substantially higher steroid concentrations in male athletes relative to females. For example, testosterone in males averaged 30.67 ± 20.94 ng/mL in 2021, 30.39 ±

	Male ($n = 17100$)						Female ($n = 7410$)					
Steroid parameter	**mean±std.**	**Min**	**IQ1**	**Median**	**IQ3**	**Max**	**mean±std.**	**Min**	**IQ1**	**Median**	**IQ3**	**Max**
A	2519.73 ± 1266.60	0.82	1594.74	2308.66	3210.18	7625.75	1764.32 ± 1072.63	4.76	972.07	1538.83	2300.40	6280.00
Etio	1986.37 ± 942.98	2.28	1294.49	1816.94	2509.90	5550.16	1973.44 ± 1069.75	4.76	1172.37	1759.54	2550.97	6037.32
E	27.70 ± 17.97	0.59	14.18	23.30	37.15	94.04	7.97 ± 5.87	0.30	3.56	6.34	10.82	33.20
T	30.67 ± 20.94	0.40	15.90	27.48	41.85	120.46	5.87 ± 4.86	0.00	2.67	4.79	7.74	62.17
5α Adiol	54.37 ± 33.44	0.09	31.08	47.10	69.63	303.15	20.43 ± 15.89	0.54	10.06	16.26	26.09	178.33
5β Adiol	143.58 ± 104.78	0.33	65.94	117.02	193.50	610.49	66.93 ± 60.96	1.57	24.76	47.11	87.60	389.27
T/E	1.39 ± 1.02	0.02	0.71	1.18	1.85	5.99	1.01 ± 0.84	0.00	0.42	0.81	1.35	5.82
A/Etio	1.37 ± 0.60	0.01	0.94	1.28	1.71	3.72	0.95 ± 0.44	0.00	0.64	0.88	1.20	2.59
A/T	162.17 ± 210.45	0.17	55.59	84.21	138.00	1148.41	581.46 ± 779.22	0.00	203.69	310.94	514.32	4438.41
5α Adiol/5β Adiol	0.51 ± 0.34	0.01	0.26	0.41	0.66	1.92	0.44 ± 0.31	0.02	0.20	0.36	0.58	1.99
5α Adiol/E	2.55 ± 1.88	0.01	1.29	2.01	3.18	15.62	3.40 ± 2.60	0.05	1.64	2.71	4.34	23.29

Table 4. Descriptive statistics of urinary steroid parameters for male ($n = 17100$) and female ($n = 7410$) athletes in the 2021 dataset. The table reports the mean ± standard deviation (std.), minimum (min), first quartile (IQ1), median, third quartile (IQ3), and maximum (max) for each parameter. The data shows significant gender-based differences in concentration levels and variability across most biomarkers.

	Male ($n = 15779$)						Female ($n = 7641$)					
Steroid parameter	**Mean±std.**	**Min**	**IQ1**	**Median**	**IQ3**	**Max**	**Mean±std.**	**Min**	**IQ1**	**Median**	**IQ3**	**Max**
A	2662.86 ± 1332.24	2.63	1680.93	2424.31	3402.43	8032.49	1952.45 ± 1255.96	3.70	1046.37	1665.98	2530.78	7897.79
Etio	1952.25 ± 965.63	2.63	1235.88	1772.08	2473.36	5671.47	2024.14 ± 1113.92	3.70	1201.88	1792.78	2631.56	6165.68
E	27.59 ± 18.08	0.31	13.98	23.01	36.82	94.00	8.41 ± 6.46	0.27	3.88	6.65	11.17	48.45
T	30.39 ± 21.17	0.31	15.26	27.26	41.79	129.31	6.08 ± 5.34	0.00	2.46	4.94	8.23	79.85
5αAdiol	55.91 ± 32.15	0.52	32.61	49.40	71.66	236.89	21.76 ± 16.51	0.49	10.66	17.35	27.80	153.53
5βAdiol	133.03 ± 98.03	0.66	59.58	106.98	180.62	568.42	63.06 ± 60.69	1.48	22.73	43.50	82.28	471.49
T/E	1.40 ± 1.08	0.01	0.68	1.17	1.86	7.19	0.95 ± 0.78	0.00	0.38	0.78	1.31	5.17
A/Etio	1.49 ± 0.64	0.07	1.02	1.38	1.84	3.88	1.02 ± 0.49	0.11	0.67	0.93	1.28	3.31
A/T	180.86 ± 242.10	2.05	59.23	89.50	154.50	1425.44	819.43 ± 1540.37	0.00	214.12	332.79	580.73	20105.92
5αAdiol/5βAdiol	0.57 ± 0.36	0.04	0.30	0.47	0.74	1.93	0.50 ± 0.33	0.02	0.24	0.42	0.68	1.69
5αAdiol/E	2.64 ± 1.81	0.01	1.36	2.12	3.41	11.17	3.38 ± 2.45	0.13	1.68	2.72	4.33	17.20

Table 5. Descriptive statistics of urinary steroid parameters for male ($n = 15779$) and female ($n = 7641$) athletes in the 2022 dataset. The table reports the mean ± standard deviation (std.), minimum (min), first quartile (IQ1), median, third quartile (IQ3), and maximum (max) for each parameter. The data shows significant gender-based differences in concentration levels and variability across most biomarkers.

21.17 ng/mL in 2022, and 28.77 ± 19.84 ng/mL in 2023. In contrast, female testosterone levels remained around 6 ng/mL across all years: 5.87 ± 4.86 (2021), 6.08 ± 5.34 (2022), and 5.13 ± 4.55 (2023). Similarly, androsterone in males averaged 2519.73 ± 1266.6 (2021), 2662.86 ± 1332.24 (2022), and 2474.99 ± 1221.89 (2023), while in females it ranged between 1764.32 ± 1072.63 (2021), 1952.45 ± 1255.96 (2022) and 1809.45 ± 1073.45 (2023). Etiocholanolone followed a similar pattern, remaining 1950-2000 ng/mL for both gender. The standard deviations and ranges indicate higher biological variability in males. In 2022, male testosterone spanned from a minimum of 0.31 ng/mL to a maximum of 129.31 ng/mL, while in females, it ranged from 0.0 to 79.85 ng/mL, which shows not only gender differences but also the presence of rare high-end outliers. The ratios derived from these markers, such as A/T and T/E, amplify this variability. For example, the A/T ratio in 2023 females reached a mean of 2324.17 with a staggering standard deviation of 15416.32, driven largely by the extremely low T values in some athletes. The maximum A/T in that year was 511010.93, with a median of 230.08, suggesting that most profiles lie within a moderate range, but some of the cases dominate the average, indicative of heavy skew. Such skew is further evident when comparing mean and median values. In 2022, female testosterone had a mean of 6.08 but a median of only 4.94, and in 2023, the male A/T ratio had a mean of 192.94 but a median of just 84.62, which clearly shows the right-skewed distributions with long tails. These statistics justify the use of robust model architectures and possibly normalization during data preprocessing. Ratios such as 5αAdiol/5βAdiol and 5α Adiol/E also show gender-specific patterns and variability. In 2023, males had 5αAdiol/5βAdiol = 0.59 ± 0.38, while females had a slightly lower mean of 0.55 ± 0.36. The 5αAdiol/E ratio for males showed 2.61 ± 1.79, but for females it increased to 3.58 ± 2.76 in 2023, possibly reflecting differential metabolism or clearance rates.

	Male ($n = 13261$)						Female ($n = 6460$)					
Steroid parameter	**Mean±std.**	**Min**	**IQ1**	**Median**	**IQ3**	**Max**	**Mean±std.**	**Min**	**IQ1**	**Median**	**IQ3**	**Max**
A	2474.99 ± 1238.82	26.76	1567.56	2267.92	3165.71	7358.46	1809.45 ± 1073.46	19.38	1021.98	1581.93	2365.10	6131.27
Etio	1944.43 ± 932.77	74.45	1246.67	1792.49	2479.67	5383.35	1981.61 ± 1065.53	19.03	1183.46	1774.19	2582.11	6028.99
E	27.06 ± 18.00	0.32	13.47	22.42	36.25	94.74	7.77 ± 5.62	0.22	3.58	6.26	10.54	31.18
T	28.77 ± 19.84	0.00	14.28	26.25	40.31	104.90	5.13 ± 4.55	0.00	1.92	4.16	7.10	35.27
5αAdiol	53.81 ± 30.76	1.38	31.59	47.64	69.26	201.49	20.65 ± 14.60	0.37	10.70	16.89	26.21	120.72
5βAdiol	125.32 ± 93.77	1.24	56.04	100.63	168.18	530.57	55.97 ± 53.71	1.19	20.22	37.07	72.78	392.04
T/E	1.35 ± 1.01	0.00	0.66	1.14	1.82	5.66	0.87 ± 0.77	0.00	0.31	0.71	1.24	5.27
A/Etio	1.40 ± 0.68	0.05	0.94	1.30	1.74	3.78	0.98 ± 0.45	0.10	0.65	0.90	1.23	2.70
A/T	192.94 ± 283.14	0.00	58.03	86.72	148.33	1887.44	2324.17 ± 15416.32	0.00	230.08	365.39	673.72	511010.93
5αAdiol/5βAdiol	0.59 ± 0.38	0.01	0.31	0.48	0.77	2.00	0.55 ± 0.36	0.00	0.26	0.46	0.76	1.81
5αAdiol/E	2.61 ± 1.79	0.04	1.33	2.10	3.38	10.73	3.58 ± 2.76	0.08	1.72	2.83	4.56	25.81

Table 6. Descriptive statistics of urinary steroid parameters for male ($n = 13261$) and female ($n = 6460$) athletes in the 2023 dataset. The table reports the mean ± standard deviation (std.), minimum (min), first quartile (IQ1), median, third quartile (IQ3), and maximum (max) for each parameter. The data shows significant gender-based differences in concentration levels and variability across most biomarkers.

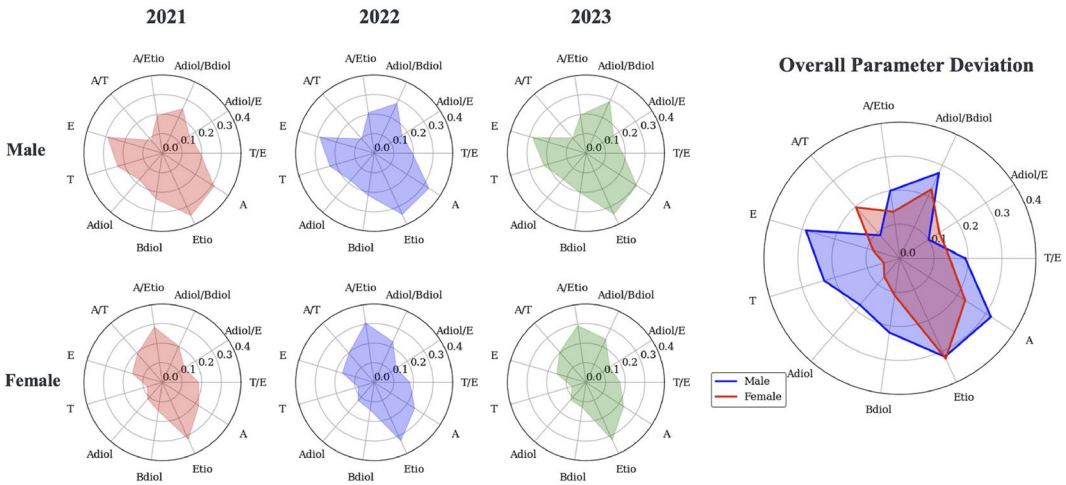

Fig. 2. Radar plots showing the normalized deviations of urinary steroid biomarkers (median) and their ratios across male and female athletes from 2021 to 2023. Each subplot visualizes inter-parameter variability within a specific year, stratified by gender. The rightmost plot summarizes overall parameter deviations between males (blue) and females (red), highlighting consistent gender-specific patterns across the three-year period.

While most biomarkers were relatively stable across years with some small shifts were observed. Male 5β Adiol levels decreased from 143.58 ± 104.78 (2021) to 125.32 ± 93.77 (2023), suggesting a drift in athlete profiles. Female Etio increased from 1973.44 ± 1069.75 (2021) to 2024.14 ± 1113.92 (2022), before slightly dropping again. Despite these variations, the overall biomarker structure remained remarkably consistent, reinforcing the feasibility of training joint models across years without major retraining. Though we do not explicitly report correlations, the similarity in mean and dispersion between Etio and A suggests a high positive correlation in both genders. In 2023, for males: A = 2474.99 ± 1238.82 and Etio = 1944.43 ± 932.77; for females: A = 1809.45 ± 1073.46 and Etio = 1981.61 ± 1065.53. Lastly, the minimum and maximum values provide crucial cues for screening. The high ratios like T/E> 5 or A/T > 10,000 (even in females) are rare but normal and do not warrant forensic investigation. Such values support the use of both threshold-based (PCA) and learning-based (CNN) methods, especially when seeking to automate the review of hundreds of thousands of samples annually.

As shown in Fig. 2, the normalized radar plots show the inter-parameter deviations in urinary steroid biomarkers across male and female athletes over three consecutive years (2021–23). Each radar plot captures the relative normalized contribution of biomarkers and their ratios for a specific gender and year. The normalization was performed per parameter to remove scale biases and emphasize relative inter-feature variability. The rightmost radar plot aggregates deviations across years, highlighting persistent gender-based differences. The radar plots show consistent separation between male and female profiles across all years.

We observed that the male profiles dominate in A, T, E, and 5βAdiol, while female profiles have higher A/T, and 5αAdiol/E. This differential landscape validates the hypothesis that any similarity detection system must consider gender-aware differences or apply feature-wise normalization. The radar plots also show that some parameters exhibit cross-gender consistency, particularly Etio and A/Etio ratio, which show less deviations in both genders and all years. This makes them valuable as anchoring features in unsupervised similarity detection, where class boundaries are unknown. Meanwhile, derived ratios like A and 5αAdiol/5βAdiol show high cross-gender divergence and thus are well suited for discriminative tasks.

To further investigate inter-individual variability and gender-based patterns in urinary steroid biomarkers, we analyzed the distribution of each parameter using histogram plots (Fig. 3). These histograms provide insight into the range, skewness, and overlap of each parameter's distribution and help identify parameters that exhibit strong gender specificity. Across nearly all biomarkers, we observe a right-skewed distribution in both male and female populations, with the majority of data concentrated at lower values and a long tail extending toward higher values. This skewness is especially apparent in parameters like T, A, 5βAdiol, and most ratio features such as A/T and 5αAdiol/E. Importantly, the degree and direction of overlap between male and female distributions vary by feature, with some parameters showing near-total separation and others demonstrating considerable shared range. Among the steroid concentrations, T displays one of the most clearly bimodal distributions. Males dominate the distribution beyond ~10 ng/mL, with the mode around 20-25 ng/mL and values extending up to 150 ng/mL. In contrast, the female distribution peaks below 5 ng/mL and rapidly decays. The overlap is minimal, which reinforces the utility of testosterone as a strong discriminative biomarker between genders and a potential feature in similarity assessment.

Conversely, certain metabolites such as Etio and 5βAdiol show greater overlap between genders. Although males still exhibit higher overall concentrations, the distributions of Etio and 5βAdiol in females are wider and span across much of the male range, though with lower frequencies. This suggests that while Etio is included in key ratios like A/Etio, it may not offer the same standalone discriminative power. Such partially overlapping distributions can be informative for identifying subtle similarities or potential mismatches, particularly in edge cases or manipulated samples. The derived ratios provide additional insight. For example, the T/E ratio is strongly right-skewed in both populations, but males show a wider spread and higher frequency above the 1.5 threshold, which makes it the highest sensitivity to detect testosterone administrations. A/T shows a massive right tail in males, driven by extremely low testosterone values, resulting in some ratios exceeding 1000. This long tail highlights the importance of ratio-based features in uncovering extreme phenotypes or potential measurement errors. Similarly, the 5αAdiol/E ratio shows a broader distribution in females, with values reaching above 20, compared to a tighter distribution in males around 1-5.

Principal component analysis

We applied on the 2023 dataset and visualized the results in PCA-reduced space for a selected test sample (ID: 47933). Figure 4 shows how sample similarity evolves with varying distance thresholds D_T ranging from 0.95 to 0.98. Each panel shows a 3D scatter plot of the principal component space, with red dots denoting samples identified as similar to the testing sample under the respective threshold, and the green dot representing the testing sample itself.

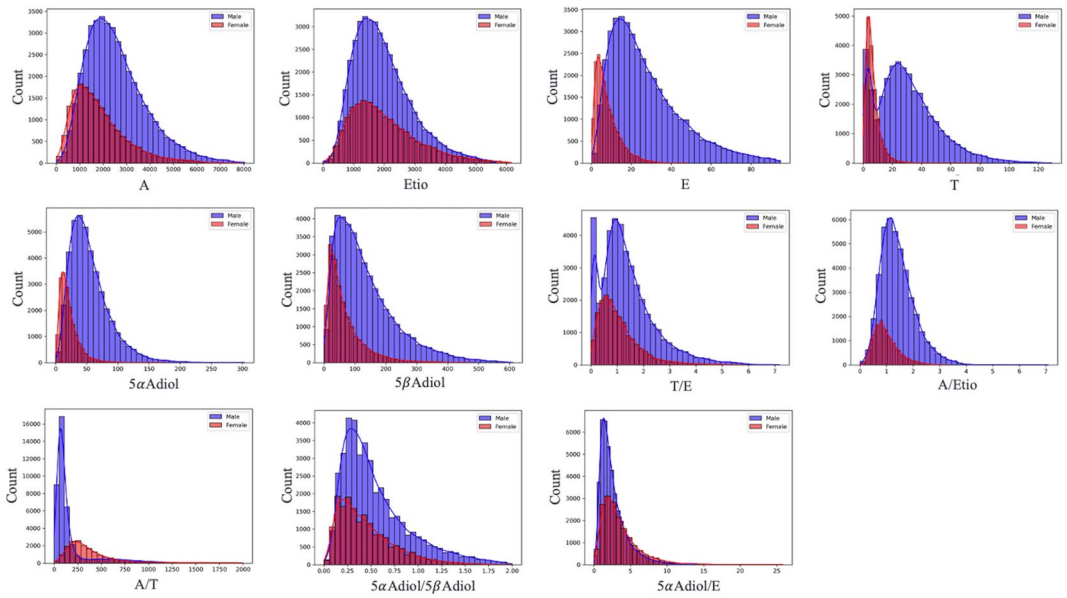

Fig. 3. Distribution plots of urinary steroid biomarkers and their ratios for male (blue) and female (red) athletes. Each subplot shows the distribution for a steroid parameter which highlights distinct gender-based differences in steroid patterns and ratio distributions, providing insight into physiological variation and potential implications for sample similarity detection.

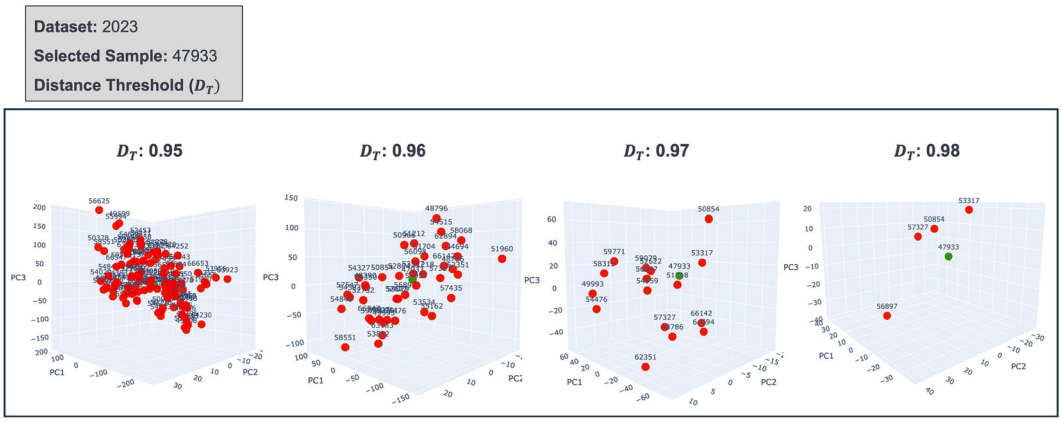

Fig. 4. Visualization of sample similarity detection using PCA for a selected test sample (ID: 47933) from the 2023 dataset. Each 3D scatter plot shows the distribution of steroid profiles in PCA-reduced space with increasing distance thresholds D_T from 0.95 to 0.98. Red dots indicate samples identified as similar under the given threshold, while the green dot represents the selected testing sample. As D_T increases, more samples fall within the similarity region, reflecting the trade-off between sensitivity and specificity in the distance-based detection approach.

We observed that at the lowest threshold ($D_T = 0.95$), a relatively dense cloud of red points surrounds the green test sample. This indicates that with a lower threshold, the method adopts a more permissive similarity region, allowing a larger number of samples to be classified as similar. However, many of these samples are distributed widely in the PCA space, suggesting that the sensitivity is high but the specificity may be compromised, as some distant samples may not truly reflect close biological similarity. As the threshold is increased to $D_T = 0.96$, the set of similar samples becomes more compact. We observe that many of the previously marked red points are excluded, and only a subset remains in proximity to the test sample. This refined selection suggests a tightening of the similarity criteria, increasing specificity while slightly reducing sensitivity. The spatial cluster of red points begins to center more clearly around the green sample, indicating that samples closer in principal component space are being prioritized. At $D_T = 0.97$, the similarity region becomes even more selective. Only those samples that are mathematically closer to the testing sample remain within the L1 distance threshold. Approximately a dozen samples are retained, forming a localized cluster in the PCA space around the test sample. This scenario balances sensitivity and specificity effectively, making it suitable for high-confidence similarity detection where false positives must be minimized. The density and cohesion of this cluster also indicate that the profiles are not only mathematically close but potentially biologically similar, given their alignment along the major principal axes. At the highest threshold shown ($D_T = 0.98$), the similarity region becomes extremely narrow. Only a few samples fall within this high-precision boundary. While this setting minimizes the risk of false positives, it may be overly conservative in some practical scenarios, potentially excluding relevant samples that lie just beyond the threshold. This reflects a classic trade-off: increasing D_T improves specificity but reduces coverage, whereas lowering D_T does the opposite. By adjusting the distance threshold D_T, we can flexibly control the tightness of the similarity region to prioritize broad recall (e.g., in screening) or precision (e.g., in forensic verification).

Performance analysis

We conducted experiments across all the datasets and compared its performance against multiple baseline models. These include models such as Support Vector Machines (SVM)[26], Logistic Regression (LR)[27], Random Forest (RF)[28], Extra Trees (ET)[29], and Extreme Gradient Boosting (XGB)[30]. The evaluation is performed based on accuracy, sensitivity, specificity and area under ROC curve.

Overall training performance

Firstly, we performed evaluations on the held-out testing data (20%) from 2021–23 for male and female athletes with $\pm 15\%$ manipulations on samples. As shown in Table 7, the CNN model consistently outperforms all baseline models across all four metrics for all three years. For 2021 data, CNN achieves perfect sensitivity (1.00 ± 0.02), with accuracy (0.99 ± 0.01), specificity (0.99 ± 0.01), and AUC (0.99 ± 0.02). This trend continues for 2022 and 2023, where CNN maintains similarly high performance, with sensitivity and specificity both at or above 0.98, and AUC consistently close to 0.99. This balance of high sensitivity and specificity is particularly important in anti-doping applications, where both false negatives (missed matches) and false positives (wrongly matched samples) must be minimized.

In comparison, baseline models such as SVM and RF also show high sensitivity (>0.97 across years) and respectable AUCs (>0.95), but their specificity lags behind the CNN. For example, SVM achieves a specificity of only 0.90 ± 0.03 in 2023, and RF achieves 0.97 ± 0.01. This implies that while these models are effective at detecting similar samples (high sensitivity), they are more prone to false positives compared to the CNN model. ET and XGB perform competitively, with high scores across all metrics, but CNN slightly edges them

Model	2021				2022				2023			
	Acc	Sens	Spec	AUC	Acc	Sens	Spec	AUC	Acc	Sens	Spec	AUC
SVM	0.95 ± 0.01	0.98 ± 0.02	0.91 ± 0.01	0.96 ± 0.00	0.95 ± 0.03	0.97 ± 0.01	0.91 ± 0.02	0.95 ± 0.01	0.95 ± 0.02	0.97 ± 0.00	0.90 ± 0.03	0.95 ± 0.01
LR	0.51 ± 0.03	0.81 ± 0.01	0.22 ± 0.02	0.52 ± 0.01	0.49 ± 0.00	0.51 ± 0.02	0.46 ± 0.01	0.49 ± 0.03	0.48 ± 0.01	0.57 ± 0.02	0.40 ± 0.01	0.48 ± 0.00
RF	0.99 ± 0.02	0.98 ± 0.01	0.97 ± 0.01	0.98 ± 0.00	0.98 ± 0.03	0.99 ± 0.01	0.97 ± 0.02	0.98 ± 0.01	0.98 ± 0.01	0.98 ± 0.00	0.97 ± 0.01	0.98 ± 0.03
ET	0.98 ± 0.01	0.99 ± 0.02	0.96 ± 0.03	0.98 ± 0.01	0.98 ± 0.00	0.97 ± 0.01	0.96 ± 0.02	0.98 ± 0.01	0.98 ± 0.02	0.97 ± 0.01	0.96 ± 0.00	0.98 ± 0.01
XGB	0.98 ± 0.03	0.98 ± 0.01	0.98 ± 0.02	0.98 ± 0.01	0.98 ± 0.00	0.99 ± 0.02	0.98 ± 0.01	0.98 ± 0.03	0.98 ± 0.01	0.97 ± 0.00	0.99 ± 0.01	0.98 ± 0.02
CNN	**0.99 ± 0.01**	**1.00 ± 0.02**	**0.99 ± 0.01**	**0.99 ± 0.02**	**0.99 ± 0.03**	**0.99 ± 0.01**	**0.98 ± 0.01**	**0.99 ± 0.03**	**0.98 ± 0.00**	**0.98 ± 0.00**	**0.99 ± 0.01**	**0.98 ± 0.00**

Table 7. Performance comparison of different models on held-out testing-data from 2021–23. The evaluaton metrics are accuracy (Acc), sensitivity (Sens), specificity (Spec), area under ROC curve (AUC). While models such as SVM, RF, and XGB achieve high sensitivity and AUC, the CNN consistently outperforms them across all three years (2021–2023), particularly in maintaining high specificity alongside sensitivity.

Model	Male				Female			
	Acc	Sens	Spec	AUC	Acc	Sens	Spec	AUC
SVM	0.97 ± 0.06	0.95 ± 0.01	0.98 ± 0.07	0.98 ± 0.03	0.93 ± 0.01	0.97 ± 0.09	0.88 ± 0.02	0.94 ± 0.01
LR	0.51 ± 0.09	0.54 ± 0.41	0.47 ± 0.36	0.64 ± 0.12	0.53 ± 0.04	0.60 ± 0.47	0.45 ± 0.40	0.65 ± 0.14
RF	0.98 ± 0.08	0.97 ± 0.05	0.99 ± 0.02	0.96 ± 0.09	0.97 ± 0.01	0.95 ± 0.01	0.98 ± 0.06	0.97 ± 0.10
ET	0.98 ± 0.02	0.98 ± 0.04	0.99 ± 0.07	0.95 ± 0.03	0.98 ± 0.05	0.98 ± 0.08	0.98 ± 0.01	0.96 ± 0.09
XGB	0.98 ± 0.10	0.96 ± 0.06	0.98 ± 0.04	0.97 ± 0.01	0.97 ± 0.01	0.94 ± 0.01	0.99 ± 0.07	0.98 ± 0.05
CNN	**0.99 ± 0.02**	**0.98 ± 0.01**	**1.00 ± 0.00**	**0.99 ± 0.00**	**0.98 ± 0.02**	**0.97 ± 0.01**	**0.99 ± 0.01**	**0.99 ± 0.00**

Table 8. Performance comparison of different models on laboratory-confirmed samples, stratified by gender. The evaluaton metrics are accuracy (Acc), sensitivity (Sens), specificity (Spec), area under ROC curve (AUC). The CNN consistently outperforms all other methods across both male and female samples, achieving better specificity and AUC.

out consistently, particularly in specificity. On the other hand, LR underperforms significantly across all metrics and years. In 2021, LR achieves an accuracy of only 0.51 ± 0.03 and a specificity of just 0.22 ± 0.02. The model shows limited improvement in subsequent years, indicating its inability to capture the complex non-linear patterns necessary for reliable similarity detection in longitudinal biosample data. The AUC remains low (0.48-0.52), further confirming its unsuitability for this task. While dataset characteristics may shift from year to year due to varying athlete demographics or laboratory conditions, CNN's performance remains stable. From 2021 to 2023, the variation in its metrics remains within ±0.02, demonstrating high reliability and generalizability. This robustness suggests that the CNN has learned latent structural representations of steroid profiles that are invariant to dataset drift. It is an essential property for large-scale, real-world anti-doping monitoring as here samples analyzed in different laboratories over a time frame of years may have to be compared.

Performance on confirmation samples

To validate the robustness of our model under real-world conditions, we conducted a evaluation on laboratory-confirmed samples. These samples represent ground-truth instances of biological similarity and are stratified by gender to examine model consistency across physiological contexts. Table 8 compares the performance of CNN against different baseline models. The CNN model achieves the improved performance across all metrics, underlining its capability to identify biologically similar samples even in the presence of measurement variability. For male samples, CNN yields an accuracy of 0.99 ± 0.02, sensitivity of 0.98 ± 0.01, perfect specificity (1.00 ± 0.00), and an AUC of 0.99 ± 0.00. These values indicate that the model is highly effective at identifying truly similar profiles without falsely labeling dissimilar ones, which is important in anti-doping where misclassifications can lead to unnecessary follow-up tests. The perfect specificity suggests that the CNN model has learned to capture subtle, biologically meaningful cues that distinguish true matches from background noise, possibly by exploiting local dependencies and latent feature interactions learned during convolutional processing. For female samples, the CNN achieves 0.98 ± 0.02 accuracy, 0.97 ± 0.01 sensitivity, 0.99 ± 0.01 specificity, and 0.99 ± 0.00 AUC. This consistency is particularly notable given the higher inter-individual variability and lower average steroid concentrations observed in female athletes. Female profiles often exhibit more extreme ratio values (e.g., A/T and 5αAdiol/E), which can complicate threshold-based classification. CNN's ability to generalize to these conditions suggests that it is not overfitting to dominant male-driven patterns but rather learning gender-independent structural representations of similarity in the biomarker space. While baseline models such as RF, ET, and XGB also perform well (with accuracies and AUCs above 0.95 in both male and female samples), they consistently fall short of the CNN in at least one metric (mostly specificity). The minor drop in specificity among tree-based models may be attributed to their tendency to capture shallow rules that do not generalize well when noise or

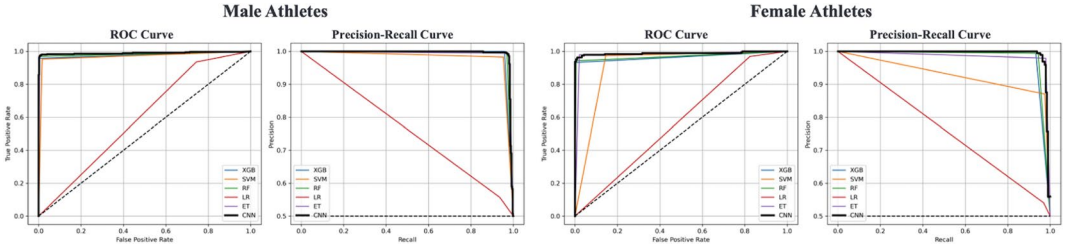

Fig. 5. ROC and Precision-Recall (PR) curves for male (left) and female (right) athletes, comparing the performance of CNN with baseline models. In both gender groups, the CNN (black line) demonstrates the best overall trade-off between true positive rate and false positive rate, as well as higher precision at high recall levels. The PR curves especially highlight the CNN's ability to maintain high confidence in positive predictions across different recall thresholds.

physiological variability is present. In contrast, the CNN's capacity to model fine-grained local patterns and non-linear interactions provides better discrimination.

In contrast, LR performs poorly on both male and female confirmation samples, with accuracy as low as 0.51 ± 0.09 in males and 0.53 ± 0.04 in females. Its sensitivity and specificity are inconsistent and unstable, marked by large standard deviations (e.g., specificity in males: 0.47 ± 0.36), reflecting the model's inability to generalize in this complex domain. These findings further reinforce that linear models are insufficient for capturing the nuanced biomarker interactions required for accurate similarity detection. The poor performance of LR may also reflect the high dimensionality and non-Gaussian distribution of the steroid data, which violates key LR assumptions. Importantly, these results were obtained on real-world, lab-validated samples, making them the most clinically relevant benchmark. The high performance of CNN across both male and female subgroups suggests that the model is not only accurate but robust to the kind of biological, procedural, and analytical variability encountered in operational settings.

Precision-recall analysis

We plotted the Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves for all the models (Fig. 5). These curves offer additional insight beyond scalar metrics like accuracy or AUC, capturing how sensitivity and precision behave across different decision thresholds. For male athletes, the ROC curve shows that all models except LR achieve high true positive rates at very low false positive rates. The CNN consistently outperforms all baselines, reflecting high sensitivity and specificity across different thresholds. Tree-based algorithms like XGB, RF, and ET also demonstrate strong ROC performance, with curves nearly overlapping CNN's but diverging slightly at lower thresholds. In contrast, LR shows a near-diagonal ROC curve, suggesting it performs close to random chance under threshold variations. In the PR curve for males, the CNN maintains precision above 0.99 until nearly the highest recall values, indicating a minimal drop in confidence even as more positive samples are retrieved. ET and RF also achieve excellent PR curves, while SVM and XGB follow closely. LR, however, suffers from rapidly decreasing precision with increasing recall, reflecting its high false positive rate and low discriminative power in practice.

For female athletes, the ROC curve for CNN dominates the plot. Even with lower testosterone concentrations and greater ratio variability in female profiles, CNN maintains higher separation of classes. Tree-based algorithms like ET and RF again perform well, but with slightly reduced AUC compared to the male cohort. SVM shows a sharp drop in ROC performance in the early range of FPR, and LR again performs worst, tracking along the diagonal with limited true discrimination. The PR curve for females highlight the robustness of CNN under more variable biological conditions. XGB also shows strong PR characteristics, but begins to degrade earlier than CNN. These plots collectively shows that the CNN model provides the most balanced performance and ability to maintain high performance across all threshold values and both genders makes it particularly well-suited for large-scale, real-world deployment in anti-doping systems.

Conclusion

In this paper, we present the challenge of detecting identical or reused urine biosamples within the context of anti-doping analysis, specifically focusing on the steroidal module of the Athlete Biological Passport. In scenarios where athletes attempt to circumvent doping detection by substituting freshly collected samples containing prohibited substances with earlier, drug-free sample (either their own or from another athlete), such manipulations can be difficult or even impossible to identify using conventional techniques, especially when the steroid profiles fall within an expected biological range.

To address this issue, we proposed a similarity detection framework capable of screening large-scale steroid profile data for highly similar or identical entries. We use a convolutional neural network, trained on synthetic and confirmed sample pairs, to learn complex pairwise relationships between profiles under biologically realistic variations. By leveraging simulated measurement uncertainty and laboratory-confirmed samples, the CNN model was able to generalize well and distinguish actual similarity from natural inter-individual variability. Across all evaluation benchmarks including gender-specific analysis, and real-world confirmation samples, the CNN model consistently outperformed all the baseline models. This research also shows that the underlying steroid profile data is not overly complex, as even standard machine learning methods such as Random Forests

and Extra Trees achieved reasonably good results. However, the CNN demonstrated consistently higher performance and greater robustness across datasets and laboratory settings, making it a more reliable choice for large-scale similarity detection. This work contributes a scalable and data-driven methodology to support forensic investigations and routine doping control, particularly in identifying suspicious entries in ABP databases. By integrating it into existing workflows, APMU and WADA can improve their ability to detect sample reuse, strengthen the evidentiary value of longitudinal profiles, and improve the overall integrity of sports drug testing.

Limitations

Despite the promising results, several limitations of the present study must be acknowledged, both from methodological and practical perspectives. *i) Simulation of variability vs. real-world complexity:* While synthetic data generation allowed us to simulate realistic variability in measurement values (±15%), this controlled noise may not fully capture the range of biological and procedural deviations observed in real-world doping control. Factors such as storage conditions, time delays, and lab-specific sample preparation methods can introduce additional, non-random variation not accounted for in the training process. *ii) Limited interpretability:* Although the CNN achieved higher performance across several metrics, it functions as a black-box model, offering limited transparency regarding the specific features that influence its similarity predictions. In contexts like forensic science and regulatory decisions, where interpretability and auditability are essential, this may limit the model's acceptance unless combined with explainability techniques. This potential limitation may be addressed by further investigations of suspicious pairs of samples including (but not limited to) the analysis of the presence and abundance of recreational drugs or forensic parameters like urinary concentration of different salts or urea.

Data availability

The data that support the findings of this study are available from the World Anti-Doping Agency (WADA), but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the World Anti-Doping Agency (WADA).

Received: 26 November 2025; Accepted: 4 March 2026
Published online: 17 March 2026

References

1. Lauritzen, F. & Solheim, A. The purpose and effectiveness of doping testing in sport. *Front. Sports Active Liv.* **6**, 1386539 (2024).
2. Thevis, M. *Mass spectrometry in sports drug testing: Characterization of prohibited substances and doping control analytical assays* (Wiley, 2010).
3. McLaren, R. H. WADA Investigation of Sochi Allegations – The Independent Person Report. https://www.bundestag.de/resource/blob/503234/1bd2a1ed434a3474927548d1d58d23eb/McLaren-Report.pdf (2016).
4. WADA. Wada athlete biological passport operating guidelines (2024). https://www.wada-ama.org/en/resources/world-anti-doping-program/athlete-biological-passport-abp-operating-guidelines#resource-download.
5. Sottas, P.-E. et al. Bayesian detection of abnormal values in longitudinal biomarkers with an application to t/e ratio. *Biostatistics* **8**, 285–296. https://doi.org/10.1093/biostatistics/kxl009 (2007).
6. Rahman, M. R. *et al.* Sacnn: Self attention-based convolutional neural network for fraudulent behaviour detection in sports. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI 24)* (2024).
7. Rahman, M. R. *et al.* Data analytics for uncovering fraudulent behaviour in elite sports. In *2022 International Conference on Information System (ICIS)* (2022).
8. Thevis, M. et al. Detection of manipulation in doping control urine sample collection: A multidisciplinary approach to determine identical urine samples. *Anal. Bioanal. Chem.* **388**, 1539–1543. https://doi.org/10.1007/s00216-006-1112-z (2007).
9. Thevis, M., Geyer, H., Sigmund, G. & Schänzer, W. Sports drug testing: Analytical aspects of selected cases of suspected, purported, and proven urine manipulation. *J. Pharm. Biomed. Anal.* **57**, 26–32. https://doi.org/10.1016/j.jpba.2011.09.002 (2012).
10. Committee, I. P. Russian sochi 2014 paralympian found to have committed an anti-doping rule violation (2025).
11. World Anti-Doping Agency. 2022 anti-doping testing figures – executive summary. Tech. Rep., World Anti-Doping Agency (2024).
12. Piper, T. et al. Current insights into the steroidal module of the athlete biological passport. *Int. J. Sports Med.* **42**, 863–878. https://doi.org/10.1055/a-1481-8683 (2021).
13. Rahman, M. R. *et al.* AI-based approach for improving the detection of blood doping in sports. arXiv:2203.00001 (2022).
14. Rahman, M. R. *et al.* Modelling metabolism pathways using graph representation learning for fraud detection in sports. In *2023 IEEE International Conference on Digital Health (ICDH)*, 158–168, https://doi.org/10.1109/ICDH60066.2023.00031 (2023).
15. Rahman, M. R. *et al.* Detection of erythropoietin in blood to uncover doping in sports using machine learning. In *2022 IEEE International Conference on Digital Health (ICDH)*, 193–201, https://doi.org/10.1109/ICDH55609.2022.00038 (2022).
16. WADA Science/EAAS Working Group. Wada technical document – td2021eaas (2022). Available at: https://www.wada-ama.org/sites/default/files/2022-01/td2021eaas_final_eng_v_2.0.pdf, Accessed 10.07.2024.
17. Agency, W. A.-D. Td2023apmu: Athlete passport management unit requirements and procedures. Tech. Rep., World Anti-Doping Agency (2022). Technical Document published on November 17, 2022.
18. Jolliffe, I. T. *Principal Component Analysis*. Springer Series in Statistics (Springer, 2002).
19. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324. https://doi.org/10.1109/5.726791 (1998).
20. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)* (2015).
21. World Anti-Doping Agency. *World Anti-Doping Code*. World Anti-Doping Agency, Montreal, Quebec, Canada (2021). Effective 1 January 2021.
22. World Anti-Doping Agency. *International Standard for Laboratories*. World Anti-Doping Agency, Montreal, Quebec, Canada (2021). International Standard, effective 1 January 2021.
23. Ayotte, C., Goudreault, D. & Charlebois, A. Testing for natural and synthetic anabolic agents in human urine. *J. Chromatogr. B Biomed. Appl.* **687**, 3–25. https://doi.org/10.1016/S0378-4347(96)00032-1 (1996).

24. Van Renterghem, P., van Eenoo, P., Geyer, H., Schänzer, W. & Delbeke, F. T. Reference ranges for urinary concentrations and ratios of endogenous steroids, which can be used as markers for steroid misuse, in a caucasian population of athletes. *Steroids* **75**, 154–163. https://doi.org/10.1016/j.steroids.2009.11.008 (2010).
25. Saad, K. et al. Population reference ranges of urinary endogenous sulfate steroids concentrations and ratios as complement to the steroid profile in sports antidoping. *Steroids* **152**, 108477. https://doi.org/10.1016/j.steroids.2019.108477 (2019).
26. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297. https://doi.org/10.1007/BF00994018 (1995).
27. Hosmer, D. W., Lemeshow, S. & Sturdivant, R. X. *Applied Logistic Regression* (Wiley, 2013), 3rd edn.
28. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32. https://doi.org/10.1023/A:1010933404324 (2001).
29. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **63**, 3–42. https://doi.org/10.1007/s10994-006-6226-1 (2006).
30. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794, https://doi.org/10.1145/2939672.2939785 (ACM, 2016).

Acknowledgements
We thank Mattes Warning and Patrick Reck for their support in conducting the analysis.

Author contributions
M.R. developed the research idea, designed the methodology, conducted the analyses, and prepared the manuscript. T.P. was responsible for data preparation and provided manuscript feedback. W.M. supervised the project, provided conceptual guidance, and reviewed all stages of the work. M.T. reviewed the manuscript and approved the submission. All authors reviewed and approved the final version of the paper.

Funding
This study was conducted as part of the MARVIN project. We gratefully acknowledge the World Anti-Doping Agency (WADA) for funding and supporting this work.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to M.R.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.