# Modelling Metabolism Pathways using Graph Representation Learning for Fraud Detection in Sports

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

Mohammed Hussain
*German Research Center for Artificial Intelligence*
Germany
mohammed.hussain@dfki.de

Thomas Piper
*German Sports University Cologne*
Germany
t.piper@biochem.dshs-koeln.de

Hans Geyer
*German Sports University Cologne*
Germany
h.geyer@biochem.dshs-koeln.de

Tristan Equey
*World Anti-Doping Agency (WADA)*
Canada
tristan.equey@wada-ama.org

Norbert Baume
*World Anti-Doping Agency (WADA)*
Canada
norbert.baume@wada-ama.org

Reid Aikin
*World Anti-Doping Agency (WADA)*
Canada
reid.aikin@wada-ama.org

Wolfgang Maass
*German Research Center for Artificial Intelligence*
Germany
wolfgang.maass@dfki.de

***Abstract*—Modelling biological pathway plays an important role in understanding different processes for decision making, especially in forensic investigations on doping activities in sports. Recently, the issue of sample swapping has arisen as a potential fraudulent behaviour by athletes to avoid a positive doping test result. The current detection models neglect an important factor, i.e., leveraging the steroid metabolism pathway of the human body. The spatial relationships between different metabolites within the steroid metabolism pathways are important and cannot be merely treated as linear correlations when assessing similarities among the samples obtained from athletes. To address this challenge, we propose the GRAMP model based on graph representation learning to incorporate domain knowledge into the model decision for the detection of sample swapping. Our model takes into account the spatial structural dependencies of different metabolites using a graph attention mechanism and generates high-level embeddings to detect fraudulent behaviour. We evaluate our approach through extensive experiments on real-world datasets and find that our proposed model outperforms existing state-of-the-art models for fraud detection tasks in sports, demonstrating the effectiveness of our approach and its potential impact on decision making.**

***Index Terms*—Metabolism Pathways, Graph Attention Network, Fraud Detection, Sample Swapping, Doping, Sports**

## I. Introduction

Modelling biological pathways is an essential aspect of bioinformatics and biochemical research. Biological pathways represent a series of interconnected molecular events that occur within a cell to carry out specific functions, such as signal transduction, metabolism, and gene regulation [19]. Understanding these pathways can provide insights into the underlying mechanisms of various cellular processes and aid in the discovery of novel therapeutic targets. There are several approaches to modelling biological pathways, ranging from qualitative to quantitative methods [12], [15], [30]. However, these methods have challenges like parameter estimation, model complexity, dynamic behaviour, etc. Therefore, using these methods for modelling biological pathways leads to inaccurate predictions and limited applicability.

Many forensic investigations primarily focus on analysing these biological pathways to identify the fraudulent behaviour of the individual, especially doping activities in sports [4]. Recent investigation at the 2014 Olympics Games in Sochi discovered a new form of fraudulent behaviour by athletes. Athletes were found attempting to replace their doping samples with clean samples obtained from other individuals to avoid positive test results, known as 'sample swapping' [17]. This fraudulent activity poses a substantial challenge in the forensic investigations of the World Anti-Doping Organisation (WADA) and other organisations.

WADA maintains a longitudinal profile for every athlete, which includes a record of all the samples collected from that athlete so far for the purpose of doping tests. Identifying sample swapping in sports events can be a difficult task, and the conventional method involves conducting DNA analysis on all samples [16], which is costly and time-consuming. Furthermore, the majority of instances involving sample swapping remain undetectable. Alternative methods, such as monitoring each sample and comparing it to the athlete's reference range to detect abnormally high values are available [22], [25], [32]. In addition, machine learning has attracted considerable attention for detecting doping activities.

[24], [32] Nevertheless, these approaches neglect an important factor, i.e., steroid metabolism pathways [29]. In other words, the spatial relationship of different metabolites in the steroid metabolism pathways of the athlete is essential to consider these dependencies when comparing similarities within an athlete's longitudinal profile. Therefore, there is a need for a better method that incorporates the information about domain knowledge into the model decision making.

Over the past decade, several new scenarios from sciences or everyday life have benefited from formulating a relationship between entities as a graph. Therefore, graph networks have become increasingly popular in modelling complex systems due to their ability to capture intricate relationships [34]. They can be used to model complex real-world networks like biological pathways, where vertices represent biological entities, and edges indicate underlying connectivity [11]. Employing graph networks to model domain knowledge facilitates comprehensive coverage of essential properties and theories in the field. Additionally, it helps to comprehend the semantics in pathways, such as the functionalities among data and the species associated with the data. Therefore, in this paper, we propose the following research questions:

- **RQ1:** What is the potential impact of leveraging and integrating domain knowledge of the steroid metabolism pathway into the machine learning model for improving decision making?
- **RQ2:** How can graph representation learning be effectively employed for modelling the steroid metabolism pathway?

To answer these questions, we propose a method to incorporate the spatial relationships between different metabolites and leverage their intricate connections into the model. By doing so, we aim to enhance the understanding and predictive capabilities of machine learning models for better decision making. The key contributions of this work are summarised as follows:

- We present a novel GRAph-based modelling for Metabolism Pathway (GRAMP model), which is capable of integrating the domain knowledge of biological pathways into a machine model. It is comprised of an attention mechanism that captures the direct relationships between different metabolites in the metabolism pathway to improve decision making.
- Unlike previous solutions, we propose a method that leverages the spatial and temporal relationship of steroid metabolism to achieve a more informative representation. To the best of our knowledge, this is the first time that a fraud detection problem in sports has been addressed by considering metabolism pathways.
- We focused on a particular fraud detection problem in sports, i.e., sample swapping. Our method is extensively evaluated on a real-world dataset collected by anti-doping organisations and laboratories. The experimental results show the efficacy of our proposed model, which could detect more fraud athletes with relatively high specificity compared with state-of-the-art baseline models.

## II. Related Work

In this section, we discuss the related research work and the state-of-the-art methods in the following categories:

### A. Graph Representation Learning

Graph representation learning (GRL) [40] automates the discovery of meaningful vector representations for nodes, edges, or entire graphs to facilitate downstream graph mining applications. There are three main groups of GRL methods: (1) network embedding models [7], [9], [21], which preserve the proximities among contextual nodes to capture graph structure information; (2) graph neural networks (GNNs) [14], [35], [39], which aggregate neighbour feature information to learn node embeddings; and (3) knowledge graph embedding methods [3], [6], [31], which model the acceptability score of each fact triplet to learn node and edge (i.e., entity and relation) embeddings by constructing the graph as a collection of fact triplets. The GRL backbone is most commonly built using GNNs, which are currently the state-of-the-art in GRL. Recent advancements in GNNs, such as Graph Convolutional Networks (GCNs) [14], Graph Attention Networks (GATs) [35], and GraphSAGE [10], have further improved their expressive power and scalability. GATs incorporate an attention mechanism to calculate the weights of node neighbourhoods during the aggregation of feature information. By considering the correlations between different samples, it effectively captures the interdependencies and relationships within the data. GraphSAGE is a semi-supervised model that learns node embeddings by sampling neighbouring nodes and aggregating their features using functions like mean or max pooling.

### B. Fraudulent Detection in Sports

Anti-doping organisations have long struggled to fight against fraudulent activities in sports, such as doping. Recently, the use of machine learning techniques for detecting such activities has gathered significant attention. For example, [32] proposed a Bayesian approach that detects abnormal values in longitudinal profiles based on the reference population. Other studies, including [13], [24], [27], [37], have used various machine learning algorithms to detect anomalous samples in the profile. However, these studies fail to address the issue of sample swapping. Current detection methods for sample swapping involve laboratory-based methods such as gas chromatography-mass spectrometry and DNA-STR analysis [22], [33].

Although [25] proposed a visualisation model for detecting fraudulent behaviour in athletes, this approach neglects the metabolism pathways of different steroid parameters that can be useful in detecting fraudulent activities. In this paper, we propose a novel approach for detecting fraudulent activities in sports, i.e., sample swapping using graph representation learning that incorporates the domain knowledge of metabolism pathways into ML-based decision making. This approach

can assist anti-doping organisations for detecting fraudulent activities in sports.

## III. Preliminaries

### A. Problem Statement

In this paper, we address the problem of the need for a method capable of modelling the knowledge about biological pathways to detect fraudulent behaviour in sports, especially sample swapping by athletes. Specifically, the study aims to develop an approach that incorporates structural information about human steroid metabolism into the decision-making of machine learning models.

### B. Sample

WADA and other anti-doping organisations across the world conduct doping tests throughout the year at various national and international athletic events, which results in large-scale historical blood and urine data for each athlete. A urine sample collected from a given athlete for performing a doping test can be denoted as $x_i = \{f_1, f_2, \cdots, f_k\} \in \mathbb{R}^k$, where $k$ represents the total number of parameters. Each sample contains a set of parameters that reflect the concentration levels of various steroid metabolites in the human metabolism, as listed in Table I.

TABLE I
COMPREHENSIVE LIST OF METABOLISM PARAMETERS PRESENTS IN EACH SAMPLE, ALONG WITH THEIR CORRESPONDING MOLECULAR FORMULAS, REPRESENTING EACH PARAMETER'S CHEMICAL COMPOSITION.

| Parameter | Description | Molecular Formula |
|---|---|---|
| T | Testosterone | $C_{19}H_{28}O_2$ |
| E | Epitestosterone | $C_{19}H_{28}O_2$ |
| Etio | Etiocholanolone | $C_{19}H_{30}O_2$ |
| A | Androsterone | $C_{19}H_{30}O_2$ |
| $5\alpha$Adiol | $5\alpha$-androstane-$3\alpha, 17\beta$-diol | $C_{19}H_{32}O_2$ |
| $5\beta$Adiol | $5\beta$-androstane-$3\alpha, 17\beta$-diol | $C_{19}H_{32}O_2$ |

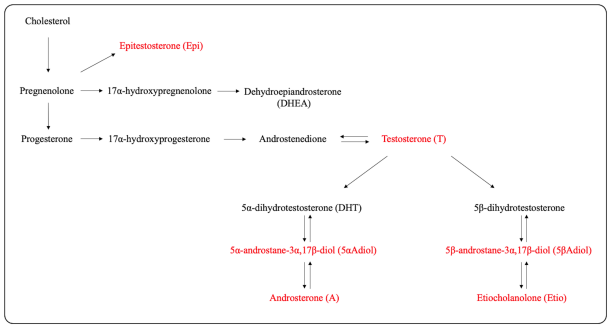


Fig. 1. Simplified human steroid metabolism pathway based on measured urinary steroids, illustrating the intricate interplay of metabolites involved in the synthesis and breakdown of steroid hormones.

### C. Longitudinal Profile

The athlete's longitudinal profile $s$ is defined as a sequence of samples collected over time and is represented by $X^{(p)} = \{x_1, x_2, \cdots, x_n\} \in \mathbb{R}^{n \times k}$, where $n$ is the total number of samples collected. The longitudinal profile is unique to each athlete and helps to track the steroid metabolites and their levels over time in athletes' biological samples, such as urine (or blood). Longitudinal profiling provides a comprehensive understanding of an athlete's steroid metabolism patterns and can be used as a tool for anti-doping agencies to the monitoring of changes in steroid profiles and the detection of potential doping practices or irregularities in athletes' hormone levels.

### D. Fraudulent Behaviour

In this study, we focus on one of the major fraudulent behaviour, i.e., sample swapping, where an athlete exchanges their contaminated sample with a clean sample from another individual. This results in a discrepancy between the sample under consideration, $x_T$, and the rest of the samples in the athlete's longitudinal profile. Therefore, this problem can be well formulated as a graph classification problem where each graph represents an athlete's longitudinal profile. The goal is to classify whether the given graph is suspicious of sample swapping or not. In addition, the prevalence of sample swapping in the real-world situation is very less compared to the clean athletic population. Therefore, this task can be formulated as fraud detection problem.

### E. Steroid Metabolism

Steroid metabolism refers to the processes involved in the synthesis, transportation, and breakdown of steroids in the body. Steroids are lipids that are essential for a variety of physiological processes, including the regulation of metabolism [29]. Steroid hormones, such as Testosterone and estrogen, are synthesised in the gonads and adrenal glands and transported through the bloodstream to target tissues. Epitestosterone is a steroid that is structurally similar to Testosterone but is considered inactive. It is produced in small amounts in the body and is primarily used as a marker for detecting the use of performance-enhancing drugs, such as Testosterone. Etiocholanolone and Androsterone are mainly produced in the adrenal glands and are only partly derived from the liver. $5\alpha$Adiol and $5\beta$Adiol are assumed to be direct metabolites of Testosterone and are, therefore, good markers, while Etiocholanolone and Androsterone represent the end-products of androgen metabolism, and their urinary concentrations are therefore definitely elevated after exogenous Testosterone administrations. Fig. 1 represents a (strongly) simplified pathway which was chosen based on the urinary steroids measured. The real metabolism is much more complicated, involving a multitude of additional enzymatic reactions, intermediate metabolites, and regulatory mechanisms.

Steroid metabolism plays a significant role in athletic doping because it involves the use, detection, and potential abuse of anabolic-androgenic steroids (AAS) by athletes to enhance

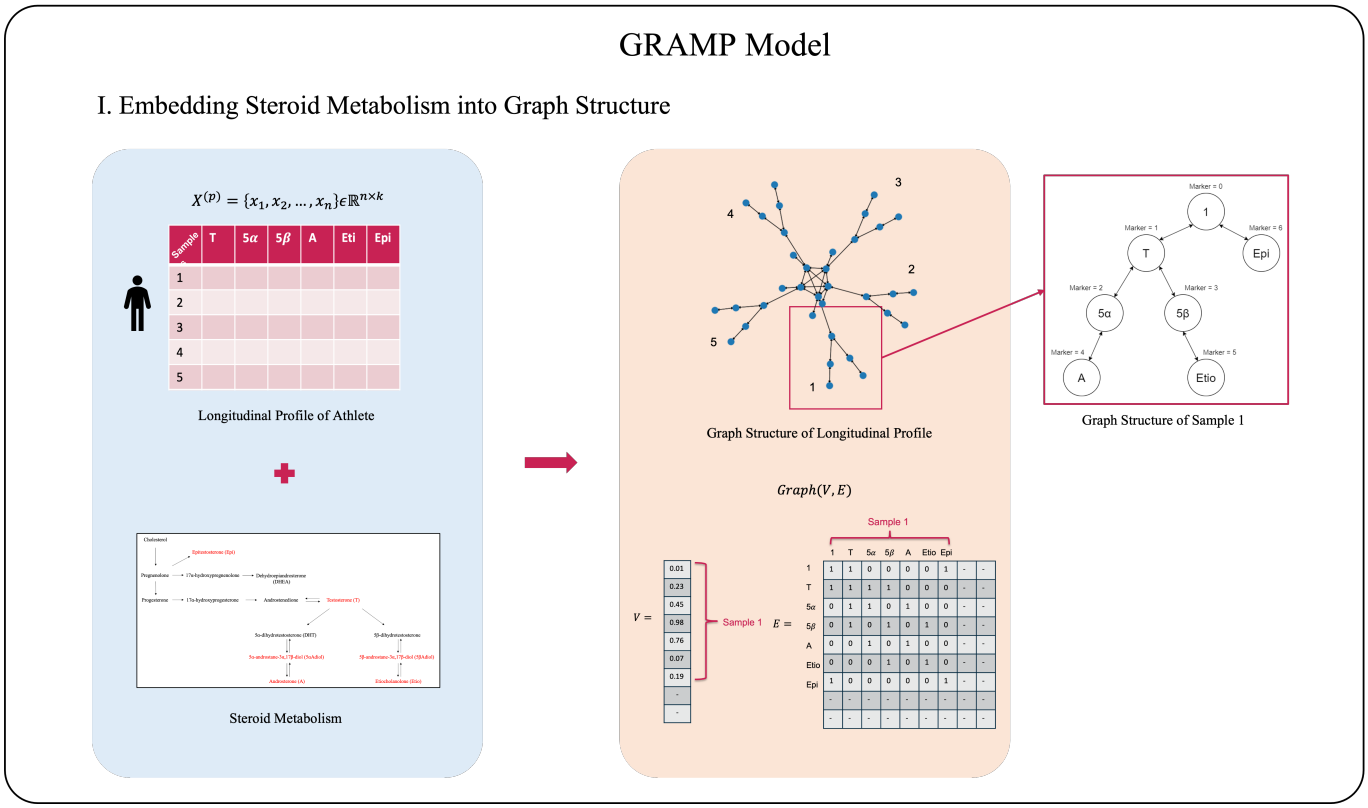


Fig. 2. GRAMP model: Embedding the steroid metabolism pathway into a graph structure by representing metabolites as nodes and capturing their interactions through edges.

their performance [2]. Anabolic steroids are synthetic derivatives of testosterone, a naturally occurring hormone in the body. They are known to promote muscle growth, increase strength and endurance, and improve recovery time. In the context of doping, athletes may misuse steroids in various ways, such as:

- *Performance-Enhancing Substance:* Anabolic steroids are used to enhance athletic performance by increasing muscle mass, strength, and power. This can provide athletes with a competitive edge over their opponents [2].
- *Fat Reduction:* Steroids can promote the breakdown of fat and increase the metabolic rate, leading to reduced body fat percentages. This can be advantageous for athletes participating in sports where weight categories are a factor.
- *Increased Red Blood Cell Production:* Administration of Testosterone can stimulate the production of red blood cells. This can improve oxygen-carrying capacity and endurance performance [28].

The significance of steroid metabolism in athlete doping lies in the detection and prevention of illicit usage. Anti-doping organisations, such as WADA, employ various methods to identify the presence of steroids or their metabolites in athletes' samples. These methods include urine and blood tests, which can detect the misuse of steroids even if they have been administered in different forms or masked through metabolism.

## IV. GRAMP Model

We propose a GRAph-based modelling for Metabolism Pathway (GRAMP model) that incorporates the domain knowledge for the identification of sample swapping in sports. Our model consists of two main steps: 1) Embedding Steroid Metabolism into Graph Structure and 2) Model Architecture for Graph Classification.

### A. Embedding Steroid Metabolism into Graph Structure:

In this step, we aim to transform the steroid metabolism pathway into a graph structure, as shown in 2. To embed the steroid metabolism pathway into a graph structure, we consider each metabolite (such as testosterone, epitestosterone, androsterone) as individual nodes in the graph. The edges between these nodes represent the connections and interactions between metabolites and reactions. For example, an edge might represent the conversion of testosterone to androsterone catalysed by a specific enzyme. By representing the pathway as a graph, we can capture the spatial relationships and dependencies between metabolites, allowing us to uncover important patterns and interactions within the steroid metabolism process.

*1) Graph Construction:* A graph $G = (V, E)$ with directed edges consists of nodes $V = \{v_1, v_2, \cdots, v_m\}$ and edges $E \subseteq V \times V$, where $e_{i,j} \in E$ represents an edge from node

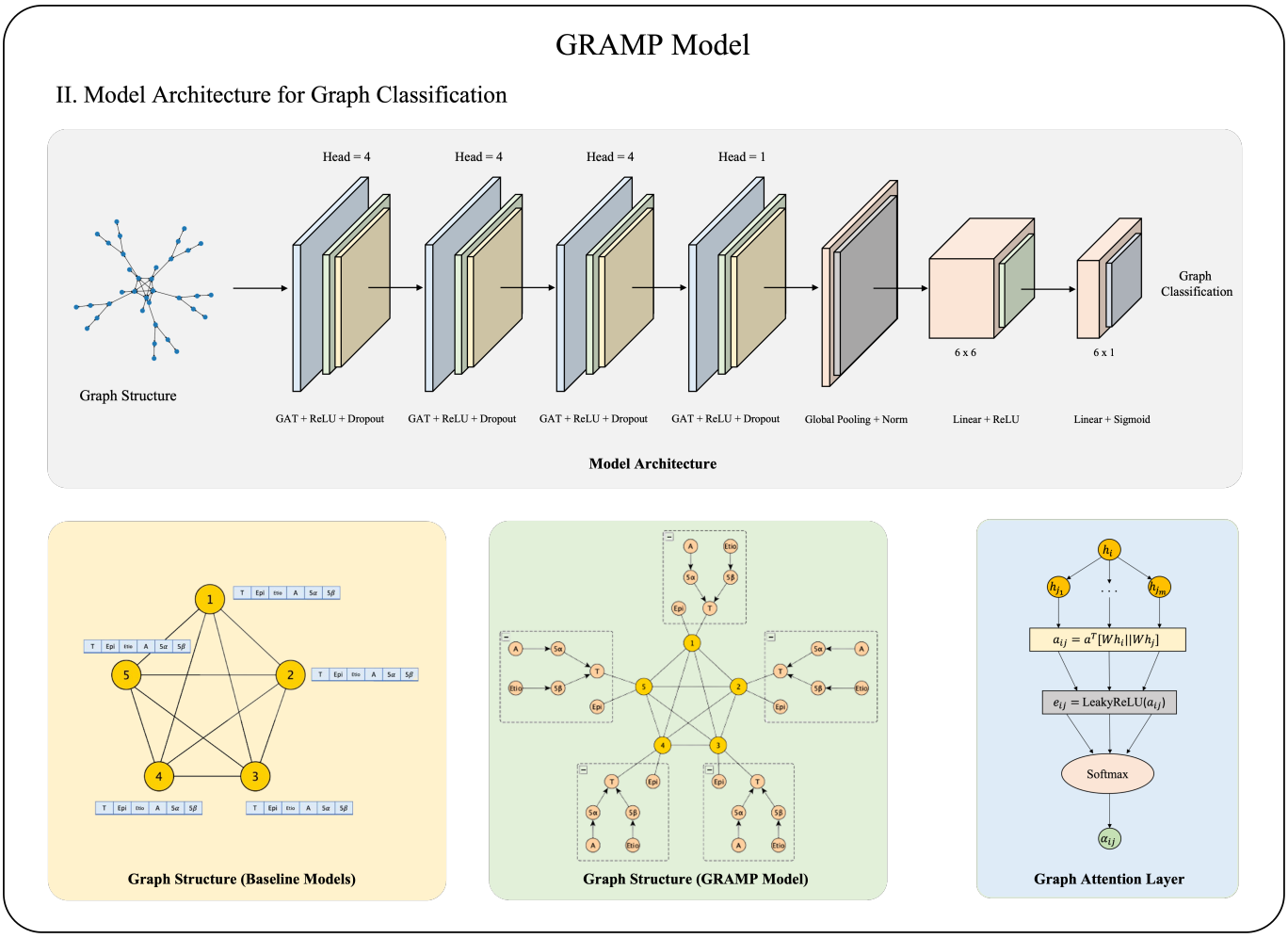


Fig. 3. GRAMP model: Model architecture for the graph classification incorporating GAT layers followed by a ReLU activation function and dropout regularisation. The model concludes with a fully connected layer and a sigmoid activation function for classification into anomalous and normal longitudinal profiles.

$j$ to node $i$. Each node $v_i \in V$ is assumed to have an initial representation $h_i^l \in \mathbb{R}^F$, where $F$ is the number of features in each node representation of $l^{th}$ layer. The neighbours of node $v_i$ are defined as $\mathbb{N}(i) = j \in V | e_{i,j} \in E$.

We construct one graph for each longitudinal profile of the athlete $p$:

$$V^{(p)} = \Big\|_{s=1}^{n} V^{(p,s)} = \{V^{(p,1)} \,\|\, V^{(p,2)} \,\|\, \cdots \,\|\, V^{(p,n)}\} \quad (1)$$

$$E^{(p)} = \Big\|_{i=1}^{n} \Big\|_{j=1}^{n} E_{i,j}^{(p)} \subseteq V^{(p)} \times V^{(p)} \quad (2)$$

where $n$ denotes the total number of samples in the longitudinal profile, and $\|$ represents the concatenation symbol.

Each sample can be represented as:

$$V^{(p,s)} = \{v_0^{(p,s)}, v_1^{(p,s)}, v_2^{(p,s)}, ..., v_k^{(p,s)}\} \in \mathbb{R} \quad (3)$$

$$E_{i,j}^{(p)} \subseteq V^{(p,s)} \times V^{(p,s)} \quad (4)$$

where $v_0^{(p,s)}$ represents the master node for each sample and $v_1^{(p,s)}$ to $v_k^{(p,s)}$ nodes represent each metabolites.

*2) Master Node:* We define a master node for every sample in the longitudinal profile of the athlete. These master nodes are interconnected in a homogeneous graph representation. Considering that all metabolites originate from a common parent compound, we define the master node as the cumulative representation of all metabolites within a given sample.

$$v_0^{(p,s)} = \sum_{i=1}^{k} v_i^{(p,s)} \quad (5)$$

where $k$ represents the total number of steroid parameters, $p$ represents the athlete and $s$ denotes the sample number within the longitudinal profile.

### *B. Model Architecture for Graph Classification:*

Once we have transformed the steroid metabolism pathway into a graph structure, we need a suitable model architecture for graph classification. The goal is to effectively utilise the

learned graph representations to classify whether the graph representing the longitudinal profile of the athlete is normal or anomalous. If there is an anomalous case, it means at least one sample is manipulated and swapped with a clean sample from another individual.

The GCN assigns equal importance to all neighbouring nodes, which may not be suitable for this graph classification task as certain nodes or metabolites could contain more important information than others. Hence, the Graph Attention Network (GAT) model architecture proves to be an optimal choice, which incorporates attention mechanisms to focus on important nodes and edges within the graph during the learning process. It assigns different attention weights to neighbouring nodes based on their relevance to the current node, enabling the model to effectively aggregate and learn from the graph's structural information. By applying the GAT model to our graph representation of the steroid metabolism pathway, we can effectively capture the relevant features and interactions between metabolites.

The GAT model is trained using labelled data, optimising the attention weights and model parameters to achieve high-performance graph classification on the steroid metabolism pathway data. Fig. 3 and Fig. 4 show the detailed model architecture of the GRAMP model, including the graph attention mechanism acting on different nodes of a graph structure.

*1) Graph Attention Layer:* The graph attention layer takes a collection of node features as input, denoted as $h_i = \{h_1, h_2, ..., h_m\}$, where $m$ is the total number of nodes, and $F$ is the number of features associated with each node representation. In our case, since we represent each node with a single metabolism parameter, we have $F = 1$ and $h_i \in \mathbb{R}$. The layer then generates a set of node features, $h'_i = \{h'_1, h'_2, ..., h'_m\}$, where $h'_i \in \mathbb{R}'$.

We perform self-attention on the nodes, i.e., a shared attentional mechanism that computes attention coefficients:

$$a_{ij} = a^T[Wh_i||Wh_j] \tag{6}$$

where $W \in \mathbb{R}^{1\times1}$ is learnable shared weight matrix applied to each node. The attention coefficient, $a^T$ indicates the importance of the node $j$'s value to node $i$. The model allows every node to attend to every other node. We inject the graph structure discussed in the previous section into the mechanism. In the next step, a non-linear activation function is added.

$$e_{ij} = LeakyReLU(a_{ij}) \tag{7}$$

To make coefficients easily comparable across different nodes, we normalise them across all choices of $j$ using the softmax function:

$$\alpha_{ij} = softmax(e_{ij}) = \frac{exp(e_{ij})}{\sum_{q\in N_i} exp(e_{iq})} \tag{8}$$

where $\alpha_{ij}$ represents the pairwise attention coefficients of each metabolites in the metabolism structure.

| Layer | Input | Output | Heads |
|---|---|---|---|
| GATConv1 | 1 x 1 | 1 x 6 | 4 |
| ReLU + Dropout | - | - | - |
| GATConv2 | 4 x 6 | 1 x 6 | 4 |
| ReLU + Dropout | - | - | - |
| GATConv3 | 4 x 6 | 1 x 6 | 4 |
| ReLU + Dropout | - | - | - |
| GATConv4 | 4 x 6 | 1 x 6 | 1 |
| ReLU + Dropout | - | - | - |
| GlobalPooling | - | - | - |
| Linear1 | 1 x 6 | 1 x 6 | - |
| ReLU | - | - | - |
| Linear2 | 1 x 6 | 1 x 1 | - |
| Sigmoid | - | - | - |

Fig. 4. Detailed architecture of each layer in the GRAMP model showing the input and output dimensions.

*2) Loss Function:* For the graph classification task of distinguishing anomalous and normal longitudinal profiles, we employed the binary cross-entropy (BCE) loss function. This loss function is defined as the negative logarithm of a categorical likelihood, which is parameterised by the softmax output. Let $p$ represent the output of softmax layer for a given graph, and $y$ denote the true label of the graph. Then, the BCE loss can be mathematically expressed as follows:

$$\mathcal{L}(y, p) = -\sum_{i=1}^{C} y_i \log p_i \tag{9}$$

where $C \in \{anomalous, normal\}$.

## V. Experiments

### A. Datasets

The dataset represents the longitudinal profile of real-world male and female athletes [25]. It consists of 1432 longitudinal profiles corresponding to 7545 samples where each athlete could have between 3-20 samples in their profile. We randomly partitioned the dataset such that 80% of the data was used for training and 20% for testing the algorithm. Table II shows the summary of the number of samples belonging to male and female athletes. Each sample consists of a set of biomarkers called steroid metabolism parameters that show significant changes in the administration of steroids, as listed in Table I. Fig. 5 shows the data distribution of the number of samples in the longitudinal profile per athlete for male and female athletes. We observe that the majority of the longitudinal profiles are from young or new athletes, i.e., with only 3-4 samples.

### B. Baseline Models

We selected a set of baseline models that serve as a performance benchmark for comparing our proposed GRAMP model. These baselines consist of both non-graph and graph-based models that do not incorporate domain knowledge

TABLE II
THE TABLE PROVIDES A COMPREHENSIVE OVERVIEW OF THE DATA STATISTICS USED FOR TRAINING AND TESTING THE PROPOSED MODEL.

| | **Male** | | **Female** | |
|---|---|---|---|---|
| | Profile | Sample | Profile | Sample |
| **Training** | 846 | 4349 | 301 | 1594 |
| **Testing** | 211 | 1121 | 74 | 481 |
| **Total** | 1057 | 5470 | 375 | 2075 |

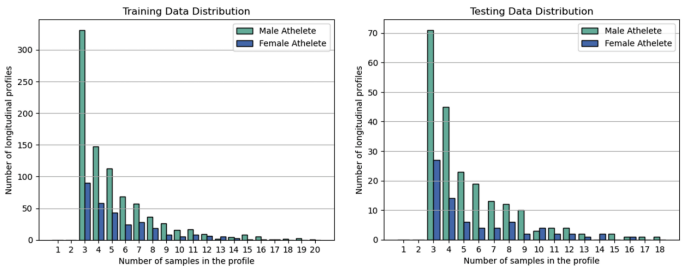


Fig. 5. Distribution of Samples in longitudinal profile per athlete for male and female athletes in training and testing datasets.

into the model training. This performance comparison will help us to explore the potential impact of leveraging the steroid metabolism pathway into the decision making using the GRAMP model. These baseline models were trained and optimised using the training dataset.

- *Bayesian Method (SoTA)* [32]: use to determine the personalised threshold for each steroid parameter which is used to compare the new samples. These thresholds are calculated from the prior distribution based on the reference population.
- *Random Forest (RF)* [23]: uses multiple decision trees and combines their output for classification problems, achieving high accuracy and interpretability.
- *eXtreme Gradient Boosting - XGBoost (XGB)* [5]: uses an optimized distributed gradient boosting algorithm to achieve high performance on structured data.
- *Graph Convolutional Network (GCN)* [14]: can learn representations of nodes in a graph, where each node represents a sample in the longitudinal profile.
- *Graph Isomorphism Network (GIN)* [38]: can learn node embeddings by aggregating local and global substructure information of graphs, where each node represents a sample.
- *Graph Attention Network (GAT)* [35]: uses attention mechanisms to learn node embeddings in graphs, achieving state-of-the-art performance in a variety of graph-based tasks, where each node represents a sample.

Table III shows the values of the different hyperparameters selected to train all the baselines and GRAMP model. These values are selected after performing the optimisation step.

TABLE III
THE TABLE PRESENTS THE HYPERPARAMETER VALUES OF ALL THE BASELINES AND THE PROPOSED GRAMP MODEL AFTER PERFORMING MODEL OPTIMISATION.

| **M**odel | Parameter value |
|---|---|
| Random Forest (RF) | *n estimators* = 100 |
| | *criterion* = gini |
| | *min samples split* = 2 |
| | *bootstrap* = True |
| XGBoost (XGB) | *objective* = binary logistic |
| | *learning rate* = 0.1 |
| | *max depth* = 7 |
| | *n estimators* = 200 |
| Graph Convolutional Network (GCN) | *n GNN layers* = 1 (1 hop) |
| | *n linear layers* = 2 |
| | *n LayerNorm* = 1 |
| | *n Epochs per fold* = 25 |
| | *Dropout* = 0.25 |
| Graph Isomorphism Network (GIN) | *n GNN layers* = 1 (1 hop) |
| | *n linear layers* = 2 |
| | *n LayerNorm* = 1 |
| | *n Epochs per fold* = 25 |
| | *Dropout* = 0.25 |
| Graph Attention Network (GAT) | *n GNN layers* = 1 (1 hop) |
| | *n linear layers* = 2 |
| | *n LayerNorm* = 1 |
| | *n Epochs per fold* = 25 |
| | *Dropout* = 0.25 |
| Our Method (GRAMP) | *n GAT layers* = 4 (4 hops) |
| | *n linear layers* = 2 |
| | *n LayerNorm* = 1 |
| | *n Epochs per fold* = 75 |
| | *Dropout* = 0.25 |
| | *optimiser* = Adam |

### C. Implementation

Given that we have framed our fraud detection problem as a supervised graph classification task, it is essential to have a labeled dataset that includes samples for each class, specifically normal and anomalous profiles. We performed a random selection on the dataset, choosing 50% of the profiles. In each selected profile, we manually replaced one sample with a sample from a different profile. These modified profiles were labelled as anomalous profiles (labelled as '1'). The remaining 50% of the profiles were considered normal profiles (labelled as '0'). To ensure consistency, we normalised each profile to the unit norm separately.

All the models are implemented based on the SCIKIT-LEARN [20], XGBOOST [5], and PYTORCH-GEOMETRIC [8] packages. One significant challenge during model training was

overfitting, which limits the model's generalisation capability. Since we have a small training dataset, addressing overfitting became a critical concern in our analysis. Therefore, we performed the $k$-fold cross-validation method [26] to train our models, with $k$ set to 5. Each fold was used as a validation set, while the remaining folds were collectively employed as the training dataset, and the overall performance was determined by computing the mean performance across all the folded models.

Each model comprises a set of hyperparameters that can be adjusted to improve the training process. Consequently, conducting a coarse grid search is necessary to determine the optimal combination of these hyperparameters. We used a hyperparameter optimisation framework to efficiently explore a substantial grid space while promptly eliminating unpromising trials and implemented this framework using OPTUNA package [1]. The optimised trained model is deployed on the testing set, enabling predictions for previously unseen profiles. Finally, the model's performance was evaluated by calculating various evaluation metrics, facilitating a comprehensive assessment of its efficacy.

### D. Performance Metrics

The performance evaluation of each model was conducted using accuracy, sensitivity, specificity, and area under the ROC curve (AUC). Sensitivity indicates the percentage of accurately identified anomalous longitudinal profiles, while specificity measures the percentage of accurately identified normal longitudinal profiles.

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

where $TP$ and $TN$ denote the number of longitudinal profiles classified correctly by the model as anomalous and normal, respectively, while $FN$ and $FP$ denote the number of misclassified anomalous and normal longitudinal profiles, respectively.

## VI. Results

We compared the performance of our proposed GRAMP model with all the baseline models for detecting fraudulent behaviour, i.e., sample swapping on both male and female datasets, as presented in Table IV and Table V respectively. The uncertainties are calculated using a 5-fold cross-validation approach. Among the baselines, SoTA and XGB demonstrated better performance, highlighting the importance of bayesian and boosting models for fraud detection. Despite an accuracy of over 60%, GIN was not able to successfully detect any anomalous profiles (sensitivity below 40%). In case of female athletes, a similar trend can also be seen where SoTA and XGB showed better performance among baselines. Graph models (GCN, GIN, GAT) show high specificity values but less accuracy compared to other baselines. This shows that the homogenous graph structure, where each node representing the sample is unable to leverage the metabolism pathways well. Our proposed GRAMP model outperformed all baselines, showing that adding domain knowledge by defining a graph structure based on the metabolism pathway is effective. We achieved sensitivity values over 80% and AUC values over 90% on both male and female athletes.

The ROC and PRC curves for all models evaluated on male and female datasets are presented in Figure 6 and Figure 7, respectively. As depicted, the proposed model outperforms all the baseline models in both curves. The results for graph models (GCN, GIN and GAT) are quite similar and better than non-graph model RF, possibly because the fraud activity in longitudinal profiles is too complex for a simple classification model to handle. Of all the baselines, XGB is the most competitive, likely because it generates a representation of parameters through a boosting algorithm.

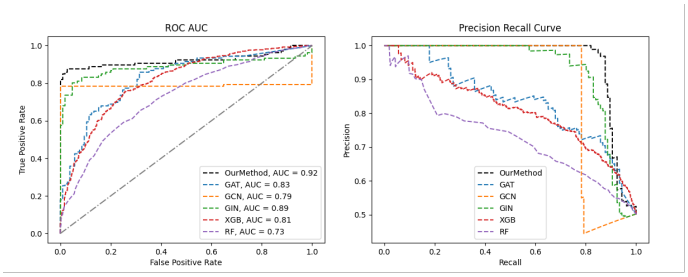


Fig. 6. Receiver Operating Characteristic (ROC) and Precision-Recall Curve (PRC) plots show the performance of our proposed model with various baseline models for male athletes. The ROC curve illustrates the trade-off between the true positive rate and the false positive rate, while the PRC curve showcases the precision-recall trade-off.

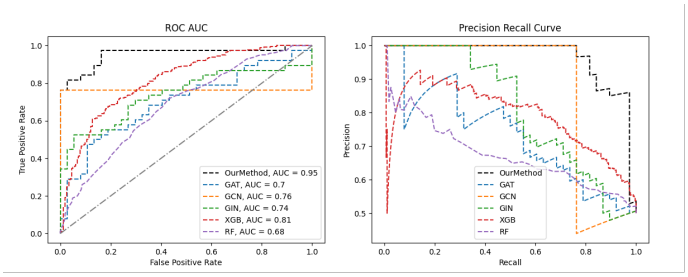


Fig. 7. Receiver Operating Characteristic (ROC) and Precision-Recall Curve (PRC) plots show the performance of our proposed model with various baseline models for female athletes.

We randomly took a longitudinal profile of both male and female athletes from the testing dataset and computed the pairwise attention coefficients for one of the samples. Fig. 8 shows the weighted contribution of the neighbourhood to each node. We can observe that testosterone and epitestosterone have the highest attention coefficient to the master node compared to the other metabolism parameters, i.e., the model gives more importance to the message passing between the master node and testosterone and between the master node and epitestosterone.

Since the data for male athletes are more sparse than for female athletes in general, i.e., there is a high variation in the concentration values of the metabolism parameters in the male body than in females, we observe high attention

TABLE IV

COMPREHENSIVE PERFORMANCE COMPARISON OF THE PROPOSED MODEL ON MALE ATHLETES. THE MEAN AND STANDARD DEVIATION VALUES, OBTAINED THROUGH CROSS-VALIDATION ON THE TRAINING SET, ARE REPORTED FOR ALL THE EVALUATION METRICS.

| **Metrics** | **SoTA** | | **RF** | | **XGB** | | **GCN** | | **GIN** | | **GAT** | | **GRAMP** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Train | Test | Train | Test | Train | Test | Train | Test | Train | Test | Train | Test | Train | Test |
| ACC | - | 0.76 | 0.65±0.01 | 0.66 | 0.73±0.01 | 0.74 | 0.68±0.02 | 0.69 | 0.62±0.04 | 0.67 | 0.66±0.08 | 0.72 | 0.89±0.04 | 0.91 |
| SN | - | 0.73 | 0.65±0.02 | 0.68 | 0.76±0.02 | 0.77 | 0.36±0.04 | 0.38 | 0.21±0.08 | 0.35 | 0.35±0.20 | 0.55 | 0.86±0.02 | 0.86 |
| SP | - | 0.82 | 0.64±0.02 | 0.65 | 0.71±0.02 | 0.70 | 0.99±0.01 | 1.00 | 1.00±0.00 | 1.00 | 0.96±0.04 | 0.90 | 0.93±0.07 | 0.97 |
| AUC | - | - | 0.65±0.01 | 0.73 | 0.73±0.01 | 0.81 | 0.67±0.05 | 0.79 | 0.83±0.04 | 0.89 | 0.75±0.08 | 0.83 | 0.91±0.04 | 0.92 |

TABLE V

COMPREHENSIVE PERFORMANCE COMPARISON OF THE PROPOSED MODEL ON FEMALE ATHLETES. THE MEAN AND STANDARD DEVIATION VALUES, OBTAINED THROUGH CROSS-VALIDATION ON THE TRAINING SET, ARE REPORTED FOR ALL THE EVALUATION METRICS.

| **Metrics** | **SoTA** | | **RF** | | **XGB** | | **GCN** | | **GIN** | | **GAT** | | **GRAMP** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Train | Test | Train | Test | Train | Test | Train | Test | Train | Test | Train | Test | Train | Test |
| ACC | - | 0.71 | 0.64±0.01 | 0.63 | 0.76±0.03 | 0.73 | 0.68±0.05 | 0.68 | 0.56±0.05 | 0.60 | 0.52±0.03 | 0.53 | 0.70±0.06 | 0.88 |
| SN | - | 0.38 | 0.66±0.03 | 0.67 | 0.79±0.03 | 0.78 | 0.35±0.08 | 0.37 | 0.08±0.06 | 0.21 | 0.08±0.04 | 0.11 | 0.60±0.17 | 0.82 |
| SP | - | 0.85 | 0.62±0.02 | 0.60 | 0.72±0.04 | 0.67 | 1.00±0.00 | 1.00 | 1.00±0.00 | 1.00 | 1.00±0.00 | 0.97 | 0.82±0.14 | 0.95 |
| AUC | - | - | 0.64±0.01 | 0.68 | 0.76±0.03 | 0.81 | 0.76±0.06 | 0.76 | 0.72±0.07 | 0.74 | 0.61±0.06 | 0.70 | 0.82±0.11 | 0.95 |

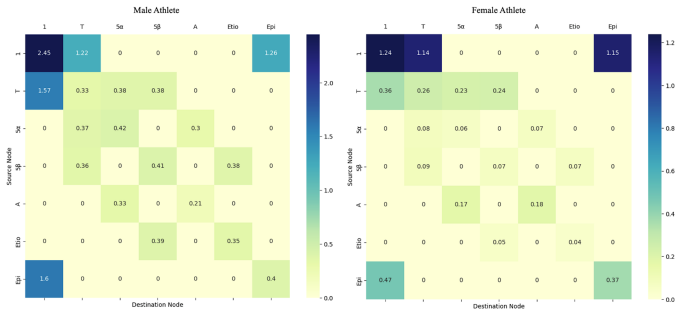


Fig. 8. Pairwise attention coefficients for the randomly selected sample from the longitudinal profile of male and female athletes from the testing dataset. The attention coefficients highlight the significance and relevance of information propagation between different metabolites within the GRAMP model, shedding light on the specific interactions and dependencies of the metabolism pathway.

coefficient values for the male athlete. In addition, we have two cases of information propagation. First, when testosterone is the source, and the master node is the destination, i.e., the message is passing from testosterone to the master node, and second, when the master node is the source and testosterone is the destination. For the male athletes, we observe relatively similar attention coefficient values in both cases suggesting that the bidirectional message passing is significant, whereas, in the case of females, we observe relatively high attention coefficients for the latter case. Similar behaviour can also be observed with epitestosterone.

Overall, the proposed GRAMP model consistently outperforms other state-of-the-art baseline models due to two factors. Firstly, our model effectively captures the spatial behaviour of longitudinal profiles through graph representation learning. Unlike other graph models such as GCN, GIN, and GAT, which treat longitudinal profiles as homogeneous graph structures, our model explicitly considers their spatial characteristics. Secondly, our model incorporates an attention mechanism that generates high-level embeddings, facilitating enhanced pattern learning. Consequently, our model outperforms other baselines, particularly at the initial stages of the curve, and exhibits remarkable accuracy in detecting fraudulent longitudinal profiles with high specificity, showcasing its promising capabilities.

### A. Ablation Studies

We performed ablation studies to study the effect of different components in the GRAMP model, like the selection of the master node and the number of graph layers/hops. First, we tried different variations in the master node by selecting different functions, i.e., $SUM$(nodes) (sum of values of all the nodes), $T/E$ (ratio between T and E), and $Avg$(nodes) (mean of values of all the nodes). Fig. 9 shows the performance of the model in all three variations on male and female athletes. Since the master node represents the entire sample, it should contain information about all the metabolism parameters. Therefore, we observe that $T/E$ shows the least performance for male athletes because it only contains information about testosterone and epitestosterone. On the other hand, $Avg$(nodes) shows the least performance for female athletes because the concentration values of all the metabolism parameters have different scales, especially for female athletes because of data sparsity, so averaging all the values would not be a feasible solution. Therefore, selecting the sum of all the values of all the metabolism parameters outperforms the other two variations for both male and female athletes.

Next, to understand the importance of the attention layer, we varied the number of GAT layers in the model network.

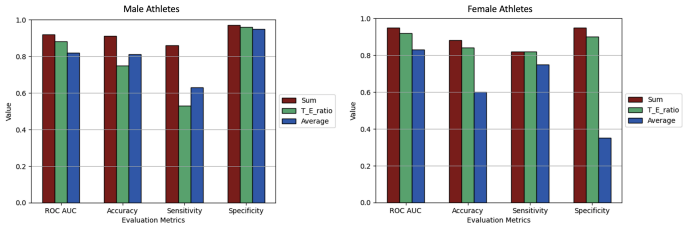


Fig. 9. Performance of the GRAMP model with variation in the master node for male and female athletes. This analysis provides insights into how different configurations of the master node influence the detection and classification of sample swapping, allowing for a comprehensive evaluation of the model's performance across testing dataset.

Fig. 10 shows the performance of the model for both male and female athletes. We observe that the performance start increasing as we keep adding the number of GAT layers until a point after which it starts decreasing. We found 4 to be the optimal number of layers for this problem, showing we need at least four hops for complete message passing in the graph network.

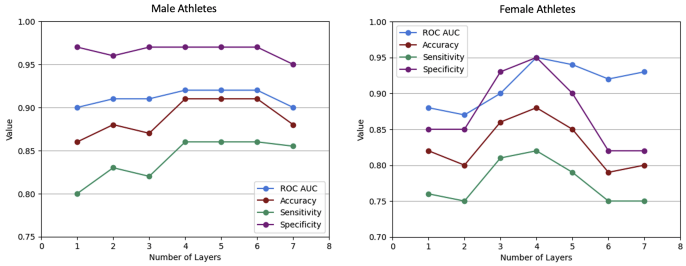


Fig. 10. Performance of the GRAMP method with respect to the number of graph attention layers for male and female athletes. By varying the number of GAT layers in the model, we analyse the impact on the overall performance and effectiveness of the proposed model and find the optimal number of layers required for this problem.

## VII. Conclusion

The objective of this analysis is to address research questions regarding the potential impact of leveraging domain knowledge in decision-making by machine learning models and employing graph representation learning to model the steroid metabolism pathway. Specifically, our investigation focuses on assessing the benefits of integrating information about metabolism pathways for enhancing anti-doping analysis and improving the detection of sample swapping in sports.

In this paper, we propose the GRAMP model, which can be helpful for the detection of fraudulent behaviour of athletes, specifically sample swapping in sports using longitudinal profiles. Our proposed model considers the spatial behaviour of the longitudinal profile and generates embedding maps using a graph attention mechanism to capture the implicit relationship among all the steroid parameters of the sample. Moreover, by using single-node representation for each metabolite, we showed how adding domain knowledge would be helpful in improving decision making. This is the first work in which a graph attention network has ever been employed to address the fraudulent behaviour detection problem in sports. The results indicate that our model outperforms other state-of-the-art methods in terms of sensitivity and specificity.

Currently, WADA follows a standardised protocol to detect sample swapping cases [36]. Firstly, all the athlete's profiles are analysed using the Adaptive model based on the Bayesian method (SoTA). The athlete's profiles flagged by the Adaptive model then undergo laboratory testing, like DNA analysis. The proposed graph-based method for incorporating domain knowledge of the metabolism pathways can help the decision makers to flag sample swapping cases 15% more accurately than the Adaptive model. The graph-based method can help the WADA experts to make decisions 17-25% more accurately than the non-graph based model. This reduces the risk of unnecessary laboratory testing by 10-15% (specificity), which provides costs and time benefits. This demonstrate that our model can effectively detect anomalous longitudinal profiles and can help anti-doping authorities trigger fraudulent practices during sports events and make the sports clean.

## References


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

[2] Bond, P., Smit, D.L., and de Ronde, W. 2022. Anabolic–androgenic steroids: How do they work and what are the risks?. Frontiers in Endocrinology, 13.

[3] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko. O. (2013). Translating embeddings for modeling multi-relational data. In Proceedings of NeurIPS 2013.

[4] Callaway, E. 2011. Sports doping: Racing just to keep up. Nature 475, 7356 (2011), 283–285.

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

[6] Dettmers, T., Minervini, P., Stenetorp, P., and Riedel. S. (2018). Convolutional 2d knowledge graph embeddings. In Proceedings of AAAI 2018.

[7] Dong, Y., Chawla, N., and Swami, A. (2017). metapath2vec: Scalable representation learning for heterogeneous networks. In Proceedings of KDD 2017.

[8] Fey, M., and Lenssen, J. E. (2019). PyTorch Geometric: An efficient graph library for deep learning in PyTorch. In International Conference on Learning Representations (ICLR) Workshop.

[9] Grover, A., and Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In Proceedings of KDD 2016.

[10] Hamilton, W.L., Ying, R., and Leskovec, J. 2017. Inductive representation learning on large graphs. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17).

[11] Hetzel, L., Fischer, D. S., Günnemann, S., Theis, F. J. 2021. Graph representation learning for single-cell biology. Current Opinion in Systems Biology, 28, 100347.

[12] Hou, J., Acharya, L., Zhu, D., Cheng, J. 2016. An overview of bioinformatics methods for modeling biological pathways in yeast. Briefings in functional genomics, 15:2, 95-108.

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

[14] Kipf, T. N., Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In International Conference on Learning Representations (ICLR).

[15] Le Novere, N. 2015. Quantitative and logic modelling of molecular and gene networks. Nature Reviews Genetics, 16:3, 146-158.

[16] Marques, M.A.S., Damasceno, L.M.P., Pereira, H.M.G., Caldeira, C.M., Dias, B.F.P., Vargens, D.G., Amoedo, N.D., Volkweis, R.O., Viana, R.O.V., Rumjanek, F.D., Aquino Neto, F.R. 2005. DNA typing: an accessory evidence in doping control. Journal of Forensic Science 50, 3 (May 2005), 587–592.

[17] McLaren, R.H. 2016. WADA investigation of Sochi allegations: The independent person 2nd report. World Anti-Doping Agency (Dec 2016).

[18] Ng Tang Fui, M., Prendergast, L. A., Dupuis, P., Raval, M., Strauss, B. J., Zajac, J. D., and Grossmann, M. 2016. Effects of testosterone treatment on body fat and lean mass in obese men on a hypocaloric diet: a randomised controlled trial. BMC medicine, 14, 153.

[19] O'Hara, L., Livigni, A., Theo, T., Boyer, B., Angus, T., Wright, D., ... Freeman, T. C. 2016. Modelling the structure and dynamics of biological pathways. PLoS biology, 14:8.

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

[21] Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). Deepwalk: Online learning of social representations. In Proceedings of KDD 2014.

[22] Piper, T., Geyer, H., Haenelt, N., Hülsemann, F., Schaenzer, W., and Thevis, M. (2021). Current Insights into the Steroidal Module of the Athlete Biological Passport. International Journal of Sports Medicine 42.

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

[24] Rahman, M.R., Bejder, J., Bonne, T.C., Andersen, A.B., Huertas, J.R., Aikin, R., Nordsborg, N.B., and Maass. W. (2022). Detection of Erythropoietin in Blood to Uncover Doping in Sports using Machine Learning. In Proceedings of IEEE International Conference on Digital Health (ICDH). 193–201.

[25] Rahman, M.R., Piper, T., Geyer, H., Equey, T., Baume, N., Aikin, R., and Maass, W. (2022). Data Analytics for Uncovering Fraudulent Behaviour in Elite Sports. In Proceedings of International Conference on Information System (ICIS).

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

[27] Renterghem, P.V., Sottas, P.-E., Saugy, M., and Eenoo, P.V. (2013). Statistical discrimination of steroid profiles in doping control with support vector machines. Analytica Chimica Acta 768, 41–48.

[28] Robinson, N., Giraud, S., Saudan, C., Baume, N., Avois, L., Mangin, P., and Saugy, M. 2006. Erythropoietin and blood doping. British journal of sports medicine. 40, 30-34.

[29] Schiffer, L., Barnard, L., Baranowski, E. S., Gilligan, L. C., Taylor, A. E., Arlt, W., Shackleton, C. H. L., Storbeck, K. H. (2019). Human steroid biosynthesis, metabolism and excretion are differentially reflected by serum and urine steroid metabolomes: A comprehensive review. The Journal of steroid biochemistry and molecular biology, 194.

[30] Simao, E., Remy, E., Thieffry, D., Chaouiya, C. 2005. Qualitative modelling of regulated metabolic pathways: application to the tryptophan biosynthesis in E. coli. Bioinformatics, 21.

[31] Socher, R., Chen, D., Manning, C., and Ng, A. (2013). Reasoning with neural tensor networks for knowledge base completion. In Proceedings of NeurIPS 2013.

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

[33] Thevis, M., Geyer, H., Sigmund, G., and Schänzer, W. (2012). Sports drug testing: Analytical aspects of selected cases of suspected, purported, and proven urine manipulation. Journal of Pharmaceutical and Biomedical Analysis 57, 26–32.

[34] Torres, L., Blevins, A. S., Bassett, D., Eliassi-Rad, T. 2021. The why, how, and when of representations for complex systems. SIAM Review, 63:3, 435-485.

[35] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y. (2018). Graph attention networks. In International Conference on Learning Representations (ICLR).

[36] WADA. (2021). Athlete Passport Management Unit Requirements and Procedures. WADA Technical Document – TD2021APMU.

[37] Wilkes, E.H., Rumsby, G., and Woodward, G.M. (2018). Using Machine Learning to Aid the Interpretation of Urine Steroid Profiles. Clinical Chemistry 64, 11, 1586–1595.

[38] Xu, K., Hu, W., Leskovec, J., Jegelka, S. (2019). Learning graph isomorphism with graph convolutional networks. In International Conference on Machine Learning (ICML).

[39] Zhang, C., Song, D., Huang, C., Swami, A., and Chawla, N. (2019). Heterogeneous graph neural network. In Proceedings of KDD 2019.

[40] Zhang, Z., Cui, P., and Zhu, W. (2020). Deep learning on graphs: A survey. IEEE Transactions on Knowledge and Data Engineering, 2020.