# Anomaly Detection in Longitudinal Clinical Profile

**Maxx Richard Rahman**

A dissertation submitted towards the degree
*Doctor of Engineering (Dr.-Ing.)*
of the Faculty of Mathematics and Computer Science
of Saarland University

Saarbrücken, 2025

# PhD Defense Details

**Dean of the Faculty MI:** Prof. Dr. Roland Speicher

**Defense Date:** 18th Dec 2025

**Examination Committee Members**

- **Chair:** Prof. Dr. Antonio Krüger
- **Reviewers:**
  Prof. Dr.-Ing. Wolfgang Maaß
  Prof. Dr. Hans-Peter Lenhof
- **Academic Assistant:** Dr.-Ing. Sabine Janzen

# Declaration

I hereby declare that this dissertation is the result of my own independent work and investigation, except where otherwise stated in the case of collaborations that resulted in joint publications. All sources used have been appropriately acknowledged and referenced. This work has not been submitted, either in whole or in part, for any other degree or qualification at this or any other university. As a non-native English speaker, I have used AI-based services (Grammarly, LLM, etc.) for grammar correction and language editing during the preparation of this thesis. These tools were used solely to improve the clarity and readability of the text and did not contribute to the generation of original scientific content, research ideas, model development, or interpretation of results. I affirm that the intellectual contribution and substantive content presented in this dissertation are entirely my own.

Maxx Richard Rahman
Saarbrücken, 2025


**Eidesstattliche Versicherung**

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form in einem Verfahren zur Erlangung eines akademischen Grades vorgelegt.

**Declaration of original authorship**

I hereby declare that this dissertation is my own original work except where otherwise indicated. All data or concepts drawn directly or indirectly from other sources have been correctly acknowledged. This dissertation has not been submitted in its present or similar form to any other academic institution either in Germany or abroad for the award of any other degree.

Saarbrücken, 18.09.2025

# Acknowledgements

I would like to express my sincere gratitude to my advisor, Prof. Dr.-Ing. Wolfgang Maaß, for his guidance and invaluable advice throughout this work. I thank the members of the examination committee and the Saarbrücken Graduate School of Computer Science for their support and consideration. I am grateful to my colleagues and the members of our research group for the many stimulating discussions that helped refine my ideas.

My gratitude extends to my co-authors, Dr. Reid Aikin, Dr. Norbert Baume, Dr. Tristan Equey, Dr. Hans Geyer, and Dr. Thomas Piper, for their valuable collaboration and contributions to the research projects that shaped this dissertation. I also acknowledge the German Research Center for Artificial Intelligence (DFKI) and the World Anti-Doping Agency (WADA) for their financial and institutional support.

Finally, I wish to acknowledge the constant encouragement of my parents, as well as the support of my family and friends, without which this dissertation would not have been possible.

# Zusammenfassung

Diese Arbeit untersucht die Erkennung von Anomalien in longitudinalen klinischen Daten, mit besonderem Schwerpunkt auf Anti-Doping-Anwendungen, bei denen die Überwachung von Sportlern die Identifizierung subtiler, zeitlich eingebetteter Abweichungen in biologischen Profilen erfordert. Im Gegensatz zu Einzelprobenbewertungen ermöglichen Längsschnittdaten die Analyse der intraindividuellen Dynamik im Zeitverlauf und unterstützen so die Erkennung abnormaler Muster, die sonst möglicherweise verborgen bleiben würden. Eine große Herausforderung im Anti-Doping-Bereich ist der Einsatz ausgeklügelter Manipulationsstrategien durch einige Sportler, um positive Dopingtests zu umgehen. Ein Beispiel hierfür ist der Probenaustausch, bei dem die biologischen Proben von Athleten absichtlich durch die Proben einer anderen Person oder durch zuvor gelagerte „saubere" Proben ersetzt werden. Solche Praktiken untergraben die Zuverlässigkeit herkömmlicher Testmethoden, bei denen in der Regel davon ausgegangen wird, dass jede Probe authentisch und unverfälscht ist. Im Gegensatz dazu ermöglicht die longitudinale Anomalieerkennung die Identifizierung von Unstimmigkeiten innerhalb der biologischen Entwicklung eines Athleten und bietet somit eine Möglichkeit, Unregelmäßigkeiten aufzudecken, die auf einen möglichen Probenaustausch hindeuten. Die Erkennung solcher Anomalien ist aufgrund verschiedener Herausforderungen im Zusammenhang mit Längsschnittdaten oder dem Bereich der Dopingbekämpfung selbst schwierig. Um diesen Herausforderungen zu begegnen, ist diese Arbeit in drei Teile gegliedert.

Der erste Teil der Arbeit stellt Methoden zur Längsschnitt-Anomalieerkennung vor, die sich mit den zentralen Herausforderungen unregelmäßiger Probenahmeintervalle, heterogener Profillängen, begrenzter Probenanzahlen pro Athlet und der Knappheit von Ground-Truth-Labels befassen. Es werden zwei sich ergänzende Architekturen vorgeschlagen. Das Self Attention-based Convolutional Neural Network (SACNN) geht diese Probleme an, indem es aus unregelmäßigen Profilen strukturierte Teilsequenzen konstruiert und aufmerksamkeitsgewichtete Faltungsschichten anwendet, um strukturelle und zeitliche Abhängigkeiten zu lernen und so subtile kontextuelle Anomalien wie Probenvertauschungen zu erfassen. Parallel dazu bewältigt das Subsampling-based Convolutional Neural Network (SCNN) die Herausforderungen durch eine Subsampling- und Aggregationsstrategie, bei der tripletbasierte

Segmente verwendet werden, um unterschiedliche Konsistenzen zu erfassen, was eine zuverlässige Anomalieerkennung selbst bei Profilen mit nur zwei Proben ermöglicht. Beide Modelle reduzieren die Abhängigkeit von expliziten Anomalie-Labels, indem sie individualisierte Baselines lernen. Sie werden unter hochspezifischen Einschränkungen trainiert, wobei die Bewertungen anhand realer Anti-Doping-Datensätze einschließlich DNA-verifizierter Anomalien durchgeführt werden.

Der zweite Teil der Arbeit bezieht Stoffwechselwegstrukturen in das Modelllernen ein und stellt so sicher, dass die Modellausgaben nicht nur genau, sondern auch biologisch plausibel sind. Es werden zwei sich ergänzende Ansätze vorgeschlagen. Structural-Temporal Tokenization for Large Language Models (STT-LLM) führt eine neuartige Tokenisierungsstrategie ein, die das metabolische strukturelle und zeitliche Verhalten klinischer Parameter aus Längsschnittprofilen codiert, sodass ressourceneffiziente Sprachmodelle klinische Daten unter Beibehaltung des biologischen Kontexts verarbeiten können. Parallel dazu bettet Graph-based Modelling for Metabolism Pathways (GRAMP) das Steroid-Stoffwechselnetzwerk in eine Graph-Attention-Architektur ein, wodurch das Modell durch Informationsverbreitung über metabolisch verknüpfte Biomarker wegkonsistente Anomalien erkennen kann.

Der dritte Teil der Arbeit konzentriert sich auf interpretierbare und domäneninformierte Argumentation zur Entscheidungsunterstützung. Es werden zwei sich ergänzende Erklärungsinstrumente vorgeschlagen. Metabolism Pathway-driven Prompting (MPP) verwendet strukturierte Graphen des Steroidstoffwechselwegs, um Sprachmodelle bei der Generierung von textuellen Erklärungen für markierte Anomalien anzuleiten und erkannte Abweichungen mit plausiblen biologischen Mechanismen zu verknüpfen. Digital Athlete Passport (DAP) bietet einen visuellen Analyseansatz, bei dem hochdimensionale longitudinale klinische Profile in niedrigdimensionale Räume projiziert werden, um Abweichungen und Trajektorienverschiebungen zu visualisieren, unterstützt durch PCA-basierte Interpretation und Zentroid-Tracking. Alle Modelle sind in CASPIAN integriert, einem Software-Framework, das es Fachleuten ermöglicht, Erkennungs-, strukturbewusste Modellierungs- und Interpretierbarkeitsmethoden flexibel zu kombinieren. Zusammen bieten diese Beiträge einen umfassenden Ansatz zur Anomalieerkennung in longitudinalen klinischen Profilen und ermöglichen eine biologisch fundierte und erklärbare Überwachung in Bereichen mit hohem Risiko, wie z. B. Anti-Doping und darüber hinaus.

# Abstract

This thesis investigates anomaly detection in longitudinal clinical data, with a particular focus on anti-doping applications where athlete monitoring requires identifying subtle, temporally embedded deviations in biological profiles. Unlike single-sample assessments, longitudinal data allows the analysis of intra-individual dynamics over time, supporting the detection of abnormal patterns that may otherwise remain hidden. A major challenge in anti-doping is the use of sophisticated manipulation strategies by some athletes to evade positive doping tests. An example is sample swapping, in which athletes' biological samples may be deliberately substituted with those of another individual or with previously stored "clean" samples. Such practices undermine the reliability of conventional testing methods, which typically assume each sample to be authentic and unaltered. In contrast, longitudinal anomaly detection allows for the identification of inconsistencies within an athlete's biological trajectory, thereby offering a means of uncovering irregularities suggestive of potential swapping events. Detecting such anomalies is difficult due to different challenges related to longitudinal data or the domain of anti-doping itself. To address these challenges, this thesis is categorized into three parts.

The first part of the thesis introduces methods for longitudinal anomaly detection that address the key challenges of irregular sampling intervals, heterogeneous profile lengths, limited numbers of samples per athlete, and the scarcity of ground-truth labels. Two complementary architectures are proposed. The Self Attention-based Convolutional Neural Network (SACNN) addresses these issues by constructing structured subsequences from irregular profiles and applying attention-weighted convolutional layers to learn structural-temporal dependencies, thereby capturing subtle contextual anomalies such as sample swapping. In parallel, the Subsampling-based Convolutional Neural Network (SCNN) handles the challenges through a subsampling and aggregation strategy, where triplet-based segments are used to capture differential consistency, allowing reliable anomaly detection even in profiles with as few as two samples. Both models reduce reliance on explicit anomaly labels by learning individualized baselines. They are trained under high-specificity constraints, with evaluations performed on real-world anti-doping datasets including DNA-verified anomalies.

The second part of the thesis incorporates metabolic pathway structures into model learning, ensuring that outputs are not only accurate but also biologically plausible. Two complementary approaches are proposed. Structural–Temporal Tokenization for Large Language Models (STT-LLM) introduces a novel tokenization strategy that encodes both the metabolic structure and temporal behaviour of clinical parameters from longitudinal profiles, enabling resource-efficient language models to process clinical data while preserving biological context. In parallel, Graph-based Modelling for Metabolism Pathways (GRAMP) embeds the steroid metabolic network into a graph attention architecture, allowing the model to detect pathway-consistent anomalies through information propagation across metabolically linked biomarkers.

The third part of the thesis focuses on interpretable and domain-informed reasoning for decision support. Two complementary explanation tools are proposed. Metabolism Pathway-driven Prompting (MPP) uses structured graphs of the steroid metabolism pathway to guide language models in generating textual explanations for flagged anomalies, linking detected deviations to plausible biological mechanisms. Digital Athlete Passport (DAP) offers a visual analytics approach, projecting high-dimensional longitudinal clinical profiles into lower-dimensional spaces to visualize deviations and trajectory shifts, supported by PCA-based interpretation and centroid tracking. All models are integrated into CASPIAN, a software framework that allows domain experts to flexibly combine detection, structure-aware modelling, and interpretability methods. Together, these contributions provide a comprehensive approach to anomaly detection in longitudinal clinical profiles, allowing biologically grounded and explainable monitoring in high-stakes domains such as anti-doping and beyond.

# Table of contents

# List of publications

## 2025

- **[NeurIPS 25]** Rahman, M.R., Hammouda, M., Maass, W. (2025). Structural-Temporal Tokenization for Resource-Efficient Language Models on Clinical Time-Series. *In Proceedings of the Neural Information Processing Systems (NeurIPS 2025), Main Track*. (under review)

## 2024

- **[IJCAI 24]** Rahman, M.R., Khaliq, L.A., Piper, T., Geyer, H., Equey, T., Baume, N., Aikin, R. Maass, W. (2024). SACNN: Self Attention-based Convolutional Neural Network for Fraudulent Behaviour Detection in Sports. *In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI 2024), Main Track*.

- **[NeurIPS 24]** Rahman, M.R., Liu, R., Maass, W. (2024). Incorporating Metabolic Information into LLMs for Anomaly Detection in Clinical Time-Series. *In Workshop on Time Series in the Age of Large Models: Neural Information Processing Systems (NeurIPS 2024)*.

- **[ICIS 24]** Rahman, M.R., Khaliq, L.A., Piper, T., Geyer, H., Equey, T., Baume, N., Aikin, R., Maass, W. (2024). Analysing the Unseen: Leveraging Data Analytics to Combat the Societal Challenge of Doping in Sports. *International Conference on Information Systems, (ICIS 2024), Main Track*.

- **[Springer 24]** Rahman, M.R., Maass, W. (2024). Generative Artificial Intelligence in Anti-doping Analysis in Sports. In Artificial Intelligence in Sports, Movement, and Health (pp. 81-93). Cham: Springer Nature Switzerland.

# 2023

- **[IEEE ICDH 23]** Rahman, M.R.., Hussain, M., Piper, T., Geyer, H., Equey, T., Baume, N., Aikin, R., Maass, W. (2023). Modelling Metabolism Pathways using Graph Representation Learning for Fraud Detection in Sports. *In Proceedings of the IEEE International Conference on Digital Health, (ICDH 2023), Main Track.*

# 2022

- **[ICIS 22]** Rahman, M.R., Piper, T., Geyer, H., Equey, T., Baume, N., Aikin, R., Maass, W. (2022). Data Analytics for Uncovering Fraudulent Behaviour in Elite Sports. *In Proceedings of the International Conference on Information Systems (ICIS 2022), Main Track.*

- **[IEEE ICDH 22]** Rahman, M.R., Bejder, J., Bonne, T. C., Andersen, A. B., Huertas, J. R., Aikin, R., Nordborg, N. B., Maass, W. (2022). Detection of Erythropoietin in Blood to Uncover Doping in Sports using Machine Learning. *In Proceedings of the IEEE International Conference on Digital Health (ICDH 2022), Main Track.*

# List of figures

# List of tables

# Section I: Introduction and Preliminaries

# Chapter 1

# Introduction

*"The thing that doesn't fit is the thing that is most interesting."* - Richard Feynman

The famous quote by Richard Feynman reflects a foundational principle of scientific exploration, i.e., meaningful insights often emerge from deviations that challenge expectations. Longitudinal clinical data refers to repeated observations collected from the same subject over time, offering a unique perspective for analyzing individual-level dynamics that are not visible in cross-sectional datasets [254, 6]. Unlike single measurements, longitudinal data captures temporal dependencies and trends, making it particularly valuable for understanding gradual changes in health or physiology. In clinical research, longitudinal analysis has become essential for identifying early indicators of disease progression and underlying physiological changes [114, 246, 196]. The increasing availability of temporal health data through electronic health records, diagnostic laboratory systems, and home-based monitoring technologies further improves its potential [208, 79, 259]. For example, studies on masked hypertension show that patients with normal blood pressure values during clinical visits may exhibit abnormal rises in home environments, with trends over weeks revealing patterns that individual measurements overlook [222]. Such examples highlight the diagnostic value of longitudinal context for establishing patient-specific baselines and monitoring subtle deviations. These deviations, seemingly minor in isolation, can signal clinically meaningful shifts when viewed against an individual's temporal trajectory. Detecting them requires robust anomaly detection methods that can distinguish true pathological changes from normal physiological variability. By flagging irregularities embedded within longitudinal data, anomaly detection enables earlier interventions, supports personalized treatment decisions, and improves the reliability of monitoring systems in high-stakes domains such as anti-doping analysis in sports.

Anomaly detection methods are developed to identify patterns that do not fit the expected behaviour and can flag these early signs, allowing earlier intervention and closer monitoring before clinical symptoms appear [127, 49, 113]. This approach has become increasingly valuable in remote health monitoring, where wearable devices produce continuous data streams that capture heart rate, oxygen saturation, and physical activity [252, 293]. In conditions such as atrial fibrillation or bipolar disorder, symptoms may occur occasionally and remain undetected in clinic visits [63, 95]. Detecting early signs of deterioration requires models that can separate true physiological changes from background variability. When analyzed longitudinally, shifts in sleep cycles, activity patterns, or heart rhythm can provide the earliest signal of an acute event or relapse [299, 142]. Personalized anomaly detection supports timely intervention [216], improving patient outcomes [13] while reducing strain on healthcare systems. These represent only a few examples; many other domains also benefit from anomaly detection in longitudinal clinical data, which highlights its broad applicability and importance.

This thesis mainly focuses on anomaly detection in longitudinal clinical data, particularly for the application in anti-doping analysis in sports [192, 279]. Broadly, doping can be understood in two ways: first, the practice of using prohibited substances or methods to enhance performance, and second, the deliberate hiding of such practices to avoid detection. Athletes seeking a competitive advantage often engage in strategies such as micro-dosing or timing interventions to reduce the likelihood of detection during scheduled tests [237, 180]. To counter these practices, the World Anti-Doping Agency (WADA) introduced the Athlete Biological Passport (ABP), a longitudinal monitoring system that tracks variables such as haemoglobin concentration and steroid hormones over time [273, 272]. The ABP aims to detect changes in an athlete's biological parameters that may indicate manipulation. For example, a suppressed epitestosterone level followed by an uncharacteristic rise in testosterone may individually fall within population norms, yet represent a manipulated pattern when evaluated longitudinally [187]. Detecting such trends requires models that account for short-term irregularities and long-term drifts relative to an athlete's historical profile.

Blood doping is one of the most prominent examples of manipulation that can be detected through longitudinal monitoring [205, 225, 131]. Athletes may increase their red blood cell mass using erythropoiesis-stimulating agents like human recombinant erythropoietin (rhEPO) or by receiving stored blood transfusions [247, 76]. These interventions improve endurance but change haematological parameters such as haemoglobin concentration and reticulocyte percentage. While direct detection is challenging, such practices often leave physiological signatures that deviate from an athlete's normal range [94]. The case of Lance Armstrong,

a Tour de France champion, shows the limitations of traditional testing and the potential of longitudinal analysis [224, 64]. Reports show that Armstrong never failed a doping test during his career, yet retrospective analysis of blood data combined with testimony and stored samples revealed fluctuations in haematological markers inconsistent with natural variation or training effects [60, 34]. This case highlights the importance of temporal models that assess intra-individual consistency rather than relying on population thresholds.

Another form of doping is sample swapping, which cannot be detected through substance analysis alone but shows itself through longitudinal inconsistency [281]. In such cases, an athlete may submit a urine or blood sample that does not originate from their own body, often timed to overlap with periods of high doping risk [283]. At the time of collection, the substituted sample may appear analytically clean and meet all laboratory criteria, but it changes the longitudinal trajectory of the athlete's biological profile [249]. Such manipulation creates abrupt changes in biomarker patterns, particularly in variables like steroid ratios or haematological markers that are physiologically implausible when compared with prior data. One of the most extensively documented examples occurred during the 2014 Sochi Winter Olympics [110], where investigations revealed that clean urine samples were swapped behind laboratory walls through a "mouse hole" to evade detection [213]. While the substituted samples appeared clean in isolation, analysis of athlete profiles over time revealed inconsistencies that triggered suspicion. Later, the investigation report by Richard McLaren confirmed these irregularities through forensic and longitudinal evidence [193]. This case demonstrated the need for integrating anomaly-aware models with temporal biomarker analysis that can uncover the most sophisticated forms of manipulation. Together, these examples demonstrate not only the ingenuity of doping practices but also the necessity of longitudinal anomaly detection methods that are robust and context-aware.

From disease monitoring to high-stakes sports regulation, longitudinal anomaly detection plays an important role in identifying meaningful deviations, enabling early intervention, and safeguarding system integrity. Yet, working with longitudinal clinical data introduces significant methodological challenges. Measurements are often collected at irregular intervals, profiles vary in length across individuals, and the number of samples available per subject can be very limited. In addition, biological data are inherently noisy, with missing values and natural variability that complicate the distinction between normal fluctuations and clinically relevant anomalies. In anti-doping, these challenges are compounded by the strategic behaviours of athletes who deliberately seek to evade detection. Doping practices such as micro-dosing or blood transfusions can create only subtle deviations that unfold gradually, while concealment strategies like sample swapping introduce abrupt but deceptive shifts in biomarker profiles. Ground-truth labels of confirmed doping cases are scarce,

making supervised model training difficult, and false positives carry serious ethical and legal consequences. At the same time, any detected anomalies must be interpretable and biologically plausible to be admissible in regulatory or forensic contexts. Despite existing tools such as the Athlete Biological Passport, current approaches are still limited in their ability to address these complexities, highlighting a gap for methodological innovation. This work is motivated by these challenges and aims to develop methods for interpretable, domain-aware anomaly detection in longitudinal clinical data. By addressing both the data-related complexities and the domain-specific needs, the work aims to contribute new approaches that improve the reliability, fairness, and practical utility of anomaly detection in real-world clinical and sports settings.

## 1.1 Research Problem

In anti-doping analysis, detecting anomalies has particular significance for identifying drug abuse and monitoring atypical physiological patterns that may arise from manipulation or pathology [127]. With the increasing availability of longitudinal data collected from different sporting events, there is growing potential to detect subtle deviations in an athlete's longitudinal profile [93]. However, working with such data poses a set of methodological and domain-specific challenges that require models capable of understanding complex, dynamic, and often incomplete behaviour. This thesis focuses on two main categories of challenges: those arising from the nature of longitudinal clinical data itself, and those specific to the domain of anti-doping.

The first set of challenges arises from the structural properties of longitudinal clinical data. Biomarker levels naturally fluctuate due to circadian rhythms, environmental conditions, training cycles, or transient illnesses, creating a background of variability that is unrelated to doping practices [156]. For example, the urinary testosterone/epitestosterone (T/E) ratio may rise temporarily after intensive training sessions or shift due to inter-individual enzymatic differences, yet such changes can still fall within an athlete's normal physiology [290]. Therefore, anomaly detection cannot rely on static population thresholds but should be evaluated against each athlete's historical trajectory. This difficulty is compounded by the sparsity and irregularity of testing data, i.e., athletes are not sampled at fixed intervals, and months may pass between urine collections, leaving profiles with large temporal gaps and missing entries. In practice, many athletes, particularly those who are younger or less frequently tested, may have only two or three recorded samples over several years. Under such conditions, distinguishing between genuine manipulations, such as sample swapping or micro-dosing, and natural physiological variation becomes highly non-trivial.

Models that overlook irregular sampling, heterogeneous profile lengths, or the individualized temporal context risk either falsely flagging clean athletes or missing sophisticated doping interventions.

Next, unlike clinical domains where ground-truth labels may be available through diagnostic confirmation, in anti-doping a flagged profile rarely comes with definitive proof of doping unless validated by DNA testing or expert investigation [318]. This scarcity of labelled anomalies limits the use of supervised learning approaches. Moreover, athletes actively employ concealment strategies such as micro-dosing or sample swapping, which produce either subtle shifts or abrupt inconsistencies that are difficult to capture using standard algorithms. For instance, during the Sochi Winter Olympics, urine sample substitution initially produced profiles that looked analytically normal but later revealed longitudinal inconsistencies inconsistent with genuine physiology [193]. In this high-stakes setting, models should achieve high specificity to avoid unfairly sanctioning clean athletes, while also remaining interpretable enough for regulatory and forensic use.

Another challenge lies in the limited integration of biological knowledge into machine learning systems [348, 5]. Many existing approaches treat biomarkers as independent, static features, ignoring the structured biochemical pathways in which they operate [4]. For example, testosterone, epitestosterone, and their downstream metabolites are linked through enzymatic reactions that constrain how they co-vary [199]. Ignoring these interdependencies risks generating anomalies that contradict known biology or overlooking those that emerge only through coordinated deviations across markers. Therefore, pathway-aware models are needed to improve both the performance and interpretability of anomaly detection in longitudinal analysis.

Finally, an important challenge lies in the interpretability and reasoning capabilities of anomaly detection systems. In anti-doping, flagged anomalies should not only be statistically significant but also accompanied by explanations that clinicians and legal experts can understand. For example, when a longitudinal steroid profile is flagged, decision-makers require a clear rationale grounded in physiology (e.g., implausible metabolite ratios or sudden deviations inconsistent with known training effects). Without transparent reasoning, even accurate models risk being unusable in practice, as results that cannot be explained or defended are unlikely to be admissible in regulatory or forensic settings.

To address these challenges, this thesis develops anomaly detection methods tailored to the complexities of longitudinal clinical data. The work emphasizes three key aspects: (i) modeling irregular and individualized temporal profiles to distinguish true manipulations from natural variability; (ii) incorporating domain-specific constraints, including metabolic pathway structures, to ensure biological plausibility; and (iii) improving interpretability

and reasoning so that detected anomalies can be explained and supported in regulatory and clinical contexts. Therefore, the following research questions are addressed:

**RQ1 (What):** *What kinds of deviations can be considered anomalies in longitudinal clinical data, and how they can be modeled to capture the inherent complexity of longitudinal data?*

This question addresses the fundamental challenge of defining and detecting anomalies in longitudinal clinical data. Unlike cross-sectional measurements, longitudinal profiles are shaped by complex temporal dynamics, individual variability, and irregular sampling. Anomalies may appear as gradual drifts within an individual's profile (intra-individual) or as deviations from expected physiological patterns observed across populations (inter-individual). For example, a sudden drop in testosterone may be abnormal for one athlete, even if the value still lies within population reference ranges. Similarly, irregular fluctuations in steroid ratios may seem plausible in isolation but become suspicious when viewed against an athlete's established baseline. Capturing these variations requires models that can represent the inherent complexity of longitudinal data and distinguish between natural variability and true manipulation. Therefore, studying this question is important for developing robust anomaly detection approaches that go beyond fixed thresholds and traditional rule-based systems.

**RQ2 (Why):** *Why incorporating domain knowledge (such as metabolic pathways) is important for improving the performance of anomaly detection methods?*

Longitudinal clinical data are inherently structured by the underlying biology of the human body, particularly in domains like metabolism and endocrinology. Many clinical parameters are not independent, i.e., they follow well-established biochemical pathways. For example, changes in the steroid biosynthesis pathway may result in coordinated changes across several hormones, such as testosterone, epitestosterone, etc. Ignoring such dependencies can lead to false positives or biologically implausible anomaly detection results. This research question examines the significance of incorporating domain-specific structures into anomaly detection models. By incorporating this knowledge, models can better differentiate between pathological patterns and normal physiological adaptations. Studying this question is required for creating biologically meaningful algorithms that align with expert knowledge and improve trust in model predictions across anti-doping and other clinical applications.

**RQ3 (How):** *How can interpretable anomaly detection methods be designed to provide domain-informed reasoning that supports decision-making?*

In high-stakes environments like anti-doping, anomaly detection is only useful if its decisions are interpretable and actionable. Black-box models that flag samples as abnormal without providing clear explanations are of limited use to regulatory bodies. This research question addresses the design of models that not only detect anomalies but also provide transparent reasoning, grounded in the domain's terminology. For example, a system detecting abnormal steroid profiles in athletes should explain which specific markers deviated from expected trends, how they relate within a metabolic pathway, and why the pattern is statistically or biologically suspicious. Studying this question is important for bridging the gap between complex systems and practical decision-making, ensuring that the models developed are not just technically accurate but also clinically trustworthy and ethically usable.

## 1.2 Thesis Objectives

The objectives of this thesis are defined based on research gaps in the literature on anomaly detection in longitudinal clinical data. While existing approaches in statistical modeling and machine learning have made progress, they often struggle to fully capture the complexity of longitudinal profiles, adapt to the domain-specific requirements of anti-doping, and produce outputs that are both reliable and interpretable. Therefore, this thesis aims to develop methods that address these limitations by advancing temporal modeling, integrating biological knowledge, and ensuring interpretability and reasoning. Achieving these objectives is important for improving the robustness and practical utility of anomaly detection in real-world applications such as anti-doping and clinical monitoring.

*Objective 1: Anomaly detection under the complexity of longitudinal data*

Longitudinal data offer the ability to detect anomalies by tracking intra-individual variation over time [83, 62]. However, such data are inherently complex: measurements are often irregularly spaced and heterogeneous in length [296, 150]. In anti-doping, for example, an athlete may be tested only a few times per year, resulting in profiles that contain long gaps and very few observations [270, 313]. Biomarker values also fluctuate naturally due to training cycles, circadian rhythms, and environmental factors, complicating the task of distinguishing manipulation from normal variability [156, 290]. Traditional statistical approaches, such as mixed-effects or autoregressive models [151, 261], typically assume regular sampling and sufficient density, while many machine learning methods require large, balanced datasets to perform reliably [135, 236]. In addition, the domain itself presents the challenge of limited ground-truth labels, since suspicious profiles in anti-doping are rarely confirmed as anomalies without expert or forensic validation. This scarcity makes supervised learning approaches difficult to apply directly. Together, these challenges highlight the need for methods that can

flexibly handle irregular sampling, heterogeneous profile lengths, and limited labelled data, while still enabling anomaly detection that reflects the true complexity of longitudinal clinical profiles. Therefore, the first objective of this thesis is to design methods that capture these temporal dynamics and reliably identify anomalies without depending on dense, uniformly collected, or fully labelled data.

*Objective 2: Incorporation of metabolic pathway structure into anomaly detection*
Most existing approaches treat biomarkers as independent parameters, neglecting the biochemical interactions between them. In reality, these markers are interconnected through metabolic pathways and enzymatic processes that constrain how they can vary under normal physiology. For example, the testosterone/epitestosterone (T/E) ratio has long been recognized as an indicator of doping precisely because of its metabolic coupling [69, 272]. Ignoring such dependencies risks detecting statistical outliers that are biologically implausible or, conversely, missing coordinated deviations that signal manipulation. Research in graph learning and structured modeling demonstrates the value of embedding prior knowledge into machine learning systems [298, 195, 28]. Yet, the application of such strategies to longitudinal clinical profiles remains limited, despite the availability of well-characterized pathways in fields such as metabolism and endocrinology. This gap highlights the importance of aligning anomaly detection methods with biological plausibility, so that detected anomalies are not only statistically robust but also interpretable in a physiological and regulatory context. Therefore, the second objective of this thesis is to design methods that incorporate metabolic pathway structure into anomaly detection frameworks, ensuring that results are both statistically valid and biologically consistent.

*Objective 3: Interpretability and domain-informed reasoning in anomaly detection*
The increasing use of machine learning in anomaly detection has raised significant concerns about interpretability. While black-box models can achieve strong predictive accuracy [112], their ambiguity limits their utility in sensitive biomedical and regulatory settings. In anti-doping, flagged anomalies should be scientifically defensible and accompanied by transparent reasoning to withstand legal and forensic scrutiny [270, 249]. Similarly, in clinical contexts, decision-support systems must provide outputs that clinicians can assess against established medical knowledge [263, 204]. Existing explanation methods such as SHAP and LIME [240, 179] offer feature-level insights, but these are often abstract, domain-agnostic, and difficult for experts to translate into meaningful physiological narratives. This gap highlights the need for anomaly detection frameworks that combine statistical modeling with domain knowledge, enabling outputs that are interpretable, biologically plausible, and aligned with expert reasoning. Therefore, the third objective of this thesis is to develop

interpretable anomaly detection approaches that embed domain-informed reasoning, ensuring that explanations are transparent and trusted in both biomedical and regulatory applications.

These objectives directly address gaps identified in the literature and show the methodological developments presented in the subsequent chapters. By pursuing these objectives, this thesis contributes an anomaly detection framework that is both data-driven and biologically grounded. The models developed here are intended to serve as practical tools for real-world longitudinal monitoring tasks.

## 1.3 Contributions of the Thesis

This thesis presents a set of significant contributions in the field of anomaly detection in longitudinal clinical analysis, with particular focus on applications in anti-doping analysis. The research addresses three key challenges: i) the detection of anomalies that exhibit challenges related to the longitudinal data complexity, ii) the integration of biological domain knowledge into model architectures to ensure physiological plausibility, and iii) the development of interpretable outputs that align with domain reasoning and support expert decision-making. These works are combined into a unified framework that allows robust and transparent analysis of anomalies in complex longitudinal biomarker profiles. The main contributions of this thesis can be summarized as follows:

### 1. Anomaly Detection for Longitudinal Data

Detecting anomalies in longitudinal clinical data is inherently difficult due to these key challenges: (i) irregular sampling intervals, (ii) heterogeneous profile lengths, (iii) limited numbers of samples per athlete, and (iv) the scarcity of labelled anomalies. These challenges are particularly critical in anti-doping, where athlete monitoring data often contain only a handful of urine samples collected over long periods, making it difficult to distinguish between natural variability and deliberate manipulation. To address these challenges, this thesis introduces two complementary neural architectures that adopt different strategies while targeting the same set of constraints.

The Self Attention-based Convolutional Neural Network (SACNN) addresses irregular sampling and heterogeneous profile lengths by constructing structured subsequences from each athlete's profile and applying attention-weighted convolutional layers to capture both temporal dependencies and structural relationships among biomarkers. The attention mechanism enables the model to focus on contextually informative samples within irregular trajectories, while convolutional filters capture local interactions across biomarkers. An

adversarial training module further improves robustness under covariate shifts, reducing the dependency on scarce ground-truth anomalies. Profile-level anomaly scores are obtained through aggregation across subsequences, allowing SACNN to detect subtle irregularities such as sample swapping with high sensitivity.

The Subsampling-based Convolutional Neural Network (SCNN) addresses the same challenges from a different angle by leveraging a triplet-based subsampling strategy. Each athlete's profile is decomposed into multiple temporally ordered subsamples, which are processed by convolutional layers to extract implicit differential features. This design is particularly effective for limited-data settings, where even profiles with only two samples can be expanded into informative subsamples. By aggregating predictions across subsamples, SCNN produces reliable anomaly scores while operating under high-specificity constraints, thereby minimizing costly false positives in anti-doping investigations. Evaluation on both synthetic manipulations and DNA-verified real-world cases shows that SCNN consistently outperforms baseline methods, enabling cost-effective pre-screening of suspicious profiles and reducing the reliance on extensive DNA testing.

Together, SACNN and SCNN provide two complementary strategies for anomaly detection in longitudinal anti-doping data: one emphasizing attention-based representation of irregular trajectories, and the other relying on data-efficient subsampling for sparse profiles. Both models show that it is possible to overcome the inherent limitations of longitudinal athlete monitoring data without requiring large collections of labelled anomalies.

## 2. Incorporation of Metabolism Pathway Structure into Anomaly Detection

Anomaly detection in longitudinal clinical data requires models that follow the underlying biological information. In clinical analysis, different clinical parameters do not behave independently but are connected through biochemical pathways and temporal dynamics. Models that ignore these structured dependencies often produce biologically implausible results or fail to capture complex anomalies. This thesis introduces two complementary approaches to address this challenge by embedding domain structure into the model architecture at different levels of representation.

Structural-Temporal Tokenization for Large Language Models (STT-LLM) is an approach that extends the capabilities of large language models (LLMs) to longitudinal clinical data by converting structured temporal sequences into LLM-compatible token representations. STT-LLM constructs joint embeddings that capture both temporal dynamics and pathway-informed dependencies. These embeddings are then discretized through a two-

stream tokenization process: a structural tokenizer that encodes node-level relationships based on steroid metabolism pathway graphs, and a temporal tokenizer that encodes longitudinal changes over time. The resulting token sequences are directly input to the LLM backbone without modifying its architecture. STT-LLM has been evaluated on real-world steroid datasets for tasks such as anomaly detection and sequence forecasting, where it outperforms pretrained and fine-tuned LLM baselines. The model allows resource-efficient deployment under strict privacy and computation constraints while preserving interpretability through its token structure.

GRAph-based modeling for Metabolism Pathway (GRAMP) addresses biological structure integration through a graph neural network approach. It represents the steroid metabolism pathway as a directed graph, where nodes correspond to biomarkers and edges encode enzymatic interactions. GRAMP leverages graph attention networks to propagate contextual information through the graph, allowing the model to detect changes that may not be visible when examining individual parameters independently. Each node representation is informed by its upstream and downstream biochemical neighbours, allowing the model to capture physiologically consistent patterns of co-variation. GRAMP is particularly effective in identifying pathway-level anomalies such as those induced by hormonal suppression or exogenous substance use, where multiple connected markers deviate coherently. The model improves anomaly detection performance by reducing false positives, and offers interpretable subgraph-level explanations, making it suitable for expert-driven evaluation in regulatory settings.

Together, STT-LLM and GRAMP show two different but synergistic strategies for incorporating biological structure into machine learning models for longitudinal clinical analysis: (i) embedding-guided tokenization for language models and (ii) graph-based neural interpretation. Both approaches allow biologically grounded anomaly detection in complex longitudinal clinical data.

## 3. Interpretability and Domain-Informed Reasoning for Decision Support

In domains such as anti-doping, where decisions have legal and reputational consequences, anomaly detection models should provide more than accurate predictions. They should offer explanations that are transparent and comprehensible to domain experts. To address this requirement, the thesis introduces two complementary methods: (i) a language-based reasoning method, and (ii) a visual trajectory analysis framework that improves interpretability from both linguistic and visual perspectives.

Metabolism Pathway-driven Prompting (MPP) is a prompting method that guides language models to reason about anomalies in longitudinal clinical data using structured biological knowledge. The method combines two sources of information: temporal differences between samples and a domain-specific metabolic pathway graph. Steroid biomarkers are modeled as nodes in a directed graph, with edges representing enzymatic relationships. In parallel, a temporal graph captures distance-based dynamics across time points. These representations are translated into structured textual prompts, allowing the LLM to integrate physiological dependencies and temporal shifts when evaluating whether a test sample constitutes an anomaly. MPP is implemented as a three-stage pipeline, where an initial zero-shot prediction is refined through domain-informed prompts. The empirical results across multiple LLMs and datasets show significant performance improvements over conventional zero-shot and in-context learning prompting methods. MPP improves both anomaly detection and reasoning capability, producing pathway-consistent rationales that align with domain knowledge in anti-doping.

Digital Athlete Passport (DAP) is an unsupervised visual analytics method developed to support the detection and interpretation of suspicious steroid profiles, with a focus on identifying sample swapping cases in anti-doping. The model addresses a major limitation of existing statistical methods, which lack the capability to visualize and quantify intra-individual profile consistency when no ground-truth labels are available. The model projects high-dimensional longitudinal clinical data into three-dimensional space using principal component analysis, capturing variance driven by key physiological trends. A trajectory is formed by connecting an athlete's historical samples in the reduced space, and new samples are scored based on their deviation from the centroid of past observations. DAP computes both cumulative and consecutive distance metrics and flags outliers using a defined rule informed by historical distributions. The method also quantifies feature contributions to the principal components, offering marker-level interpretability. By evaluating real-world steroid datasets, DAP identifies sample swaps and profile inconsistencies, including DNA-verified cases, while providing clear visual diagnostics for the domain expert.

Together, MPP and DAP provide a dual-layered interpretability interface for the anomaly detection framework, one that leverages the generative reasoning capabilities of LLMs and another that offers geometric and statistical insight into profile evolution. These methods enable domain experts to validate model outputs with confidence and to ground decisions in biologically plausible and explainable evidence.

### 4. Software Framework for Anomaly Detection

For the practical deployment of the methods proposed in this thesis, a software framework for longitudinal anomaly detection (CASPIAN) has been developed. CASPIAN is designed as a flexible software tool that brings together all the individual models into a single platform for longitudinal anomaly detection in anti-doping and clinical contexts. Rather than implementing a single pipeline, the framework allows domain experts to select appropriate models based on their specific requirements and analytical goals. For example, an anti-doping official reviewing an athlete's steroid profile can begin with SCNN for anomaly detection in limited sample cases, apply GRAMP to assess whether observed deviations align with plausible biochemical pathways, and use MPP to generate a textual explanation for regulatory reporting. As a tool, it bridges the gap between data-driven anomaly detection and human-centred decision-making, promoting adaptability and operational usability in longitudinal clinical analysis.

## 1.4 Structure of the Thesis

This thesis is structured into five main sections. The structure reflects the cumulative nature of this thesis, with each chapter building on the methodological foundations and practical implementation.

### Section I: Introduction and Preliminaries

- *Chapter 1: Introduction*
  This chapter introduces the research problem and scope of the thesis. It establishes the importance of anomaly detection in longitudinal clinical data, particularly in anti-doping contexts, and outlines the research questions and contributions addressed in this work.

- *Chapter 2: Theoretical Background*
  This chapter provides the theoretical foundation for the thesis. It outlines the key concepts and challenges related to anomaly detection in longitudinal clinical data. It reviews relevant existing methods and positions this work within the anti-doping context.

## Section II: Anomaly Detection for Longitudinal Data

- *Chapter 3: SACNN - Self Attention-based Convolutional Neural Network*
  This chapter introduces SACNN, a neural architecture that leverages self-attention and convolution networks to model structural-temporal behaviour. It addresses the challenge of detecting subtle anomalies in irregularly sampled data without enough labelled supervision.

- *Chapter 4: SCNN - Subsampling-based Convolutional Neural Network*
  This chapter presents SCNN, which extends anomaly detection to athletes with limited profiles. The model uses a subsampling strategy to construct profile embeddings and compute anomaly scores when conventional methods are not applicable due to limited data.

## Section III: Incorporation of Metabolism Pathway Structure

- *Chapter 5: STT-LLM - Structural-Temporal Tokenization for Large Language Models*
  This chapter explains STT-LLM, a method for embedding metabolism pathway structure and temporal dependencies of different biomarker into language model-compatible tokens. It allows small LLMs to process longitudinal clinical data without modifications in LLM backbone, supporting tasks such as anomaly detection and forecasting under resource constraints.

- *Chapter 6: GRAMP - GRAph-based modeling for Metabolism Pathway*
  The chapter introduces GRAMP, which models the metabolism pathway as a directed graph and applies a graph attention mechanism to capture physiologically consistent marker interactions. It demonstrates how domain-specific dependencies improve anomaly detection accuracy and biological relevance.

## Section IV: Interpretability and Domain-Informed Reasoning

- *Chapter 7: MPP - Metabolism Pathway-driven Prompting*
  This chapter presents MPP, a method that integrates pathway knowledge into LLMs to generate textual reasoning for flagged anomalies. It highlights how language-based explanations can align with domain reasoning in forensic and clinical decision-making.

- *Chapter 8: DAP - Digital Athlete Passport*
  This chapter introduces DAP, a visual analytic framework for assessing different

biomarker trajectories using dimensionality reduction and centroid-based analysis. It highlights early detection and transparency in model outputs.

### Section V: Conclusion and Limitation

- *Chapter 9: Software Framework for Longitudinal Anomaly Detection*
  This chapter presents CASPIAN, a framework that contains all models developed in this thesis. It explains how each model can be used by the domain experts based on task complexity and interpretability needs.

- *Chapter 10: Conclusion*
  The final chapter summarizes the key results of the thesis, discusses limitations, and outlines directions for future work. It also discusses how the research questions are addressed in this thesis.

## 1.5 Author Contributions

This thesis presents work that has been conducted in collaboration with different co-authors across multiple research projects and publications. As the first author, Rahman, M.R. led the conceptualization and implementation of all the methodological contributions presented in this thesis. This includes formulating the core research ideas, designing novel architectures, conducting technical developments, performing experimental evaluations, analyzing the results, and drafting publications associated with each chapter. The technical developments were supported by student assistants, Khaliq, L.A., Hammouda, M., Liu, R., and Hussain, M.

Domain expertise in the anti-doping context was provided by Piper, T. and Geyer, H., who contributed specialized knowledge in steroid metabolism and interpretation of athlete biological profiles. Access to the real-world longitudinal steroid profile datasets used throughout this thesis was provided by Equey, T., Baume, N., and Aikin, R., who are members of the research consortium. Supervisory guidance was provided by Maass, W., who provided continuous feedback throughout the development of all methods and reviewed all the publications.

# Chapter 2

# Theoretical Background

## 2.1 Introduction

Longitudinal data forms the backbone of many modern analytical frameworks across disciplines such as health analytics in sports [83, 73]. In contrast to cross-sectional data, which captures a static snapshot of observations at a single time point, longitudinal data comprises repeated measurements collected from the same subjects over time [241]. This temporal dimension enables us to observe dynamic changes and infer causal relationships that are not accessible in static datasets [86, 111]. From a modeling perspective, longitudinal data introduces both methodological challenges and opportunities. Temporal dependencies violate the independent and identically distributed (i.i.d.) assumptions common in standard statistical and machine learning models [104]. Specialized approaches, such as linear mixed-effects models, autoregressive processes, Gaussian processes, and temporal deep learning architectures, are required to capture intra-subject correlations, irregular sampling intervals, and evolving dynamics [83]. The incorporation of subject-specific random effects and time-varying covariates not only allows more accurate predictions but also provides a foundation for individualized anomaly detection, which is the focus of this work.

Applications of longitudinal analysis extend far beyond health analytics. In behavioural science, longitudinal studies are instrumental in tracking cognitive development and behavioural responses across life stages or interventions [253]. These data help differentiate cohort effects from true developmental trends, allowing for more robust conclusions. In personalized medicine, repeated measurements across treatment phases enable the optimization of therapies adapted to individual biological responses, improving treatment efficacy and minimizing side effects [96]. In epidemiology, longitudinal designs have been critical for monitoring disease progression in populations, enabling early detection of risk factors and shaping preventive strategies. In the domain of sports analytics, longitudinal data has become

central to anti-doping research and management. The Athlete Biological Passport (ABP) is a prominent example, employing time-series data of haematological and steroid biomarkers to detect subtle but consistent deviations from an athlete's physiological baseline [317]. Unlike traditional snapshot-based doping tests, the ABP leverages within-subject longitudinal variability to flag suspicious patterns that might indicate the use of performance-enhancing substances. This shows how longitudinal data, when combined with anomaly detection methods, can safeguard fairness in sports and ensure the integrity of competition.

Despite its broad utility, the analysis of longitudinal data poses several computational and methodological challenges, including irregular sampling intervals, heterogeneous profile lengths, missing values, and evolving data distributions [50]. In anti-doping, for example, the irregularity of testing schedules and the limited number of available samples create profiles that are difficult to model reliably, while in clinical contexts, noisy instrumental measurements and missing visits add further complexity. These challenges require robust theoretical frameworks and advanced modeling techniques that can extract meaningful insights while avoiding false positives and negatives. Furthermore, anomaly detection in longitudinal settings is not only a statistical exercise but also a domain-driven problem that must integrate biological plausibility and expert interpretability.

This chapter provides the theoretical foundation for longitudinal clinical analysis, including formal definitions, structural representations, key statistical properties, and challenges associated with anomaly detection in temporal settings. By grounding the discussion in real-world applications and methodological rigour, the chapter sets the stage for the subsequent technical developments of anomaly detection models presented in later chapters.

## 2.2 Definition of Longitudinal Data

Longitudinal data refers to observations obtained by repeatedly measuring the same subjects across multiple time points [62, 116, 265]. Unlike cross-sectional datasets that provide a single snapshot per subject, longitudinal data captures temporal dynamics, allowing for intra-subject trend modeling and inter-subject variability analysis.

**Definition 1.** *Longitudinal data consists of repeated measurements $y_{ij}$ and associated covariates $\mathbf{x}_{ij}$ collected from the same subjects over multiple time points. Mathematically, it is defined as:*

$$\mathscr{D} = \left\{ (y_{ij}, \mathbf{x}_{ij}) \mid i = 1, \ldots, N;\ j = 1, \ldots, n_i \right\},$$

*where $y_{ij} \in \mathbb{R}$ is the scalar response variable for subject $i$ at time point $j$, $\mathbf{x}_{ij} \in \mathbb{R}^p$ is the corresponding $p$-dimensional covariate vector, $N$ is the total number of subjects, and $n_i$ is the number of time-indexed observations for subject $i$, which may vary across subjects.*

Consider a longitudinal study involving $N$ subjects indexed as $i = 1, \ldots, N$. For each subject $i$, measurements are taken at $n_i$ time points, where $n_i$ may vary across subjects (unbalanced design) or remain constant (balanced design). The key components are defined as follows:

**Response Variable ($y_{ij}$)** The scalar response $y_{ij} \in \mathbb{R}$ denotes the observed outcome for subject $i$ at time step $j$, where $j \in \{1, \ldots, n_i\}$. This variable represents the target signal in longitudinal modeling, typically reflecting temporally-evolving phenomena such as clinical severity scores, biomarker levels, or athlete performance metrics [62].

$y_{ij}$ is treated as a temporally-indexed dependent variable whose evolution is influenced by covariates $\mathbf{x}_{ij}$ and potentially latent subject-specific or time-varying stochastic processes. This aligns with dynamic systems modeling and time-series prediction tasks in which the system state is only partially observed through noisy measurements. In many longitudinal learning scenarios, the response series $\mathbf{y}_i$ is the primary output to be predicted or analyzed for anomalies or change points. For each subject $i$, the full temporal response trajectory is represented as a column vector:

$$\mathbf{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in_i} \end{bmatrix}, \quad \text{where } \mathbf{y}_i \in \mathbb{R}^{n_i}. \tag{2.1}$$

This vector-valued response series may exhibit both inter-subject heterogeneity and intra-subject temporal dependencies. Models working on longitudinal data typically assume that $y_{ij}$ is not i.i.d., but exhibits temporal autocorrelation.

**Covariates ($\mathbf{x}_{ij}$)** The covariate vector $\mathbf{x}_{ij} \in \mathbb{R}^p$ encodes the explanatory features associated with subject $i$ at time step $j$. These covariates serve as inputs for modeling the conditional distribution of the response variable $y_{ij}$, and may capture temporal or contextual information

relevant to the underlying process dynamics [62]. In longitudinal modeling, covariates are typically categorized as:

- **Time-invariant covariates**: Features that remain fixed for each subject over time, such as gender, baseline demographics, or group membership. These variables explain inter-subject heterogeneity and can be treated as global identifiers.
- **Time-varying covariates**: Features that evolve across time steps within a subject, such as drug dosage, environmental exposure, physical activity, or dynamic clinical test results. These inputs model intra-subject temporal variation and are crucial for sequence modeling tasks such as forecasting or anomaly detection.

Each covariate vector is structured as:

$$\mathbf{x}_{ij} = \begin{bmatrix} x_{ij1} \\ x_{ij2} \\ \vdots \\ x_{ijp} \end{bmatrix} \in \mathbb{R}^p, \tag{2.2}$$

where $x_{ijk} \in \mathbb{R}$ denotes the value of the $k^{th}$ covariate at time step $j$ for subject $i$. The full covariate matrix for subject $i$ can be written as:

$$\mathbf{X}_i = \begin{bmatrix} \mathbf{x}_{i1} \\ \mathbf{x}_{i2} \\ \vdots \\ \mathbf{x}_{in_i} \end{bmatrix} \in \mathbb{R}^{n_i \times p}, \tag{2.3}$$

where $n_i$ is the number of time points and $p$ is the covariate dimension. This matrix representation enables downstream learning architectures to capture structural and temporal correlations across covariates, and serves as the primary input for modeling tasks such as classification, sequence prediction, or anomaly scoring.

### Longitudinal Data Structure

Longitudinal datasets are commonly organized in the *long format*, where each row represents a single temporally-indexed observation for a specific subject [116]. This tabular layout is particularly well-suited for dynamic querying, batch processing, and integration with time-series learning frameworks. In this representation, the dataset consists of tuples $(i, j, y_{ij}, \mathbf{x}_{ij})$. A simplified illustration of this structure is shown in Table 2.1.

Table 2.1 Structure of Longitudinal Data in Long Format

| **Subject** | **Observation** | **Response** | **Covariates** | | |
|---|---|---|---|---|---|
| 1 | 1 | $y_{11}$ | $x_{111}$ | $\cdots$ | $x_{11p}$ |
| 1 | 2 | $y_{12}$ | $x_{121}$ | $\cdots$ | $x_{12p}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| 1 | $n_1$ | $y_{1n_1}$ | $x_{1n_11}$ | $\cdots$ | $x_{1n_1p}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $N$ | 1 | $y_{N1}$ | $x_{N11}$ | $\cdots$ | $x_{N1p}$ |
| $N$ | 2 | $y_{N2}$ | $x_{N21}$ | $\cdots$ | $x_{N2p}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $N$ | $n_N$ | $y_{Nn_N}$ | $x_{Nn_N1}$ | $\cdots$ | $x_{Nn_Np}$ |

## 2.3 Anomaly Detection in Longitudinal Data

Anomaly detection in longitudinal data involves identifying temporal observations that deviate from expected patterns of behaviour over time [84, 176, 23]. Unlike static anomaly detection, this problem is compounded by temporal dependencies, individual variability, and irregular sampling. Such detection is important in domains such as anti-doping and clinical trials, where early identification of abnormal dynamics can inform timely intervention or diagnosis.

**Definition 2.** *An anomaly is defined as a time-indexed observation that significantly deviates from expected temporal behaviour, either within the subject's trajectory or relative to the population baseline. Mathematically, it is defined as any observation* $\mathbf{x}_{ij}$ *for which:*

$$Anomaly(\mathbf{x}_{ij}) = \begin{cases} 1, & \text{if } D(\mathbf{x}_{ij}, \mathscr{N}_{ij}) > \delta \\ 0, & \text{otherwise} \end{cases}$$

*where* $\mathscr{N}_{ij}$ *represents the modeled normal behavior for subject* $i$ *at time* $j$*,* $D(\cdot,\cdot)$ *is a distance or deviation measure, and* $\delta \in \mathbb{R}_+$ *is a selected threshold.*

This formulation accommodates both point-level deviations at individual time points and trajectory-level deviations across the full temporal sequence. The choice of deviation measure $D$ governs the granularity of detection, whether computed per time step or over the entire sequence, and calculated as a distance, log-likelihood under a generative model, or reconstruction error from an autoencoder. The threshold $\delta$ is typically selected via domain-specific constraints or by leveraging the statistical distribution of scores under the learned normal model.

### 2.3.1 Types of Anomalies

In longitudinal data, anomalies can be categorized into two types based on their scope and nature: *global anomaly* and *local anomaly*. These distinctions are important for adapting detection methods to the specific characteristics of the data and the underlying processes being monitored.

#### Global Anomaly

A global anomaly represents a subject-level deviation where the entire temporal trajectory exhibits atypical dynamics relative to population-level patterns rather than isolated outliers [84]. Global anomalies show altered progression trends or long-term deviations in the subject's profile. These anomalies often signal underlying behavioural drift or performance deterioration and are typically detected using sequence-level modeling and population-normalized distance metrics.

**Definition 3.** *A subject i is said to exhibit a global anomaly if their entire longitudinal profile* $\mathbf{X}_i$ *diverges substantially from population-level normative behavior. This condition can be formalized as:*

$$D(\mathbf{X}_i, \mathscr{P}) > \delta_G$$

*where* $\mathscr{P}$ *represents the learned distribution or manifold capturing typical population trajectories,* $D(\cdot,\cdot)$ *is a sequence-level divergence measure, and* $\delta_G \in \mathbb{R}_+$ *is the global anomaly threshold.*

It indicates a systemic shift in the subject's behaviour, such as a significant change in performance metrics or a chronic health condition. This type of anomaly is characterized by deviations that are consistent across multiple time points, suggesting a persistent change in the underlying process. For example, in clinical monitoring, a patient exhibiting consistent deviations in vital signs across visits may signal the onset of a chronic condition [197]. Similarly, in anti-doping analysis, an athlete whose longitudinal steroid profile continuously diverges from established biological norms across competition seasons or testing periods may indicate the use of performance-enhancing substances [98]. Such anomalies are important to detect, as they often reflect long-term interventions or physiological adaptations that standard pointwise anomaly detection techniques may fail to capture.

### Local Anomaly

A local anomaly refers to a temporally isolated deviation at a specific time point within a subject's longitudinal profile [84]. Unlike global anomalies, which span multiple observations, local anomalies are confined to individual time steps and typically signify transient irregularities rather than systemic shifts. These deviations may result from short-term external influences or momentary physiological fluctuations.

**Definition 4.** *An observation $\mathbf{x}_{ij}$ is considered a local anomaly if it deviates markedly from the subject's own historical or expected temporal dynamics. This condition can be formalized as:*

$$D\left(\mathbf{x}_{ij}, \mathscr{T}_i\right) > \delta_L$$

*where $\mathscr{T}_i$ denotes the subject-specific model of temporal consistency, and $\delta_L \in \mathbb{R}_+$ is the local anomaly threshold.*

Local anomalies reflect abrupt, short-lived deviations from the expected temporal behaviour of a subject, typically confined to a single or a few consecutive time points. For example, in anti-doping, an athlete may experience a sudden spike in a specific steroid biomarker on a single test date, which potentially signals acute doping activity or biological manipulation intended to evade longitudinal tracking systems [98]. Similarly, a temporary drop in performance metrics due to fatigue, minor injury, or environmental stressors can also manifest as local anomalies. Identifying such deviations is essential for targeted retesting and maintaining the integrity of dynamic monitoring frameworks.

## 2.4 Challenges in Anomaly Detection

Anomaly detection in longitudinal data is inherently complex due to the temporal evolution of data, subject-specific patterns, and other factors [62, 116, 265]. Therefore, detecting anomalies requires careful modeling of both the temporal structure and cross-sectional variability. The key challenges are outlined below.

### Temporal Behaviour within Longitudinal Profile

Longitudinal data are inherently temporal, exhibiting strong dependencies between observations across adjacent time points [50]. For a given subject $i$, the observation at time $t$ is often conditionally dependent on past values, and this temporal structure should be explicitly modeled to avoid false anomaly detection. Formally, such dependencies can be represented

as:

$$\mathbf{x}_{it} = f(\mathbf{x}_{i(t-1)}, \mathbf{x}_{i(t-2)}, \ldots, \mathbf{x}_{i(t-k)}) + \varepsilon_{it} \tag{2.4}$$

where $f$ denotes an autoregressive function or a learned temporal transformation, and $\varepsilon_{it}$ is a stochastic noise term. In clinical monitoring, failing to account for temporal causality can result in false positives. For example, in blood glucose monitoring, a rise in glucose levels post-meal is a physiologically expected pattern [145]. Without temporal modeling, such patterns may be erroneously flagged as anomalous despite being contextually normal. Accurately capturing these temporal correlations is thus essential for improving the precision and reliability of anomaly detection systems in longitudinal settings.

### Variability Across Subjects

A fundamental challenge in longitudinal anomaly detection lies in the significant inter-subject variability inherent in real-world datasets [308]. Each subject $i$ may follow a unique distribution $P(\mathbf{X}_i)$, influenced by individual characteristics or baseline levels. A sample $\mathbf{x}_{ij}$ that appears rare or anomalous under the global distribution $P(\mathbf{X})$ may be entirely typical when evaluated under the subject-specific distribution $P(\mathbf{X}_i)$.

This heterogeneity implies that a global anomaly scoring function may lead to misclassification, particularly for subjects with naturally divergent patterns. To address this, anomaly detection models should support personalization through subject-specific scoring functions:

$$Anomaly(\mathbf{x}_{ij}) = D(\mathbf{x}_{ij}, \hat{\mathbf{x}}_{ij}) \tag{2.5}$$

where $\hat{\mathbf{x}}_{ij}$ denotes the expected value or predictive estimate of $\mathbf{x}_{ij}$ based on subject $i$'s historical trajectory.

### Irregular Time Intervals and Missing Data

Another challenge in longitudinal data analysis is the presence of irregular time intervals and missing observations [148, 150, 128]. In practice, measurements are often recorded at non-uniform time steps, and data may be intermittently missing due to equipment failure or resource limitations. Let $\Delta t_{ij} = t_{ij} - t_{i(j-1)}$ denote the time gap between consecutive observations for subject $i$. When $\Delta t_{ij}$ varies across time points, models that assume regularly sampled time series become unreliable or invalid.

Additionally, missing values denoted as $\mathbf{x}_{ij} = \text{NaN}$ further complicate anomaly detection, particularly when they occur non-randomly or in correlated patterns. These issues require models that can handle both temporal irregularity and incomplete data. Recent advances

in attention-based models with time encoding and imputation-aware training offer scalable solutions for handling these irregularities in modern deep learning pipelines [146, 36].

### High Dimensionality and Multivariate Complexity

Longitudinal observations are often multivariate, with each time point $\mathbf{x}_{ij} \in \mathbb{R}^p$ comprising multiple correlated clinical parameters. In high-dimensional settings, conventional distance-based anomaly detection methods face significant challenges due to the curse of dimensionality [71, 226]. As the feature dimension $p \to \infty$, pairwise distances between points become increasingly similar, reducing the discriminative power of distance metrics. Mathematically, for any two distinct centers $\mu$ and $\nu$, the following asymptotic behavior can be observed:

$$\frac{\|\mathbf{x}_{ij} - \mu\| - \|\mathbf{x}_{ij} - \nu\|}{\|\mathbf{x}_{ij} - \mu\|} \to 0 \tag{2.6}$$

This phenomenon of distance concentration implies that the relative difference between distances collapses, making it difficult to distinguish anomalies from normal samples based purely on Euclidean distances [149]. Moreover, anomalies may not be evident in marginal distributions but may emerge through complex interactions across subsets of parameters. This necessitates the use of dimensionality reduction techniques and feature selection mechanisms to project the longitudinal data into more informative subspaces. The attention mechanisms and multivariate temporal encoders have shown promise in isolating feature combinations that are most relevant for detecting subtle, high-dimensional anomalies without discarding semantically important signals [9, 335].

### Concept Drift and Evolving Behavior

In longitudinal settings, the underlying data distribution of a subject may evolve over time, known as *concept drift* [177]. This dynamic behaviour violates the stationarity assumption often made by conventional anomaly detection models. Mathematically, for subject $i$, the probability distribution at time $t$ may differ from that at a future time point $t+k$, such that:

$$P(\mathbf{x}_{it}) \neq P(\mathbf{x}_{i(t+k)}) \tag{2.7}$$

This drift reflects intervention-driven changes in the subject's physiological or behavioural state. For example, in clinical monitoring, a patient's clinical profile may evolve due to treatment effects, recovery processes, or even ageing [103]. Without accounting for such evolution, models may incorrectly classify expected transitions as anomalies or miss emerging

abnormal patterns altogether. Therefore, modeling concept drift is essential for maintaining the relevance and accuracy of longitudinal anomaly detection over extended monitoring periods.

### Lack of Labeled Anomalies

A major limitation in longitudinal anomaly detection is the scarcity or complete absence of labeled anomalies [50, 35]. Annotating anomalies often requires domain expertise, which is time-consuming and can be inherently subjective, especially when deviations are subtle or context-dependent. As a result, supervised learning approaches that rely on explicit ground truth labels are typically infeasible. In such cases, models should operate under unsupervised or semi-supervised paradigms, learning to characterize the normal data distribution and flagging deviations from it. A common approach involves training a model $\hat{f}(\mathbf{x}_{ij})$ to reconstruct or predict observations under the assumption of normality and then computing an anomaly score based on deviation magnitude:

$$Anomaly(\mathbf{x}_{ij}) = \|\mathbf{x}_{ij} - \hat{f}(\mathbf{x}_{ij})\| \tag{2.8}$$

Here, $\hat{f}$ may represent a generative model or a probabilistic estimator of the normal data manifold. However, the absence of ground truth labels complicates model evaluation. Consequently, alternative validation strategies are employed, including the use of synthetically injected anomalies, expert-in-the-loop validation and statistical consistency metrics. These surrogate evaluations are important for benchmarking and model selection in real-world applications where labelled anomalies are scarce or ambiguous.

### Evaluation and Interpretability

Evaluating anomaly detection methods in longitudinal settings also presents challenges to the interpretability of flagged outputs, particularly in high-stakes domains such as anti-doping. Black-box anomaly scores are insufficient when actionable decisions depend on understanding *why* an observation was flagged. Instead, models should offer explanations that identify which clinical parameters and time points contributed most to the anomaly decision. More advanced interpretability techniques, such as SHAP values [256] or attention heatmaps [106], can be integrated into deep learning frameworks to generate fine-grained explanations. These techniques support expert-in-the-loop validation, improve model trustworthiness, and facilitate compliance with transparency requirements in regulated domains.

## 2.5 Current Approaches for Anomaly Detection

### 2.5.1 Statistical Methods

Statistical methods form the foundational layer of anomaly detection in longitudinal data analysis. These approaches offer mathematically grounded techniques that model data behaviour over time to identify deviations indicative of abnormal events or systemic disruptions. Broadly, they can be categorized into six classes: univariate outlier detection, classical time series modeling, time series decomposition, statistical process control, structural change detection, and filtering and probabilistic modeling. This categorization reflects an evolution from static, point-based anomaly detection techniques to dynamic, time-aware frameworks that accommodate the complexities of longitudinal data. Each category addresses a specific aspect of temporal variation and pattern recognition, enabling domain experts to select appropriate techniques based on data structure, sampling frequency, and the nature of expected anomalies.

At the most basic level, univariate methods such as the z-score [345] and Interquartile Range (IQR) [297] detect anomalies by identifying samples that fall outside predefined statistical thresholds. The z-score assumes normality and flags data points that deviate significantly from the mean, while the IQR method isolates outliers based on percentile ranges, providing greater robustness to skewed distributions. While computationally simple and easily interpretable, these methods treat each sample in isolation, disregarding temporal correlations and context, which makes them insufficient for capturing subtle longitudinal anomalies that manifest over time. Nonetheless, they are often used in early screening stages or for parameter-specific alerts, especially in datasets with uniform sampling.

To incorporate temporal dependencies, classical time series models such as Moving Average [346], Exponential Smoothing [121], and AutoRegressive Integrated Moving Average (ARIMA) [329] are used to forecast future values and identify anomalies as deviations from expected trajectories. ARIMA models are well-suited for capturing linear autocorrelations and trends, offering a principled way to distinguish between natural temporal variation and unexpected shifts. In scenarios where seasonality is a dominant feature, such as circadian hormonal cycles or periodic clinical measurements, time series decomposition techniques like Seasonal-Trend decomposition (STL) [305] are used to extract trend and seasonal components from the residual signal. Anomalies are then detected within these residuals, assuming they represent irregular behaviour that is not accounted for by the systematic components. However, such models often rely on well-structured and complete temporal sequences, which may not be applicable in real-world clinical monitoring, where sampling is irregular.

Further temporal sensitivity can be achieved through Statistical Process Control methods, including Shewhart charts [143], and Cumulative Sum (CUSUM) [48]. These control charts were originally developed for industrial quality assurance but have been adapted for clinical and environmental monitoring. Shewhart charts are effective in detecting large, abrupt changes, whereas CUSUM excel at identifying small, persistent shifts. Although powerful, these methods typically assume independent and identically distributed observations and normality, assumptions that rarely hold in complex physiological processes. To overcome these constraints, structural change detection techniques such as Pruned Exact Linear Time (PELT) [15] identify unknown change points in the mean or variance of a sequence. This method is particularly relevant in longitudinal data settings where anomalies may correspond to regime shifts, such as transitions from healthy to diseased states or the effect of a pharmacological intervention. It is especially valuable when prior knowledge of anomaly locations is unavailable, but careful tuning of penalty functions is often required to avoid overfitting or false positives.

Filtering and probabilistic modeling frameworks such as Kalman Filters [125] and Gaussian Process Regression (GPR) [101] represent the most sophisticated statistical approaches. Kalman Filters are ideal for modeling systems governed by latent states, enabling recursive estimation and real-time tracking of physiological trends. They can adapt to changing system dynamics and measurement noise, making them suitable for applications like wearable sensor monitoring or adaptive clinical surveillance. GPR is a non-parametric Bayesian method that not only provides point estimates but also quantifies prediction uncertainty through confidence intervals. This probabilistic treatment of time series is well-aligned with the goals of anomaly detection, as observations that fall outside high-probability regions can be flagged as suspicious. While these models offer significant power and flexibility, they are computationally expensive and require rigorous tuning of kernel functions and prior assumptions. Nevertheless, they are among the most promising methods for high-resolution anomaly detection in longitudinal clinical data.

Statistical methods provide a comprehensive toolkit for detecting anomalies in longitudinal data, ranging from simple univariate thresholds to complex temporal and probabilistic frameworks. Their value lies in their theoretical grounding and adaptability to various temporal structures. While limited in their ability to capture non-linear or high-dimensional relationships inherent in clinical datasets, they provide essential baselines. They are often integrated into more advanced machine learning and hybrid systems. A detailed comparative overview of these methods is presented in Table 2.2 for selecting the most appropriate techniques in anomaly detection scenarios.

Table 2.2 Detailed comparison of statistical methods used for anomaly detection in longitudinal data.

| Statistical Method | Type | Description | Strengths | Limitations | Reference |
|---|---|---|---|---|---|
| **Z-score Method** | Univariate Analysis | The z-score method calculates the number of standard deviations a sample is from the mean of the dataset. It is computed as $$Z = \frac{\mathbf{x}_{ij} - \mu_{\mathbf{X}_i}}{\sigma_{\mathbf{X}_i}}$$ where $\mathbf{x}_{ij}$ is the sample, $\mu_{\mathbf{X}_i}$ is the mean, and $\sigma_{\mathbf{X}_i}$ is the standard deviation of longitudinal profile. Samples with a z-score beyond a certain threshold (commonly ±3) are considered anomalies. | - Easy to implement and understand.<br>- Computationally inexpensive, making it suitable for large datasets.<br>- Provides a clear measure of how unusual a sample is relative to the dataset. | - Assumes the data follows a normal distribution; may not perform well with skewed or non-Gaussian data.<br>- The mean and standard deviation can be influenced by extreme values, potentially distorting the z-score calculation. | [345, 130, 333, 332] |
| **Interquartile Range (IQR)** | Univariate Analysis | The IQR method identifies outliers by measuring the spread of the middle 50% of the data. It is calculated as the difference between the third quartile (Q3) and the first quartile (Q1): $$\text{IQR} = Q3 - Q1$$ samples below $Q1 - 1.5 \times \text{IQR}$ or above $Q3 + 1.5 \times \text{IQR}$ are considered anomalies. | - Less sensitive to extreme values and non-normal data distributions.<br>- Simple to calculate and does not require complex statistical software.<br>- Can be easily visualized using box plots. | - Does not account for temporal dependencies in the data.<br>- The multiplier (commonly 1.5) is arbitrary and may not be optimal for all datasets.<br>- Primarily used for single-variable analysis; not directly applicable to multivariate datasets. | [297, 138, 268] |
| **Moving Average / Exponential Smoothing** | Time Series Smoothing | These methods smooth longitudinal data to identify underlying trends and detect anomalies as deviations from the expected pattern. The moving average calculates the average of samples within a fixed-size window that moves over the data, while exponential smoothing applies exponentially decreasing weights to past observations, giving more importance to recent samples. | - Effective at identifying underlying trends by smoothing out short-term fluctuations.<br>- Exponential smoothing can adapt to changes more responsively by adjusting the smoothing parameter.<br>- Relatively easy to implement and understand. | - Both methods can introduce a lag, causing delays in detecting sudden changes or anomalies.<br>- Choosing appropriate window sizes or smoothing parameters can be subjective and may require domain knowledge.<br>- Standard implementations do not account for seasonal patterns without modifications. | [346, 289, 121, 105, 88] |

| Statistical Method | Type | Description | Strengths | Limitations | Reference |
|---|---|---|---|---|---|
| **AutoRegressive Integrated Moving Average (ARIMA)** | Time Series Analysis | ARIMA models capture various components of a time series: autoregression (AR), differencing (I), and moving average (MA). By modeling the dependencies between observations and differencing to achieve stationarity, ARIMA can forecast future points and detect anomalies as deviations from these forecasts. | - Effectively models linear relationships in time-series data.<br>- Provides a framework for making short to medium-term forecasts. | - Struggles with data exhibiting nonlinear patterns.<br>- Selecting appropriate values for AR, I, and MA components can be challenging and often requires expertise.<br>- The model assumes stationarity; non-stationary data require transformation, which may not always be straightforward. | [329, 147, 2, 186, 276, 202] |
| **Seasonal Decomposition (STL)** | Time Series Decomposition | Decomposes a time series into its constituent components: trend, seasonal, and residual. The Seasonal-Trend decomposition is a versatile and robust method that allows for flexible seasonal and trend extraction. By analyzing the residual component, anomalies can be detected as deviations from the expected behavior. | - Capable of managing complex and nonlinear seasonal patterns.<br>- Resistant to outliers, ensuring reliable decomposition. | - STL focuses on a single seasonal component and may not detect multiple seasonalities effectively.<br>- Requires complete seasonal cycles for accurate decomposition, which can be limiting with incomplete data. | [305, 306, 16, 72] |
| **Control Charts (Shewhart, CUSUM)** | Statistical Process Control | Control charts are used to monitor process stability over time by plotting specific statistics of process measurements and comparing them to control limits.<br>Shewhart Charts detect large shifts by monitoring if samples fall outside control limits, typically set at ±3 standard deviations from the process mean.<br>CUSUM (Cumulative Sum) Charts are designed to detect small shifts by accumulating the sum of deviations from the target value, signaling when the cumulative sum exceeds a certain threshold. | - Simple to implement and effective for detecting large process shifts.<br>- More sensitive to small and persistent shifts. | - Less effective at detecting small or gradual shifts in the process.<br>- Can be slow to detect large shifts and may be more complex to implement. | [143, 30, 320, 48, 32, 139] |

| Statistical Method | Type | Description | Strengths | Limitations | Reference |
|---|---|---|---|---|---|
| **Change Point Detection (PELT)** | Structural Change Analysis | Identifies points in time where the statistical properties of longitudinal data change significantly, indicating potential regime shifts or anomalies. | - Effective at detecting abrupt changes in the data generating process.<br>- Applicable to various types of data and can detect multiple change points. | - May produce false positives in noisy data.<br>- Performance depends on the choice of penalty parameters, which can be challenging to set appropriately. | [15, 169, 89] |
| **Kalman Filter** | Time Series Filtering | Kalman filters are used to estimate hidden states in dynamic systems by modeling the observed data as a combination of hidden states and noise. Anomalies can be detected by analyzing the residuals and identifying significant deviations. | - Well-suited for modeling time-varying processes and systems with hidden states.<br>- Provides real-time estimates and updates as new data becomes available. | - Requires accurate specification of the system dynamics and noise characteristics; incorrect models can lead to poor performance.<br>- Standard Kalman Filters assume linearity; may not perform well with nonlinear systems without modifications like Extended Kalman Filters. | [125, 228, 303, 40, 14, 115] |
| **Gaussian Process Regression (GPR)** | Probabilistic Modeling | The GPR is used to model normal temporal behavior, and anomalies are flagged when observations fall outside a specified confidence interval (e.g., 95%). GPR can flexibly model non-linear temporal dynamics by selecting appropriate kernels such as the RBF kernel. | - Provides confidence bounds for predictions, which is useful for probabilistic anomaly scoring.<br>- Capable of modeling complex, non-linear temporal dependencies by using custom kernels.<br>- Naturally incorporates prior knowledge and handles small data regimes well. | - Computationally expensive for large datasets due to $\mathcal{O}(n^3)$ complexity in training and $\mathcal{O}(n^2)$ in prediction, where $n$ is the number of observations.<br>- Performance is sensitive to the choice of kernel and its hyperparameters. | [101, 46, 24, 321, 44] |

### 2.5.2 ML-based Methods

Machine learning (ML) approaches have become increasingly central to anomaly detection in longitudinal data, particularly due to their ability to model non-linear dynamics and temporal evolution in complex systems. Unlike statistical methods that often rely on rigid assumptions or handcrafted features, ML models can learn directly from data to identify patterns and deviations. This capability is especially necessary in longitudinal clinical data, where repeated measurements across time introduce intricate dependencies that are subject- and context-specific. ML methods used in longitudinal anomaly detection can be broadly categorized into probabilistic models, distance- and density-based techniques, and representation learning methods. This taxonomy reflects both the algorithmic structure and how each method captures temporal structure and latent behavior, which are key considerations in modeling irregular and noisy real-world clinical datasets.

Probabilistic models such as Hidden Markov Models (HMMs) [163] and Bayesian Networks [67] represent some of the machine learning approaches for sequential anomaly detection. HMMs capture temporal dynamics by modeling the system as a sequence of latent states governed by transition probabilities, with observed data emitted probabilistically based on the current hidden state. This makes them particularly suitable for systems where behavior evolves through unobserved regimes, such as stages of disease progression or phases in physiological cycles. Bayesian Networks use directed acyclic graphs to encode conditional dependencies among variables and support inference under uncertainty. These models are effective in integrating prior domain knowledge, which is often available in clinical applications, but suffer from scalability issues and strong assumptions such as conditional independence and stationarity. In practice, their performance can degrade in high-dimensional, irregularly sampled data, limiting their applicability to more constrained settings.

Distance- and density-based models provide an alternative that is intuitive and unsupervised, relying on the assumption that normal data points occur in dense clusters, while anomalies lie in low-density regions. Methods like k-Nearest Neighbors (k-NN) [38] flag anomalies by comparing the distance to neighboring points, whereas DBSCAN [274] identifies outliers as points that do not belong to any sufficiently dense cluster. These methods require minimal assumptions about the data distribution and are easily interpretable, which makes them attractive in domains with limited labeled data. However, their simplicity comes at a cost, i.e., they do not explicitly model temporal structure and tend to perform poorly in high-dimensional feature spaces due to the curse of dimensionality. Additionally, they lack robustness when handling irregular longitudinal sequences, making them suboptimal for clinical monitoring tasks without extensive feature engineering or data preprocessing.

Effective and flexible ML approaches for anomaly detection in longitudinal data fall under the representation learning category, particularly deep learning-based methods. These models automatically learn latent representations of temporal dynamics, enabling them to detect complex deviations in multivariate time series. Autoencoders [334] learn to reconstruct inputs and identify anomalies as instances with high reconstruction error, while Recurrent Neural Networks (RNNs) [275] and their extensions, such as Long Short-Term Memory (LSTM) [185] networks, model sequential dependencies explicitly. These architectures are well-suited to capturing long-range temporal context in different domains. More advanced approaches like Generative Adversarial Networks (GANs) [278] learn to generate realistic data and use reconstruction or discrimination loss for anomaly detection. Additionally, ensemble methods like Isolation Forests [324] isolate anomalies using random partitioning trees and can perform well when combined with engineered clinical parameters. Despite their versatility, these models are often data-hungry and suffer from limited interpretability. These are the challenges that are particularly important in high-stakes domains such as anti-doping and clinical monitoring, where transparency and domain alignment are paramount. A detailed comparative summary of machine learning methods is presented in Table 2.3.

Table 2.3 Detailed comparison of machine learning methods used for anomaly detection in longitudinal data.

| ML Method | Type | Description | Strengths | Limitations | Reference |
|---|---|---|---|---|---|
| **Isolation Forest** | Unsupervised | Builds an ensemble of randomly selected trees. Anomalies are expected to be isolated faster due to fewer splits needed, reflecting their rarity and distinctness. The anomaly score is based on the average path length of a sample. | - Fast and efficient for large-scale datasets<br>- No assumptions about data distribution<br>- Effective in high-dimensional spaces | - Not ideal for temporal dependencies<br>- May fail with dense, clustered anomalies<br>- Hyperparameters (e.g., contamination) require tuning | [324, 307, 221, 160, 229] |
| **Autoencoders** | Unsupervised | Neural networks trained to encode and reconstruct input. Anomalies are detected based on high reconstruction error, assuming that the model is trained primarily on "normal" data patterns. | - Captures non-linear dependencies<br>- Suitable for high-dimensional and complex data<br>- Adaptable with CNNs or LSTMs for temporal sequences | - May reconstruct anomalies well, reducing detection<br>- Sensitive to architecture and threshold tuning<br>- Requires enough training data | [334, 336, 227, 284, 311, 42] |
| **One-Class SVM** | Supervised (semi) | Learns a decision boundary around normal data by projecting data into a high-dimensional space. Anything outside the learned hypersphere is flagged as anomalous. | - Strong theoretical foundation<br>- Works well in high-dimensional space<br>- Suitable when only normal class is labeled | - Computationally expensive on large datasets<br>- Requires tuning of kernel and nu parameter<br>- Sensitive to outliers in training data | [183, 118, 11, 78, 257] |
| **Hidden Markov Models (HMM)** | Unsupervised | Probabilistic model assuming that the system being modeled is a Markov process with hidden states. Observations are linked to state transitions; anomalies are flagged when observation likelihood is low. | - Well-suited for sequential data<br>- Good interpretability via hidden states<br>- Incorporates probabilistic uncertainty | - Assumes Markov property (short memory)<br>- Does not capture long-range dependencies<br>- Sensitive to number of hidden states | [163, 159, 87] |
| **Long Short-Term Memory (LSTM)** | Supervised | A type of RNN designed to capture long-term dependencies in sequence data. Learns to predict future values or reconstruct sequences; anomalies are detected via high prediction or reconstruction error. | - Captures long-range, complex temporal dependencies<br>- Effective in longitudinal and multivariate time-series data<br>- Can model both prediction and reconstruction tasks | - Requires large labeled datasets<br>- Sensitive to vanishing/exploding gradients<br>- Computationally expensive to train and deploy | [185, 167, 211, 227, 122] |

| ML Method | Type | Description | Strengths | Limitations | Reference |
|---|---|---|---|---|---|
| **Hierarchical Temporal Memory (HTM)** | Unsupervised | HTM models temporal patterns in streaming data using sparse distributed representations. It continuously learns and adapts, flagging data that doesn't match its learned patterns. | - Online learning (no retraining)<br>- Suitable for streaming and non-stationary data | - Complex configuration<br>- Less mature and limited community support | [322, 12, 243, 19, 198] |
| **k-Nearest Neighbors (k-NN)** | Unsupervised | Calculates the distance to the $k$ nearest neighbors; samples with distances above a certain threshold are considered anomalies. Typically uses Euclidean or Mahalanobis distance. | - Simple and intuitive<br>- No training phase; works well with small datasets<br>- No strong distributional assumptions | - Poor scalability with large datasets<br>- Sensitive to choice of $k$ and distance metric<br>- Does not model temporal dependencies directly | [38, 152, 158, 302] |
| **DBSCAN** | Unsupervised | Clustering method based on density. Points in low-density regions (outside dense clusters) are considered outliers. It doesn't require the number of clusters in advance. | - Detects arbitrarily shaped clusters<br>- Robust to noise and outliers<br>- No need to specify number of clusters | - Struggles with varying density in data<br>- Parameter sensitivity<br>- Not suited for very high-dimensional data | [274, 309, 280, 58] |
| **GANs** | Supervised | Trains a generator to produce synthetic sequences and a discriminator to distinguish real from fake data. Anomalies are detected via poor reconstruction or discrimination performance. | - Learns complex data distributions<br>- Powerful for high-dimensional time series | - GANs are hard to train and prone to instability<br>- Mode collapse can occur<br>- Requires careful architecture and loss design | [278, 162, 155, 347, 165, 47, 327] |
| **Bayesian Networks** | Unsupervised | Models probabilistic relationships between variables in a graph structure. Anomalies are identified based on low probability given the learned conditional dependencies. | - Encodes causal/conditional structure explicitly<br>- Supports reasoning under uncertainty<br>- Integrates domain knowledge | - Requires structural learning or domain input<br>- Scalability challenges with many variables<br>- Less common for temporal data | [67, 175, 81, 66, 209] |

## 2.6 Summary

This chapter provided the theoretical foundation for detecting anomalies in longitudinal clinical data. Longitudinal data offer unique opportunities to model within-subject variability and detect deviations that are not apparent in static cross-sectional data. The formal definition of longitudinal data introduced the structure of response variables and covariates across time points, establishing the basis for temporal modeling. The distinction between time-invariant and time-varying covariates was detailed to prepare for designing algorithms that can process irregular, multivariate time series.

A key focus of this chapter was the precise formulation of anomalies in longitudinal settings for addressing RQ1 and introduced two primary types of anomalies: global and local. Global anomalies represent sustained deviations across an individual's entire profile when compared to a population reference, while local anomalies are short-term deviations relative to the subject's own past behavior. These definitions highlight the necessity for models that can account for both intra-individual temporal consistency and inter-individual population structure. Since labeled anomalies are unavailable in real-world clinical and anti-doping settings, the chapter emphasized unsupervised and self-supervised strategies where models learn normative patterns and identify deviations without requiring explicit labels. The chapter also identified the computational challenges inherent in this domain, including temporal autocorrelation, high inter-subject variability, irregular sampling and concept drift. Each of these factors complicates anomaly detection and necessitates model architectures that go beyond standard i.i.d. assumptions. For example, capturing temporal dependencies is essential for distinguishing between natural physiological variation and meaningful deviations, while handling irregular time intervals is important for preserving signal integrity in sparsely sampled datasets. These issues motivate the design decisions in the later chapters, which introduce structure-aware, temporally adaptive models that are robust to such complexities.

The chapter concluded with a review of current methods for anomaly detection, including classical statistical models and machine learning approaches. While statistical methods offer well-defined assumptions, they often struggle with the high-dimensional, nonlinear, and irregular nature of modern clinical data. Machine learning techniques provide greater modeling capacity but require careful design to preserve interpretability and accommodate unlabeled sequences. This synthesis of conventional and modern perspectives sets the stage for the thesis contributions, which aim to bridge this gap by developing interpretable models for detecting anomalies in longitudinal clinical data, particularly in the context of athlete monitoring in anti-doping.

# Section II: Anomaly Detection for Longitudinal Data

# Chapter 3

# SACNN: Self Attention-based Convolutional Neural Network

## 3.1 Introduction

[1]Sports events, such as the Olympic Games or FIFA World Cup, attract the attention of billions of people around the world. However, the fraudulent behavior by athletes to improve their performance in these events raises many social issues due to ethical and moral reasons [245]. The impact of this can be seen at both individual and societal levels, e.g., disqualification of athletes, or even ban of a nation from competing in future events, etc. [144]. Therefore, it is a global concern that follows international sporting events worldwide, and anti-doping analysis is a crucial measure to fight against these activities in sports [31]. During the recent investigation at the Olympic Games 2014 in Sochi, a new form of fraudulent activity was found. Some athletes try to swap/exchange their doped samples with another individual's clean sample to evade positive tests. This form of doping is referred to as 'sample swapping' [193]. This simple but new form of fraudulent activity became a threat to the whole anti-doping decision-making organization. The anti-doping organization maintains a longitudinal profile of every athlete, which contains records of all the samples collected from that athlete so far for the doping tests [317].

Sports events, such as the Olympic Games or FIFA World Cup, attract the attention of billions of people around the world. However, fraudulent behavior by athletes to improve their performance in these events raises many social issues due to ethical and moral reasons [245].

[1]**Based on Publication:** Rahman, M.R., Khaliq, L.A., Piper, T., Geyer, H., Equey, T., Baume, N., Aikin, R. Maass, W. (2024). SACNN: Self Attention-based Convolutional Neural Network for Fraudulent Behaviour Detection in Sports. *In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI 2024), Main Track.*

The impact of this can be seen at both individual and societal levels, e.g., disqualification of athletes, or even ban of a nation from competing in future events, etc. [144]. Therefore, it is a global concern that follows international sporting events worldwide, and anti-doping analysis is a crucial measure to fight against these activities in sports [31]. During the investigation at the Olympic Games 2014 in Sochi, a new form of fraudulent activity was uncovered: some athletes attempted to swap/exchange their doped samples with another individual's clean sample to evade positive tests. This form of manipulation, referred to as 'sample swapping' [193], became a serious threat to the integrity of anti-doping organizations. To counter such strategies, these organizations maintain a longitudinal profile of every athlete, containing records of all samples collected for doping tests [317].

The primary way to detect sample swapping is to perform DNA analysis across all collected samples [189]. However, this method is both expensive and time-consuming, with estimated costs exceeding $300 million annually [184]. Alternative approaches monitor each sample against athlete-specific reference ranges to detect unusually high values [234, 223, 270]. However, such interpretations fail to address the fundamental challenges highlighted in RQ1: (i) the temporal behavior of biomarkers, where an athlete's physiology evolves over time and fixed ranges may introduce bias, and (ii) the structural behavior of biomarkers, where values are interdependent as part of steroid metabolism pathways [250, 231]. In addition, longitudinal profiles are typically irregularly sampled, with some athletes contributing only a few measurements per year, making it difficult to distinguish genuine manipulation from natural variability. From a domain perspective, the scarcity of confirmed ground-truth labels adds a further challenge: flagged anomalies are rarely validated as true positives unless costly DNA testing or expert review is conducted. Together, these issues demand anomaly detection approaches that are robust to limited, irregular data while also operating effectively without extensive labeled training sets.

This scenario can be well represented as an anomaly detection problem in longitudinal clinical profiles [77], where the anomalous profile is determined based on both structural and temporal behavior. Many existing models for sequential anomaly detection have been explored [167, 343, 251]. However, these models typically rely on manually defined feature spaces and fail to automatically learn the joint impact of temporal evolution and structural dependencies. Moreover, their dependence on labeled anomalies limits their applicability in anti-doping, where such labels are rare. Addressing this gap requires methods that can model the inherent complexity of longitudinal data without relying on fixed feature definitions and uniformly collected datasets, or comprehensive ground-truth annotations. Recent advances, such as attention mechanisms for automatic feature learning [292, 269] and convolutional networks for structural-temporal representations [39, 288], suggest promising directions.

Therefore, this chapter presents a novel approach, the Self-Attention Convolutional Neural Network (SACNN), which jointly models structural-temporal behavior through embedding maps that capture both intrinsic biomarker relationships and their temporal evolution. This model serves as an adaptive approach for preliminary screening, flagging suspicious profiles for confirmatory DNA testing. In doing so, it minimizes false positives while ensuring that no athlete faces unjust penalties without irrefutable evidence. The main contributions of this work can be summarized as follows:

- A novel architecture is proposed based on a self-attention mechanism, convolution layers, and adversarial attack for detecting sample swapping by capturing embeddings from the longitudinal profiles of athletes. To the best of my knowledge, this is the first time a fraud detection problem in sports has been addressed by considering structural-temporal behavior.
- The method is extensively evaluated on various real-world datasets collected by anti-doping organizations and associated laboratories. The experimental results show the efficacy of the proposed model, which could detect more fraudulent athletes with relatively high specificity compared with SoTA baseline models.
- A case study is performed to demonstrate the performance of the proposed model on real-world fraudulent athlete profiles that were tested using DNA analysis.

## 3.2 Related Work

**Attentional Convolution Neural Network**

Many recent studies have shown the advantage of combining an attention mechanism with convolutional networks for a wide range of applications [292, 269], such as medical image segmentation [102], language understanding [258], etc. For example, the attentional convolution network is well exploited in many NLP-related tasks, e.g. text classification [164, 325], sequence-to-sequence prediction [74], document understanding [210], etc. [43] employed an attention model for learning structural-temporal features in fraud detection. This approach develops from a similar intuition and integrates an attention network to generate embedding maps that consider both structural and temporal aspects and let convolutional filters learn the relationships from these embedding maps.

**Fraudulent Detection in Sports**

Detection of fraudulent activities like doping using machine learning is not new in the sports anti-doping community. A Bayesian approach was proposed for the detection of abnormal values in longitudinal profiles [270]. Several studies [239, 310, 304, 230] used different ML algorithms for detecting anomalous samples in the profile. However, the problem of investigating sample swapping has not so far been addressed by machine learning. Currently, it is mainly detected by laboratory-based methods. Studies like [282, 223] showed how different biochemical techniques like gas chromatography-mass spectrometry, DNA-STR analysis, etc., can be used to detect sample swapping. However, these methods ignore the joint feature learning on structural and temporal relationships.

## 3.3 Preliminaries

**Sample** A sample $\mathbf{x}_{ij}$ refers to a urine sample collected from the athlete for the doping test, where each parameter represents the metabolites in human steroid metabolism. This metabolism pathway is a biological mechanism that follows structural relationship [231]. Testing sample $\mathbf{x}_T$ refers to the sample under consideration for the similarity check with the other samples in the longitudinal profile.

**Longitudinal Profile** A longitudinal profile of an athlete $\mathbf{X}_i = \{\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{in_i}\}$ refers to a sequence of samples collected from that athlete at different times, where $n_i$ denotes the total number of samples collected from subject $i$. When $n_i = 2$, it is defined as limited longitudinal profile $\mathbf{X}_{i,lim} = \{\mathbf{x}_{i1}, \mathbf{x}_{i2}\}$. In this case, it is difficult to compute $\mathbf{x}_{i1} \sim \mathbf{x}_{i2}$. The fraudulent behavior refers to when an athlete performs sample swapping, i.e. exchanges their doped sample with a clean sample from another individual. In this case, if the collected sample is $\mathbf{x}_T$, it will not match other samples in the longitudinal profile.

## 3.4 Self Attention-based Convolutional Neural Network (SACNN)

The proposed model consists of three main components: i) subsequence generator, ii) attentional convolution neural network, and iii) aggregate function together with adversarial training, as shown in Fig. 3.1.

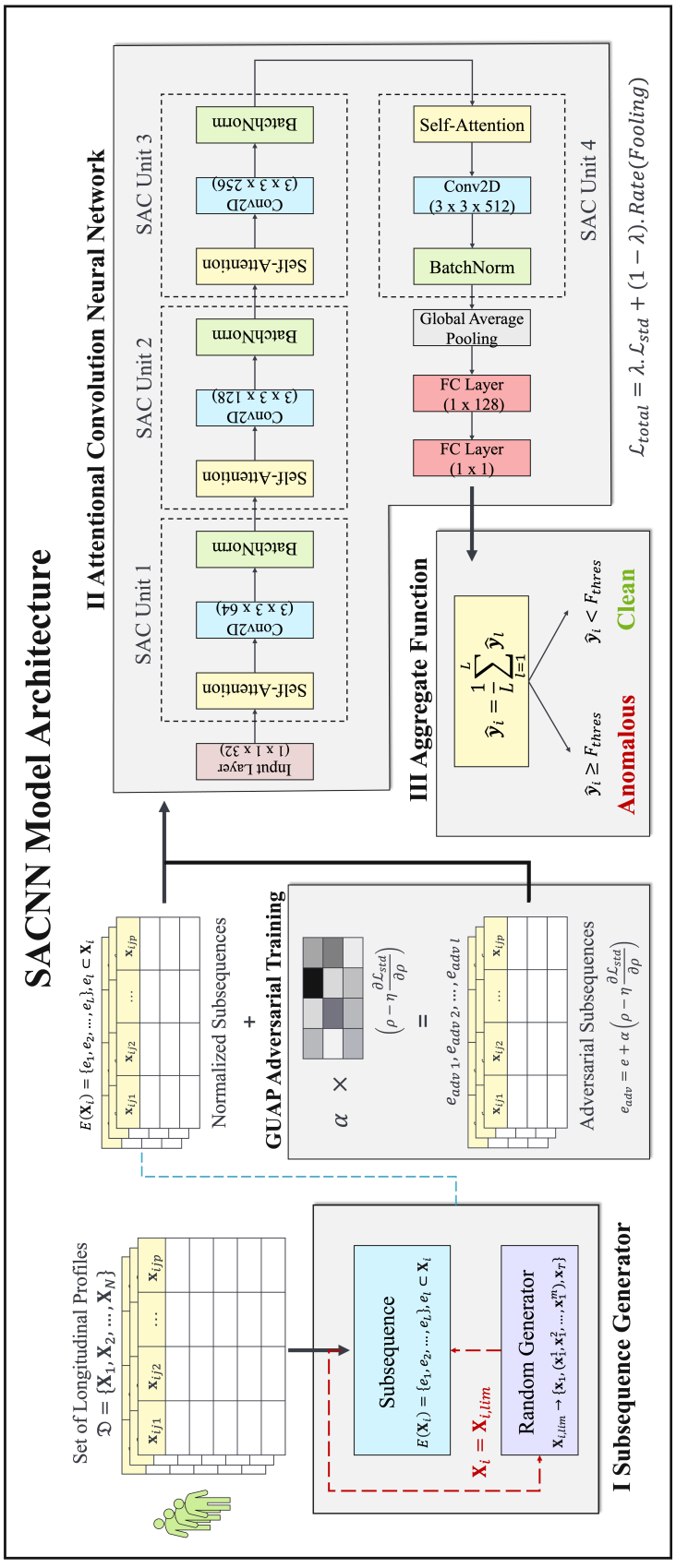


Fig. 3.1 Model architecture of the Self Attention-based Convolutional Neural Network (SACNN) for anomaly detection in longitudinal profiles. The model consists of three main components: i) a subsequence generator that segments profiles into subsequences, ii) an attentional convolutional neural network combining convolution and self-attention layers to process each subsequence, and iii) an aggregate function that integrates anomaly scores to classify the profile as clean or anomalous.

### 3.4.1 Subsequence Generator

The input data is a collection of longitudinal sequences $\mathscr{D} = \{\mathbf{X}_1, \mathbf{X}_2, ...\mathbf{X}_N\}$ with length of the sequences $n_1, n_2, ...n_N$ respectively. Since each sequence is of a different length and the model requires input of fixed dimensions, subsequences of fixed dimensions are generated from the given sequence. The subsequence generator performs this operation in two steps. First, it scans whether the sequence is a limited sequence $\mathbf{X}_{i,lim}$. In this case, it is not possible to generate the subsequences, so the random generator randomly generates $m$ additional samples based on sample $\mathbf{x}_{i1}$ within the measurement uncertainty limit of $\pm 10\%$ and $\mathbf{x}_{i2}$ can be treated as $\mathbf{x}_T$. It is a standard systematical uncertainty caused by the quantification instrument taken from biochemical domain experts [316]. The output of the random generator is the sequence consisting of $\mathbf{x}_{i1}$, $m$ generated samples, and $\mathbf{x}_{i2}$. Next, the generator encodes each sequence into a set of subsequences denoted by $E(\mathbf{X}_i)$ as shown below:

$$E(\mathbf{X}_i) = \{e_1, e_2, ..., e_L\}, \;\; e_l \subset \mathbf{X}_i \tag{3.1}$$

Each subsequence $e_l$ has a fixed length denoted by $len(e_l)$ and consists of $e_l = \{\mathbf{x}_{ij}, \mathbf{x}_{i(j+1)}, ..., \mathbf{x}_{i(\mathbf{x}_T-1)}, \mathbf{x}_T\}$, $\mathbf{x}_{ij}, ..., \mathbf{x}_{i(\mathbf{x}_T-1)} \in \mathbf{X}_i$. In this case, the similarity of $\mathbf{x}_T$ is compared with the other samples in the subsequence. The generator generates sequences corresponding to all the possible combinations of the samples by keeping the longitudinal aspect. This step is similar to the sliding window operation. However, the main difference in this case is that all possible combinations of the samples with $\mathbf{x}_T$ are considered, allowing the model to learn the structural-temporal relationships within all the combinations. The number of subsequences $L$ can be calculated by:

$$L(n_i, n_{\mathbf{x}_T}) \;=\; \left( \frac{n_i!}{(len(e_l)!(n_i - len(e_l))!)} \right)^{n_{\mathbf{x}_T}} \tag{3.2}$$

where $n_i$ represents the number of samples in the sequence $\mathbf{X}_i$, and $n_{\mathbf{x}_T}$ represents the number of testing samples under consideration. These subsequences are then normalized separately. Therefore, the output is a set of normalized subsequences.

### 3.4.2 Attentional Convolution Neural Network

The network architecture consists of an input layer, four SAC units and a fully connected layer.

**Input Layer** The input layer is the Conv1D layer with 32 filters ($1 \times 1 \times 32$). It takes subsequence $e_l$ as input to perform a convolution operation and generates low-level embeddings

for the given subsequence while preserving the structural dimension. Therefore, the output is in a tensor format $\chi \in \mathbb{R}^{N_1 \times N_2 \times N_3}$, where $N_1, N_2, N_3$ denote the length of subsequence, number of parameters, number of filters respectively.

**SAC Unit** Each unit comprises a self-attention layer, a 2D convolution layer and batch normalization.

*1) Self-Attention Layer:* Fig. 3.2 shows the layered architecture where the input tensor is first flattened using the reshaping layer. This is to make sure 2D embedding sequence is fed into the attention layer.

$$\chi \in \mathbb{R}^{N_1 \times N_2 \times N_3} \longrightarrow \mathbb{R}^{N_1 . N_2 \times N_3} \tag{3.3}$$

The self-attention layer is used for two reasons. Firstly, the structural-temporal relationships of the embedding subsequence are of interest, i.e., each parameter of each sample is compared with itself. The attentional weights represent this relationship and can be used to generate high-level embeddings. Secondly, it increases the receptive field of the convolutional layer without adding computational costs associated with very large filter sizes.

The self-attention layer maps a query $Q_i$ and a set of key-value pairs $(K_i, V_i)$ to an output. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed using the given query and the key. In this case, given the low-level embedding sequence from the input layer, the dot-product attention operation can be computed as:

$$H_i = softmax\left(\frac{Q_i K_i^T}{\sqrt{d}})V_i\right) \tag{3.4}$$

The single attention layer performing $h$ multi-head attention operation can be computed as:

$$MultiHead = Concat(H_1, H_2, ..., H_h)w^o \tag{3.5}$$

where $Q_i = \chi . w_i^Q$, $K_i = \chi . w_i^K$, $V_i = \chi . w_i^V$ and the learned attentional weights of the attention layer are:

$$w_i^Q, w_i^K, w_i^V \in \mathbb{R}^{N_1 . N_2 \times N_3} \tag{3.6}$$

$$w_i^o \in \mathbb{R}^{h . N_3 \times N_1 . N_2} \tag{3.7}$$

The output of the attention layer consists of high-level embeddings with the same structural dimensions, i.e. $\chi^{att} \in \mathbb{R}^{N_1.N_2 \times N_3}$. The reshaping layer changes back the dimension to $\chi^{att} \in \mathbb{R}^{N_1 \times N_2 \times N_3}$ for the convolution layer.

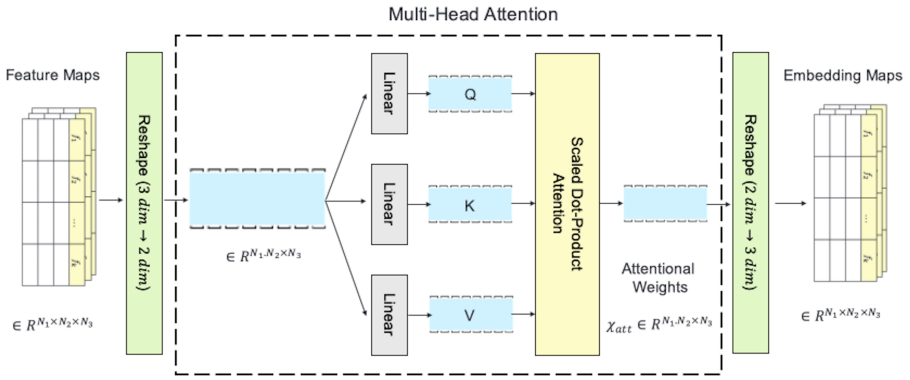


Fig. 3.2 Architecture of the self-attention layer used in SACNN. The input feature maps are reshaped from a 3D to 2D representation and passed through linear projections to generate query (Q), key (K), and value (V) matrices. Scaled dot-product attention computes attentional weights, which are then reshaped back to match the original feature map dimensions, producing the final embedding maps.

*2) 2D Convolution Layer:* The convolution layer is used for three reasons. First, it can learn the structural relationship using filters, which is useful for this task. Second, it can learn more complex patterns from structural space by stacking multiple filters. Lastly, stochastic gradient descent algorithms on commercial hardware could efficiently optimize the network.

A Conv2D layer with $v$ filters ($u \times u \times v$) and padding is used to preserve the structural dimension, where $v = len(e_l)$. It can be represented by:

$$S^c_{i,j} = \left(\chi^{att} * C\right)^c_{i,j} = \sum_{a=0}^{u-1}\sum_{b=0}^{u-1} \chi^{att}_{i+a,j+b} C^c_{a,b} \tag{3.8}$$

where $C^c_{i,j}$ is $c^{th}$ filter which convolves over the embedding map $\chi^{att}$ and represents the element-wise weights. Thus, the output feature map $S^{out}_{i,j}$ is obtained by different Conv2D filters:

$$\chi^{conv} = S^{out}_{i,j} = \sigma\left(\sum_{c=1}^{v} S^c_{i,j} + bias\right) \in \mathbb{R}^{N_1 \times N_2 \times v} \tag{3.9}$$

where $\sigma$ denotes the Leaky-ReLU activation function.

*3) Batch Normalization:* Batch normalization is used to reduce internal covariate shift caused by the attention and convolutional layers and for faster and more stable training.

$$\chi^{BN} = BN(\chi^{conv}) = \gamma\hat{\chi}^{conv} + \beta \in \mathbb{R}^{N_1 \times N_2 \times v} \tag{3.10}$$

where $\gamma$ and $\beta$ are learnable parameters and $\hat{\chi}^{conv}$ denotes normalized input to zero mean and unit variance.

**Fully-Connected Layer** The output of the last SAC unit is first flattened using a global average pooling layer to $\chi^{BN} \in \mathbb{R}^{v}$ . A fully connected layer is used, which takes the flattened input and evaluates the probability of whether it corresponds to a fraudulent trade. If the probability is greater than the threshold value $P_{thres}$, the subsequence can be classified as anomalous.

$$\hat{\mathbf{y}}_l = f(\chi^{BN}) = \sigma(\mathscr{W}.\chi^{BN} + \mathscr{B}) \tag{3.11}$$

where $\mathscr{W}$ denotes weight matrix and $\mathscr{B}$ represents bias vector. The cross-entropy loss was used, defined as:

$$\mathscr{L}_{\text{std}} = -\frac{1}{L}\sum_{l=1}^{L}\Big(\mathbf{y}_l \cdot \log\hat{\mathbf{y}}_l + \big(1-\mathbf{y}_l\big)\cdot\log\big(1-\hat{\mathbf{y}}_l\big)\Big) \tag{3.12}$$

where $\hat{\mathbf{y}}_l \in \{0,1\}$ denotes the predicted label for the subsequence, and $\mathbf{y}_l \in \{0,1\}$ denotes the ground truth label, which is set to 1 if the subsequence is anomalous and 0 otherwise. $f(\chi^{BN})$ is the detection function that maps $\chi^{BN}$ to probability of whether the current subsequence is fraudulent. The proposed model can be optimized through the standard stochastic gradient descent algorithm. The Adam optimizer was used to learn the weights, with the learning rate set to 0.001 and the batch size to 256.

### 3.4.3 Aggregate Function

Since predictions $\hat{\mathbf{y}}_l$ are obtained for each subsequence separately from the fully connected layer. Therefore, an aggregate function is required to combine each prediction and determine the final decision on the longitudinal sequence. The aggregate function tells the likelihood of the given sequence being anomalous. The final classification of the longitudinal sequence can be calculated by:

$$\hat{\mathbf{y}}_i = \begin{cases} 1, & \frac{1}{L}\sum_{l=1}^{L}\hat{\mathbf{y}}_l \geq F_{thres} \\ \\ 0, & else \end{cases} \tag{3.13}$$

where the value of $F_{thres}$ is arbitrary and can be set accordingly to achieve high specificity.

### 3.4.4 Adversarial Training

Adversarial training was used for two reasons. First, it adds novelty to the model by recognizing and classifying variations in the input data that it may not have seen before. This can improve the model's generalization capability and make it more robust to unseen profiles. Second, it is used as fairness-aware training that helps to eliminate discrimination or bias in the predictions. This can be done by incorporating fairness constraints into the training process, such as ensuring that the model's predictions are not systematically worse for certain demographic groups (e.g. based on gender, race, etc.). The Generalized Universal Adversarial Perturbation (GUAP) [340] is a SoTA adversarial attack that was employed to generate adversarial samples for the model. A GUAP attack aims to generate single perturbations to the multiple inputs that cause the model to make a mistake. The generated samples can be represented as:

$$e_{adv} = e + \alpha \left( \rho - \eta \frac{\partial \mathscr{L}_{std}}{\partial \rho} \right) \tag{3.14}$$

A loss function was used that penalizes the model for producing outputs different from the original input. So, the total loss function for the model can be represented as:

$$\mathscr{L}_{total} = \lambda . \mathscr{L}_{std} + (1 - \lambda) . Rate(Fooling) \tag{3.15}$$

where $\rho$ denotes the perturbation vector, $\alpha$ denotes perturbation constant, $\eta$ represents learning rate, and $\lambda$ controls the balance between the two loss functions.

## 3.5 Experiments

### 3.5.1 Datasets

The experiments are performed on real-world athlete datasets consisting of steroid longitudinal profiles with 11 parameters (Table 3.1) gathered by anti-doping agencies at various athletic events worldwide. The data is extracted from the Anti-Doping Administration & Management System (ADAMS) database [312], where each dataset contains $< 20\%$ anomalous profiles, i.e., one sample is swapped in each profile ($n_{\mathbf{x}_T} = 1$). In addition, the dataset Steroid-M$_{lim}$ and Steroid-F$_{lim}$ represent the case of limited longitudinal profile $X_{lim}$.

Table 3.1 Description of all the datasets used in this experiment.

| **Datasets** | **Athlete** | **Profiles** | **Samples** | $len(e_l)$ |
|---|---|---|---|---|
| Steroid-M | Male | 755 | 4214 | 3-20 |
| Steroid-F | Female | 375 | 2307 | 3-20 |
| Steroid-$M_{lim}$ | Male | 737 | 1474 | 2 |
| Steroid-$F_{lim}$ | Female | 293 | 586 | 2 |

### 3.5.2 Baseline Methods

The following SoTA models were employed to compare the performance of the proposed model.

- **Beta-VAE:** [120] Variational autoencoder uses modified reconstruction loss to find anomaly in a sequence.
- **V-LSTM:** [167] Sliding window based approach uses joint learning of VAE and LSTM to generate low-dimensional embeddings for anomaly detection.
- **SUOD:** [343] Ensemble approach produce acceleration to different heterogeneous models for anomaly detection.
- **XGBOD:** [342] Semi-supervised boosting algorithm to extract useful embeddings from the sequence to detect outlier.
- **LSCP:** [344] Unsupervised parallel ensemble algorithm which selects competent detectors in the local region of a sequential instance to detect outlier.
- **AnoGAN:** [251] Deep convolutional generative adversarial network that learns a manifold of normal anatomical variability to detect anomalies.
- **IsoForest:** [168] Unsupervised learning approach that constructs multiple trees which isolate observations with different characteristics to identify outliers.

### 3.5.3 Experimental Settings

The Steroid-All dataset [234] was used for training and validating all the models, containing 50,450 clean profiles from both male and female athletes. In this dataset, 50% of the profiles were randomly selected, and one sample in each selected profile was manually swapped with a sample from a different profile and labeled as an anomalous profile (class 1). The other 50% of the profiles were labeled as clean profiles (class 0). Each profile is normalized

separately, i.e. all the samples within the profile are normalized to the unit norm. The 80% of the dataset was used for training all the models, and 20% for the validation, and the performance of SACNN was evaluated against the baseline models. High specificity in the model's performance is required, as mandated by the anti-doping domain, to avoid false negatives and unnecessary DNA testing (thereby reducing unnecessary costs) and better reflect real-world conditions. Therefore, the baseline models' hyperparameters are optimized to achieve optimal sensitivity under high specificity $(99 \pm 0.1)\%$.

## 3.6 Results

### 3.6.1 Performance Comparison

The performance of SACNN was compared with different baseline models on different steroid datasets for the anomaly detection task, as shown in Table 3.2. The uncertainties are evaluated using the 5-fold cross-validation method. XGBOD and V-LSTM have proven competitive in all baselines, demonstrating the necessity of embedding extraction models for anomaly detection. However, even with an accuracy of $> 70\%$, Beta-VAE could not detect any anomalous profiles (sensitivity of $< 1\%$). In the case of limited profiles, it is observed that all the models except SUOD show poor performance on Steroid-$M_{lim}$ dataset (in terms of sensitivity). However, for the Steroid-$F_{lim}$ dataset, XGBOD and V-LSTM show better performance. The accuracy of all the models is much better because of the highly imbalanced nature of the datasets. The proposed SACNN outperforms all the baselines, i.e., generating structural-temporal embeddings that prove to be effective. SACNN achieves the sensitivity value of $> 50\%$ and AU value of $> 80\%$ on all the datasets.

### 3.6.2 Precision-Recall Curve

Fig. 3.3 shows ROC and PRC curves for all the models evaluated on the Steroid-M dataset. The proposed SACNN model performs better than all the baseline models concerning both curves. The results of V-LSTM, SUOD and LSCP are quite similar. All of them are much better than Beta-VAE. This might be because the fraudulent behavior in longitudinal profiles is too complex for a simple autoencoder model to address. Among all the baselines, XGBOD is shown to be the most competitive. It might be because it generates a deep representation of parameters into embeddings using a boosting algorithm.

The proposed SACNN model consistently outperforms other SoTA baseline models. The reason is: (1) it deals with both structural and temporal behavior and generates embedding

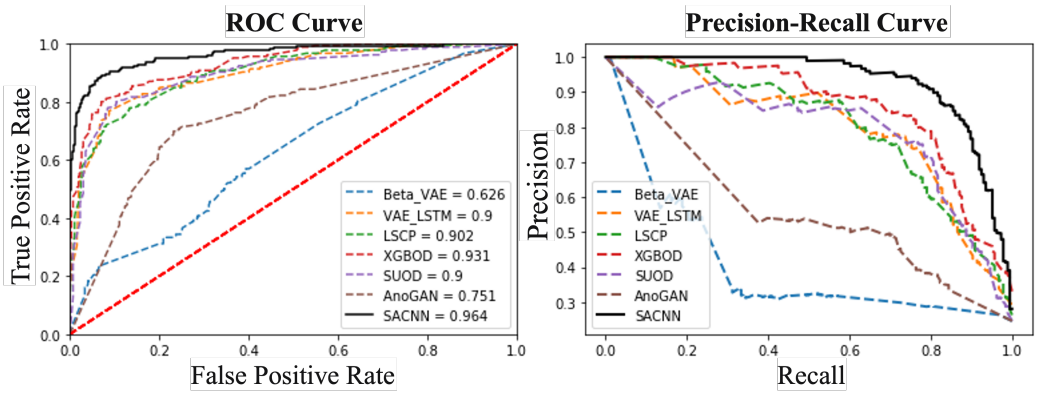


Fig. 3.3 ROC and Precision-Recall (PR) curves comparing SACNN with baseline models on the Steroid-M dataset. SACNN achieves the highest performance, with a ROC-AUC of 0.964, outperforming all baselines. The PR curve further shows SACNN's improved precision across recall values, particularly in the high-recall region, which is important for anomaly detection.

maps using an attention network, contrasted with XGBOD, which only deals with the structural pattern and cannot address the temporal behavior of the longitudinal profile; (2) SACNN uses a convolutional network for better pattern learning from the generated embedding maps. This model works even better at the very beginning of the curve compared to the other baselines. Moreover, this model can accurately detect more anomalous longitudinal profiles with a high specificity, which is quite promising.

### 3.6.3 Parameter Sensitivity

The study was performed on the impact of different values of threshold parameters on the performance of the SACNN model. Both threshold parameters ($P_{thres}$ and $F_{thres}$) were varied from 0 to 1 with a step of 0.1, and the sensitivity and specificity of the model were evaluated. As shown in Fig. 3.4, it can be easily found that these parameters greatly influence the model performance. Sensitivity is reduced as the values of $P_{thres}$ and $F_{thres}$ are increased; however, specificity is improved due to the trade-off between the two. However, it was observed that $P_{thres}$ has a greater impact than $F_{thres}$, as it serves as the threshold applied to the predictions of each subsequence individually. Therefore, $P_{thres} = 0.5$ and $F_{thres} = 0.5$ were set as the default setting, while $P_{thres} = 0.8$ and $F_{thres} = 0.6$ were set for the high specificity setting corresponding to a 99% specificity value.

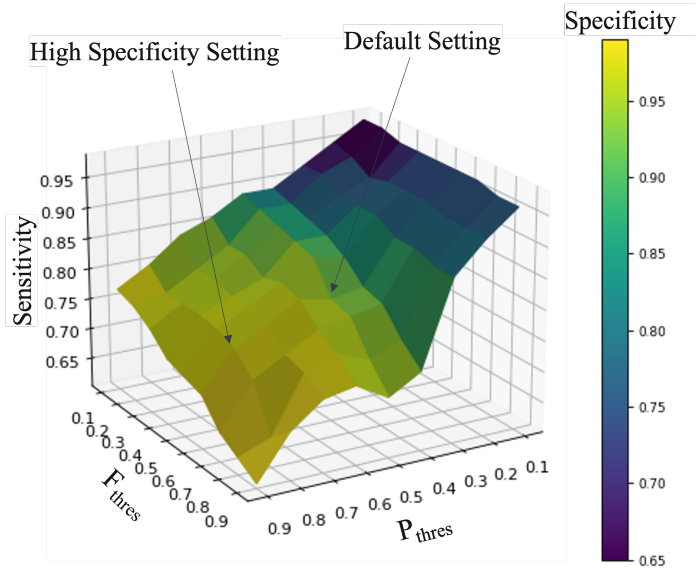


Fig. 3.4 Sensitivity analysis of SACNN threshold parameters $P_{thres}$ and $F_{thres}$. The plot shows how changes in the decision thresholds affect the sensitivity and specificity of the model. The color map encodes specificity values, with higher regions corresponding to settings that prioritize precision. Two operating points, i.e., default and high specificity, are shown for reference.

Table 3.2 Evaluation results of SACNN and all the baseline models on different datasets at high specificity setting. AC = accuracy, SP = specificity, SN = sensitivity and AU = area under ROC curve.

| **Datasets** | **Mtr** | **Beta-VAE** | **V-LSTM** | **SUOD** | **XGBOD** | **LSCP** | **AnoGAN** | **IsoForest** | **SACNN** |
|---|---|---|---|---|---|---|---|---|---|
| Steroid-M | AC | 0.75±0.04 | 0.81±0.03 | 0.79±0.02 | 0.85±0.01 | 0.78±0.03 | 0.77±0.02 | 0.79±0.00 | **0.93±0.02** |
| | SP | 0.99±0.01 | 0.99±0.01 | 0.99±0.01 | 0.99±0.01 | 0.99±0.01 | 0.99±0.01 | 0.99±0.01 | 0.99±0.01 |
| | SN | 0.01±0.01 | 0.31±0.04 | 0.20±0.04 | 0.42±0.02 | 0.13±0.05 | 0.09±0.03 | 0.30±0.01 | 0.74±0.03 |
| | AU | 0.50±0.00 | 0.75±0.03 | 0.73±0.02 | 0.79±0.00 | 0.61±0.02 | 0.60±0.01 | 0.74±0.01 | 0.92±0.01 |
| Steroid-F | AC | 0.78±0.02 | 0.83±0.03 | 0.79±0.02 | 0.84±0.03 | 0.78±0.02 | 0.78±0.01 | 0.82±0.01 | **0.90±0.03** |
| | SP | 0.99±0.01 | 0.99±0.01 | 0.99±0.01 | 0.99±0.01 | 0.99±0.01 | 0.99±0.01 | 0.99±0.01 | 0.99±0.01 |
| | SN | 0.00±0.00 | 0.38±0.05 | 0.10±0.03 | 0.40±0.04 | 0.01±0.03 | 0.00±0.00 | 0.36±0.01 | 0.65±0.03 |
| | AU | 0.50±0.01 | 0.77±0.04 | 0.65±0.01 | 0.79±0.03 | 0.53±0.01 | 0.50±0.00 | 0.78±0.01 | 0.85±0.01 |
| Steroid-$M_{lim}$ | AC | 0.72±0.04 | 0.80±0.03 | 0.81±0.01 | 0.82±0.02 | 0.79±0.02 | 0.77±0.03 | 0.77±0.02 | **0.90±0.02** |
| | SP | 0.99±0.01 | 0.99±0.01 | 0.99±0.01 | 0.99±0.01 | 0.99±0.01 | 0.99±0.01 | 0.99±0.01 | 0.99±0.01 |
| | SN | 0.02±0.01 | 0.23±0.02 | 0.34±0.04 | 0.31±0.02 | 0.29±0.01 | 0.18±0.04 | 0.28±0.01 | 0.70±0.01 |
| | AU | 0.52±0.00 | 0.78±0.01 | 0.76±0.03 | 0.77±0.00 | 0.66±0.00 | 0.60±0.02 | 0.74±0.02 | 0.90±0.00 |
| Steroid-$F_{lim}$ | AC | 0.71±0.03 | 0.79±0.02 | 0.77±0.02 | 0.79±0.01 | 0.73±0.03 | 0.73±0.02 | 0.75±0.03 | **0.84±0.01** |
| | SP | 0.99±0.01 | 0.99±0.01 | 0.99±0.01 | 0.99±0.01 | 0.99±0.01 | 0.99±0.01 | 0.99±0.01 | 0.99±0.01 |
| | SN | 0.01±0.02 | 0.47±0.04 | 0.09±0.03 | 0.50±0.00 | 0.18±0.03 | 0.14±0.03 | 0.33±0.01 | 0.52±0.00 |
| | AU | 0.51±0.00 | 0.72±0.02 | 0.54±0.01 | 0.74±0.00 | 0.61±0.02 | 0.59±0.01 | 0.70±0.01 | 0.81±0.00 |

Table 3.3 Ablation studies showing the model performance evaluated on Steroid-M dataset at high specificity setting. The first block shows the effect of removing key components such as self-attention ($w/o\ Att$), adversarial training ($w/o\ Adv$), and masking ($w\ Mask$), as well as the impact of adding additional samples ($w\ Add\ Samp$). The second block evaluates model depth by varying the number of SAC units.

| **Model** | **AC** | **SN** | **AU** | **# Parameters** |
|---|---|---|---|---|
| $w/o\ Att$ | 0.871 | 0.505 | 0.829 | 1.6M |
| $w/o\ Adv$ | 0.893 | 0.658 | 0.823 | 1.2M |
| $w\ Mask$ | 0.841 | 0.418 | 0.790 | 2.0M |
| $w\ Add\ Samp$ | 0.858 | 0.490 | 0.816 | 2.0M |
| 1 SAC | 0.860 | 0.502 | 0.815 | 50k |
| 2 SAC | 0.873 | 0.523 | 0.834 | 180k |
| 3 SAC | 0.890 | 0.642 | 0.856 | 600k |
| 5 SAC | 0.903 | 0.664 | 0.880 | 7.1M |
| **SACNN** | **0.926** | **0.737** | **0.916** | **2.0M** |

### 3.6.4 Ablation Studies

The effect of different components in the proposed model was studied. First, the attention layer in the SAC units was removed (denoted as $w/o\ Att$), and it was observed that the model's performance was degraded because the convolutional network is now learning the structural relationship of the normalized subsequences instead of structural-temporal embedding maps. This shows the importance of considering the structural-temporal behavior of the sequence. Second, the adversarial attack was removed from the model (denoted as $w/o\ Adv$), and it was observed that the model is less robust to the variation in input data. Next, the number of SAC units in the model was varied, and the performance was evaluated. It was observed that the model performed better by adding SAC units up to a certain point, after which the performance began to drop. The reason might be that adding a SAC unit helps to evolve the embedding maps, but once it is fully generated, adding more units will introduce overfitting. Moreover, adding SAC units exponentially increases the number of trainable parameters. Therefore, 4 SAC units were selected for the SACNN model. The results of the ablation study are shown in Table 3.3.

In addition, two different model variants were tested to understand the significance of the subsequence generator. Instead of generating subsequences of length $len(e_l)$: i) additional samples were generated in the profile based on other samples to achieve uniform sequence length, denoted as $w\ Add\ Samp$; ii) the additional samples were masked with padding, denoted as $w\ Mask$. It was observed that adding additional samples to shorter sequences

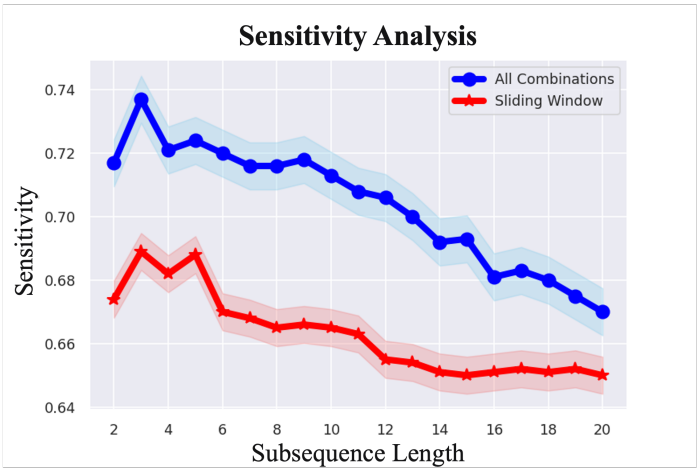


Fig. 3.5 Sensitivity of SACNN as a function of subsequence length $len(e_l)$. The plot compares two subsequence generation strategies: all combinations (blue) and sliding window (red). SACNN achieves consistently higher sensitivity using all combinations, with shorter subsequences yielding better performance. Shaded areas represent standard deviation across different runs.

introduces a bias in the model's decision, as $\mathbf{x}_T$ is known to be dissimilar to the masked or generated samples. Fig. 3.5 shows the effect of different subsequence lengths $len(e_l)$ on the model's performance for both sliding windows and this approach of considering all the combinations for defining the subsequences. It was found that 3 was an optimal subsequence length for this model.

## 3.7 Case Study

A study was performed on real-world proven cases to understand the structural-temporal patterns of the longitudinal profile from the embedding maps. These longitudinal profiles were tested using DNA analysis performed by an accredited anti-doping laboratory and found that 2 profiles were proven for sample swapping, 5 for doping and 22 for clean profiles. This model could able to detect all the sample swapping and doping cases and 20/22 clean profiles. One clean and one anomalous profile (sample swapping) were selected, and the subsequences, along with the embedding maps generated from each SAC unit, were plotted, as shown in Fig. 3.6. The total number of embedding maps generated by the attention mechanism depends on the output of the Conv2D layer of the previous SAC unit. Therefore, these maps represent high-level embeddings. One embedding map from each was plotted to understand how the attentional weights in these maps evolve. In the case of the clean subsequence, higher weights were observed in the embedding map of SAC unit 4 compared

to the anomalous subsequence, indicating a strong structural-temporal relationship among the three samples. Furthermore, the evolution of the embedding maps also demonstrates why at least 4 SAC units are necessary.

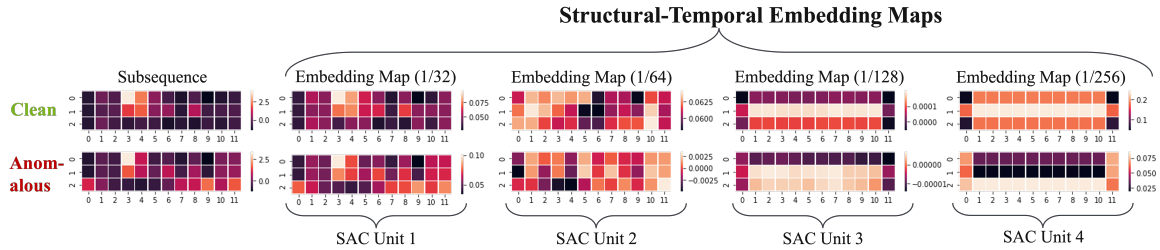


Fig. 3.6 Visualization of structural-temporal embedding maps across different SAC units for clean and anomalous subsequences. Each row shows the progression of feature representations through the SAC units. The clean and anomalous inputs produce distinctly different activation patterns, particularly in later SAC units, highlighting the model's ability to progressively refine structural-temporal features for anomaly detection.

## 3.8 Summary

This chapter introduced the Self Attention-based Convolutional Neural Network (SACNN), an architecture designed to detect anomalies in longitudinal clinical data, particularly in the context of anti-doping analysis in sports. The method addresses the limitations of existing approaches that often ignore the temporal dynamics and structural interdependencies inherent in biological profiles such as urinary steroid profiles. It responds directly to the challenges outlined in RQ1, namely: irregularly sampled data, heterogeneous profile lengths, limited numbers of samples per athlete, and the scarcity of labeled anomaly data.

SACNN solves these challenges through several key innovations. To handle irregular sampling intervals and heterogeneous profile lengths, the model uses a subsequence generator that constructs all valid longitudinal combinations leading up to the test sample, ensuring that meaningful context is preserved even when data are sparse or unevenly distributed. The integration of multi-head self-attention with 2D convolutional layers enables the model to capture both local biomarker interactions and long-range temporal dependencies, dynamically weighting time points according to their relevance for the current prediction. This design allows SACNN to adapt to variability across athletes and testing schedules, avoiding the pitfalls of static thresholds. To address the challenge of limited samples per profile, subsequence expansion using a random generator provides a richer training signal even in cases where athletes contribute only a few samples over multiple years. Finally,

the scarcity of ground-truth anomalies is mitigated by semi-supervised training objectives: contrastive learning extracts individualized patterns of normal physiology, while adversarial augmentation simulates realistic intra-individual variability, reducing overfitting and improving generalization under covariate shifts. Together, these innovations allow SACNN to detect clinically relevant anomalies in longitudinal data without enough explicit labels, by learning context-sensitive representations of normal physiology and identifying deviations from subject-specific baselines. This capability is important in anti-doping, where natural fluctuations may mask manipulation, and few ground-truth cases exist to guide supervised approaches.

As a result, SACNN can detect subtle manipulations, such as sample swapping, by modeling divergence from subject-specific baselines rather than relying on population norms or predefined cutoffs. Extensive empirical evaluation shows the model's effectiveness across multiple real-world datasets, including limited-profile scenarios and gender-specific subgroups. Compared to baseline models such as XGBoost, VAE, and ensemble-based classifiers, SACNN achieves significantly higher sensitivity at a fixed high specificity threshold, which is an essential operational requirement in anti-doping programs where false positives carry substantial investigative and legal costs. For example, SACNN maintained performance above 70% sensitivity at 99% specificity across datasets for both male and female athletes. Moreover, a detailed case study using DNA-verified doping cases highlights SACNN's ability to generate interpretable attention maps that distinguish clean from anomalous profiles based on learned structural-temporal features, offering transparency to domain experts.

In addition to its strong empirical performance, SACNN contributes conceptually by demonstrating a principled way to capture domain-specific structure in longitudinal sequence learning for anomaly detection. The use of attention not only increases performance but also improves interpretability, an essential feature in expert-governed domains like sports regulation and healthcare. By explicitly addressing the four main challenges in RQ1, SACNN shows how anomaly detection systems can remain robust and deployable in practice, while reducing reliance on costly biochemical confirmation methods such as DNA testing. The framework is also generalizable beyond anti-doping, with potential applications in endocrinology, pharmacology, and other fields that rely on longitudinal clinical monitoring.

# Chapter 4

# SCNN: Subsampling-based Convolutional Neural Network

## 4.1 Introduction

[1]Doping in sports is the practice of using prohibited substances, e.g., performance-enhancing drugs (PEDs), to gain an undue advantage over competitors [17]. Aside from the risks it poses to athletes' health, the practice affects the basic principles of fair competition. The World Anti-Doping Agency (WADA) leads the worldwide effort against doping, but the problem continues to be present in many sports events and competitions [18]. The number of athletes involved in doping activities is considerably higher than those detected using doping tests. A study conducted in 2017 determined that the percentage of elite athletes using PEDs might be between 14% and 39%, which is much higher than the small number of positive tests reported in official statistics [55]. WADA disclosed in its 2018 annual report that from 322,000 samples tested globally, only 1.43% were found to be adverse or atypical, showing a significant under-detection [313]. Economically, substantial resources are allocated annually to doping prevention, including testing and educational programs aimed at athletes and support personnel [90]. This comprehensive look at sports doping points to both individual failure and systemic problems within sports governance and societal values.

Sample swapping is an unfair practice involving the exchange of biological samples, such as urine or blood, to avoid a positive doping test. It can usually be seen when a doped athlete's sample containing a high level of a prohibited substance is swapped with a clean sample from another individual [206]. Such practices undermine the value of sports and

[1]**Based on Publication:** Rahman, M.R., Khaliq, L.A., Piper, T., Geyer, H., Equey, T., Baume, N., Aikin, R., Maass, W. (2024). Analyzing the Unseen: Leveraging Data Analytics to Combat the Societal Challenge of Doping in Sports. *International Conference on Information Systems, (ICIS 2024), Main Track.*

challenge anti-doping agencies striving to ensure fairness among athletes. Currently, the screening for prohibited substances and the possibility of sample swapping in sports is carried out through different techniques, including biochemical methods and data analytics. Biochemical techniques, including Isotope Ratio Mass Spectroscopy (IRMS) [20], Gas Liquid Chromatography-Mass Spectrometry (GLC-MS) [281], and High-Performance Liquid Chromatography (HPLC) [20], are among the most important methods for detecting prohibited substances in athletes' urine and blood samples. Although these techniques work, they are costly and take time to set up. While DNA testing is effective for verifying athlete identity, it remains expensive and involves complicated lab processes that may delay the testing process, especially during competition times [178].

Regarding data analytical techniques, the Athlete Biological Passport (ABP) monitors various biological parameters over time to detect any abnormal values in an athlete's profile [223]. Moreover, machine learning algorithms and statistical techniques are employed to analyze the longitudinal profiles of athletes to identify patterns indicative of doping practices [231, 234, 137]. However, the performance of these models is limited by the challenges of longitudinal data, i.e., irregular sampling, heterogeneous profile lengths, and the domain-specific challenge of scarce ground-truth labels. Most anomalies flagged in practice are not conclusively labeled unless verified by costly DNA testing or expert review, which restricts the applicability of supervised learning approaches. This motivates the need for adaptive anomaly detection methods that can work effectively under limited data availability and without reliance on explicit labels.

In this chapter, a method is presented that uses a convolutional-based approach to support anti-doping experts in efficiently identifying sample swapping cases. It pre-screens all samples collected during competitions and flags only suspicious samples, allowing further biochemical tests to be performed. Unlike existing baselines, the model is specifically designed to handle irregular longitudinal profiles and to detect inconsistencies even when only a few samples are available for an athlete. By learning implicit differential consistency across subsequences, the model reduces dependence on ground-truth anomalies while providing a low-cost pre-screening tool. This reduces costs and processing times associated with unnecessary testing of clean samples while strengthening the evidence base for confirmatory investigations. The main contributions of this work can be summarized as follows:

- A data-driven approach is proposed for identifying sample swapping cases in sports that applies a convolutional network on the longitudinal profiles of athletes to quantify similarity with other samples collected from the same athlete.

- The method explicitly addresses challenges of irregular and limited longitudinal profiles, allowing anomaly detection in cases where current SoTA methods have limitations.
- The approach is designed to work under scarce ground-truth labels, learning implicit patterns of consistency without requiring fully annotated anomaly datasets.
- Performance evaluations are performed on data collected from real-world athletes and DNA-proven sample swapping cases conducted by an accredited laboratory.

## 4.2 Related Work

### Laboratory-based Biochemical Testing

The laboratory-based methods include biochemical testing of the blood and urine samples collected from the athletes. A study showed how different factors influence the behavior of longitudinal profile patterns to find suspicious profiles [188]. Another study showed a multidisciplinary approach to determine identical urine samples [281]. The approach includes numerous analytical strategies, i.e., gas chromatography-mass spectrometry with steroid and metabolite profiling, gas chromatography nitrogen/phosphorus detector analysis, high-performance liquid chromatography, and DNA-STR analysis [281]. Moreover, the mass spectrometry, immunological doping control, and forensic chemistry methodologies have also been proven to be useful in finding sample manipulation by athletes [282, 223]. Their recent study showed how the large interindividual variability of the steroidal parameters could be used to find sample swapping cases. All these methods have high costs and long processing times associated with them, which makes it difficult to perform at large scale, especially during athletic competitions like the Olympic Games.

### Data-driven Methods for Anomaly Detection

Several studies in anti-doping analytics discuss the use of data-driven approaches to detect doping activities in sports events. The doping activities are mainly classified into blood doping, steroid doping, and sample swapping. While much of the research has focused on blood and steroid doping, less work has been done to address the issue of sample swapping. For example, different machine learning algorithms were applied to identify the presence of erythropoietin (doping substance) in athletes' blood samples [230]. A study employed different machine learning algorithms to determine the athletes with the highest risk of doping based on their performance data [137]. The use of machine learning to evaluate how much an

athlete's steroid profile deviates from normal population profiles [200, 310]. However, there is less research work focusing on sample swapping using a data-driven approach. The current SoTA method for detecting sample swapping remains the Bayesian method of the Adaptive Model, which is often followed by laboratory testing like DNA testing [270, 223]. In the recent works, the concept of Digital Athlete Passport (DAP) based on athletes' individual profiles was proposed to identify suspicious activity [234].

However, all these works suffer from three main limitations. First, while analyzing the athlete profile, they consider the complete longitudinal profile at the same time. This approach can lead to overfitting, where the model becomes too adapted to the specific samples already in the profile used for analysis and does not generalize well to new, unseen samples belonging to the same profile [234, 310]. Additionally, using the complete longitudinal profile at once may introduce bias, particularly if certain patterns in the profile are overemphasized, potentially leading to inaccurate decisions [214]. Second, many approaches fail to adequately consider the longitudinal aspect of athlete profiles [137, 200]. Without accounting for these temporal dynamics, model may miss significant patterns that could be indicative of fraudulent activities or natural variations, thus reducing the effectiveness of the models in identifying genuine cases of irregularities [25]. Third, these methods require more than two samples in the athlete longitudinal profile, which is problematic for new and young athletes with limited samples collected [234]. This requirement restricts the applicability of current methods to only those athletes with extensive historical data, thereby excluding a segment of newer or younger athletes. Therefore, these works suggest the need to develop a more robust adaptive model to analyze longitudinal athlete profiles in a more efficient manner.

## 4.3 Preliminaries

The aim of this study is to develop a model capable of identifying sample swapping by examining the longitudinal profiles of athletes. The model should be able to detect whether a newly collected steroid profile matches previous samples collected from the athlete over time. The model should determine the relatedness between each collected sample and all the previous samples in the longitudinal profile of the athlete to identify any suspicious activity.

**Longitudinal Profile** The longitudinal profile of each athlete consists of one or more steroid profiles (urine samples) collected at different times. Therefore, the longitudinal profile can be defined as $\mathbf{X}_i = \{\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{in_i}\} \in \mathbb{R}^{n_i \times p}$ where $\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots$ represent the steroid samples for subject $i$, $n_i$ denotes the number of samples collected from subject $i$, and

$p$ represents the total number of steroid parameters. The testing sample $\mathbf{x}_T$, i.e., the sample under consideration, is represented to determine its similarity with previous samples.

It is investigated whether $\mathbf{x}_T$ belongs to the same athlete $i$ based on the other samples contained in the longitudinal profile. For example, consider three athletes with longitudinal profiles $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$ as depicted in Fig. 4.1. Athlete $\mathbf{X}_1$ and $\mathbf{X}_3$ have clean profiles with all samples being similar to each other while athlete $\mathbf{X}_2$ has an anomalous profile. Let us suppose samples $\mathbf{x}_{24}$ and $\mathbf{x}_{27}$ of athlete $\mathbf{X}_2$ are under investigation, i.e., $\mathbf{x}_T = \{\mathbf{x}_{24}, \mathbf{x}_{27}\}$, and could either belong to athlete $\mathbf{X}_2$ or have been swapped to evade a positive doping result. It is also possible that both samples are from another athlete but not $\mathbf{X}_2$ (e.g., $\mathbf{x}_{24}, \mathbf{x}_{27} \in \mathbf{X}_1$) or even from different athletes (e.g., $\mathbf{x}_{24} \in \mathbf{X}_1, \mathbf{x}_{27} \in \mathbf{X}_3$). So, the goal is to identify such samples in the longitudinal profile of athletes. Therefore, an iterative algorithm is required to examine the similarity of each sample with every other sample in the longitudinal profile as denoted by the following expression:

$$\sum_{j=1}^{n_i} \sum_{k=1}^{n_i} \left( \mathbf{x}_{ij} \sim \mathbf{x}_{ik},\ \forall j,k \in \mathbb{N},\ j \neq k \right) \tag{4.1}$$

**Limited Profile** To determine the similarity between $\mathbf{x}_T$ and other samples in the longitudinal profile, it is necessary to have a minimum of two other samples from the same athlete. However, in real-world situations, the number of samples may be limited $\mathbf{X}_{i,lim} = \{\mathbf{x}_{i1}, \mathbf{x}_{i2}\}$. In this scenario, calculating $\mathbf{x}_{i1} \sim \mathbf{x}_{i2}$ can be challenging. Therefore, a solution is needed to address the issue of limited samples in the longitudinal profiles of athletes.

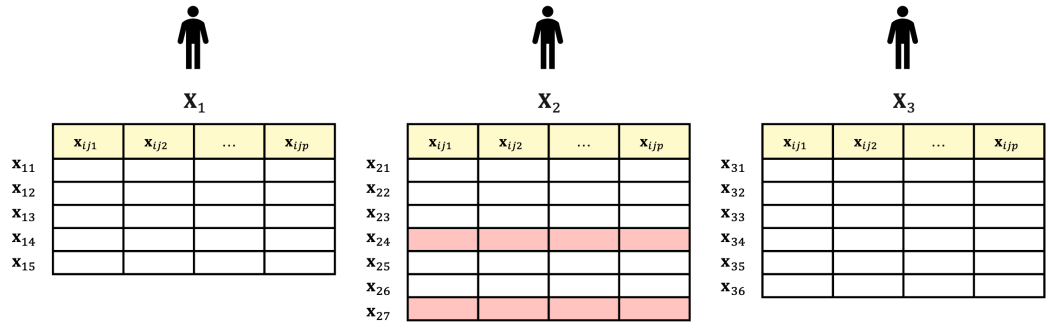


Fig. 4.1 Illustration of longitudinal steroid profiles for athletes $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$. The white color indicates observed values for the samples, while the red color in $\mathbf{X}_2$'s profile indicates anomalous samples.

## 4.4 Subsampling-based Convolutional Neural Network (SCNN)

The primary way to find the similarity among the samples is to evaluate the arithmetic difference of each steroid parameter of the samples within the longitudinal profile by the following expression:

$$\Delta = \sum_{j=1}^{n_i-1} \delta(\mathbf{x}_T, \mathbf{x}_{ij}) = \sum_{j=1}^{n_i-1} \sum_{k=1}^{p} \left| x_{T,k} - x_{ij,k} \right|, \quad p = \text{parameter space} \tag{4.2}$$

Minimizing this expression results in:

$$\frac{d^2\Delta}{d^2\mathbf{x}_{ij}} > 0 \implies \mathbf{x}_T \in \mathbf{X}_i \tag{4.3}$$

However, the longitudinal profile contains both primary and derived parameters. Therefore, calculating the arithmetic difference explicitly and using a classifier to train on the difference leads to the loss of some implicit differential information. Therefore, automatic feature learning is needed, which allows the model to learn this implicit differential information by itself without being explicitly provided with it.

The proposed Subsampling-based Convolutional Neural Network (SCNN) consists of three main components: (i) subsample generator, (ii) convolutional neural network, and (iii) aggregate function, as shown in Fig. 4.2. The model converts the longitudinal profile into a set of subsamples, which are used to create a set of embedding maps by different filters of the convolutional network. These embedding maps consist of the implicit differential information among the samples within the subsample. Finally, the model decision is computed by aggregating the predictions on the network's output over all the subsamples.

### 4.4.1 Subsample Generator

Let us consider the longitudinal profile of dimension $(n_i \times p)$, where each sample is a 1D array of $p$ parameters, i.e., $(1 \times p)$. The subsample generator consists of a subsampler and a random generator which perform two operations. First, the subsampler scans whether the profile consists of limited samples. In the case of limited samples, the subsampler passes the profile to the random generator, which randomly generates additional samples based on sample $\mathbf{x}_{i1}$ within the uncertainty limit of the measurement and quantification device, i.e., $\pm 10\%$. This limit is set by the domain experts of anti-doping [318]. This uncertainty limit is considered as a reference range of the concentration values for the particular sample, and any value within this range could be representative of the same sample. Therefore, the uncertainty

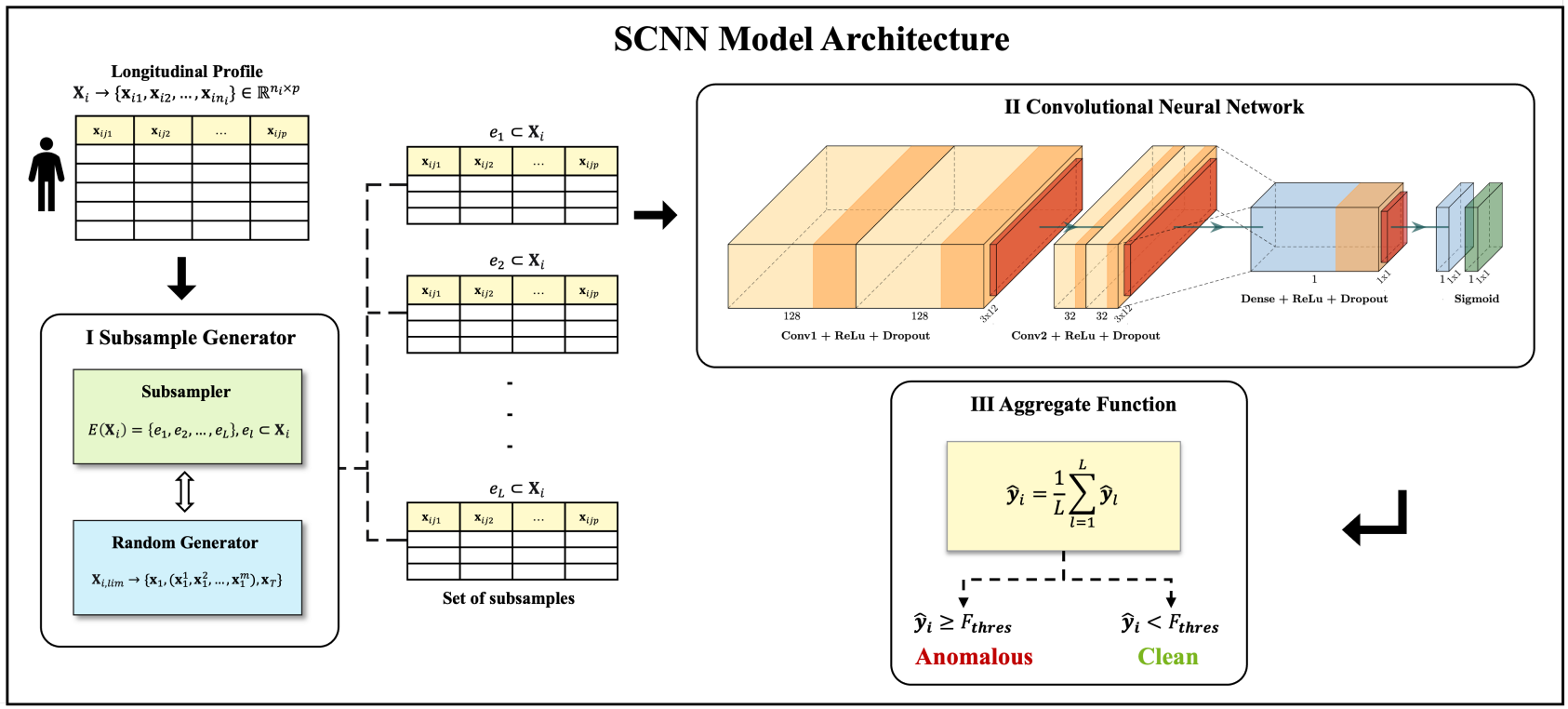


Fig. 4.2 Model architecture of the Subsampling-based Convolutional Neural Network (SCNN). The model consists of three main components: i) a subsample generator that creates multiple subsamples from each longitudinal profile, ii) a convolutional neural network that processes these subsamples through stacked convolutional and dense layers with dropout and ReLU activations, and iii) an aggregate function that averages predictions across subsamples to produce the final classification of the profile as clean or anomalous.

limit is used for the generation of additional samples that could effectively represent the same sample.

Next, the subsampler converts the longitudinal profile into a set of subsamples denoted by $E(\mathbf{X}_i) = \{e_1, e_2, ..., e_L\},\ e_l \subset \mathbf{X}_i$ where:

$$e_l = \{\mathbf{x}_{ij}, \mathbf{x}_{ik}, \mathbf{x}_T\}, \quad j,k \in \{1,2,\ldots,n_i\},\ j \neq k \tag{4.4}$$

The length of a subsample is set to 3, i.e., each subsample is a $(3 \times p)$ array ($2D$ data) consisting of 3 samples. Subsamples are generated corresponding to all the possible combinations of the samples, with their temporal order taken into account. The last sample of each subsample is always $\mathbf{x}_T$, which is compared with the other two previous samples. The number of generated subsamples $L$ can be calculated by:

$$L(n_i, n_{\mathbf{x}_T}) = \left(\frac{n_i!}{3! \cdot (n_i - 3)!}\right)^{n_{\mathbf{x}_T}} \tag{4.5}$$

where $n_i$ represents the number of samples in the longitudinal profile for subject $i$, and $n_{\mathbf{x}_T}$ represents the number of testing samples under consideration for the fraudulent trade.

### 4.4.2 Convolutional Neural Network

The generated subsamples are used as input to the convolutional neural network. The architecture of the convolutional network consists of two 2D convolutional units and two dense units. The first convolutional unit applies filters of size $(3 \times 1)$ on the input subsample with the ReLU activation function. Filter size $(3 \times 1)$ is used to compute the implicit differential information of the three samples across each parameter separately. Similarly, the second unit filters are applied to the output of the first unit. The dropout layer is added to avoid overfitting. The output embedding map is then flattened, and batch normalization is applied before passing it to the dense units. The final output is obtained from the second dense unit with one neuron and sigmoid activation function.

The output of the convolutional network $\hat{\mathbf{y}}_l$, tells whether the $\mathbf{x}_T$ is similar to the other two samples within the subsample $e_l$. A probability threshold $P_{\text{thres}}$ is defined, which can be set to achieve the desired specificity level. The binary cross-entropy loss function is used to train the convolutional network, as shown by the following equation:

$$\mathscr{L}_{\text{BCE}} = -\frac{1}{L}\sum_{l=1}^{L}\left(\mathbf{y}_l \cdot \log \hat{\mathbf{y}}_l + (1 - \mathbf{y}_l) \cdot \log\left(1 - \hat{\mathbf{y}}_l\right)\right) \tag{4.6}$$

where $\hat{\mathbf{y}}_l$ represents the predicted label and $\mathbf{y}_l$ denotes the actual label of the subsample $e_l$.

### 4.4.3 Aggregate Function

The output of the convolutional network for all the subsamples of the profile is fed into the aggregate function to determine the final prediction of whether the testing sample belongs to the same longitudinal profile of the athlete or not. The hard voting criterion is used, i.e., the voting is based on the output class predicted by the convolutional network for each subsample.

Let us consider the prediction class for each subsample to be $\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \ldots, \hat{\mathbf{y}}_L$. Then the overall prediction on the longitudinal profile by the model can be calculated by:

$$\hat{\mathbf{y}}_i = \begin{cases} 1, & \frac{1}{L}\sum_{l=1}^{L} \hat{\mathbf{y}}_l \geq F_{\text{thres}} \\ 0, & \text{else} \end{cases} \tag{4.7}$$

where the value of $F_{\text{thres}}$ can be set to achieve the desired specificity level.

## 4.5 Experiments

### 4.5.1 Datasets

WADA and international sports federations maintain historical data for each athlete through doping tests conducted at various national and international athletic events. These records are managed in the Anti-Doping Administration  Management System (ADAMS) database, as described in the International Standard for Testing and Investigations [318]. For this study, the data was extracted from the ADAMS database, which contains longitudinal profiles of different athletes (represented by anonymized unique IDs) tested between 1 September 2018 and 31 March 2021.

Each longitudinal profile consists of a collection of steroid samples that include both primary and derived steroid parameters representing the concentration values of different steroid metabolites. The six primary parameters are androsterone (A), etiocholanolone (Etio), epitestosterone (E), testosterone (T), 5$\alpha$-androstanediol (5$\alpha$Adiol), and 5$\beta$-androstanediol (5$\beta$Adiol), and their five ratios (T/E, A/Etio, A/T, 5$\alpha$Adiol/5$\beta$Adiol, and 5$\alpha$Adiol/E) have a direct first-order dependence on the primary parameters, as described in TD2021EAAS [316]. From this extracted data, two different types of datasets are defined, i.e., training and testing datasets, by stratifying based on gender and the number of samples within each profile.

**Training Dataset** The Steroid-All dataset is used for training the model. It consists of 254,478 urine samples from 65,039 athletes, with each athlete having 2-20 steroid samples in their longitudinal profile [234]. Table 4.1 summarizes the distribution of the samples across male and female athletes.

Table 4.1 Data statistics of longitudinal profiles used for training the model.

| **Athlete** | **Profiles** | **Samples** |
|---|---|---|
| Male | 52,152 | 166,237 |
| Female | 12,887 | 88,241 |
| **Total** | **65,039** | **254,478** |

**Testing Dataset** Four different datasets are used to assess the performance of the model as listed in Table 4.2. Fig. 4.3 shows the distribution of the number of steroid samples in the longitudinal profiles of athletes for the training (Steroid-All) and testing (Steroid-M and Steroid-F) datasets. A peak at 2 is observed for the Steroid-All dataset indicating that the majority of athletes have only two samples in their profiles. This suggests that most athletes are young, early in their sports careers, and have undergone fewer doping tests.

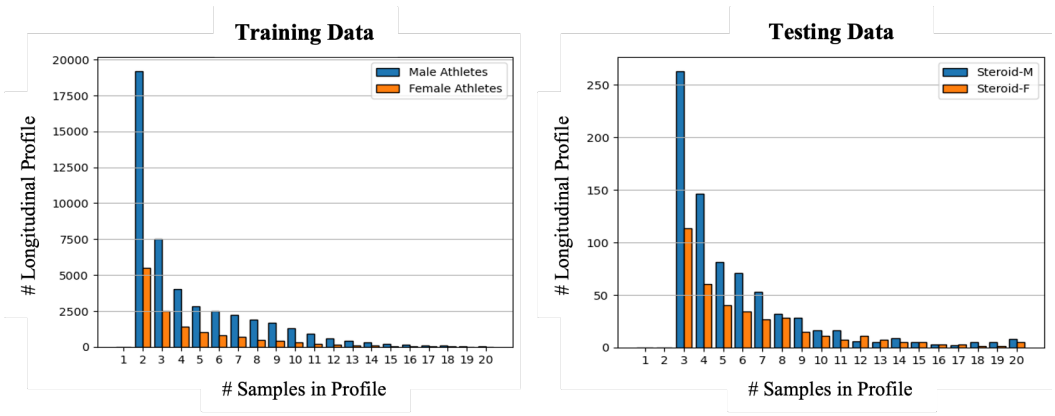


Fig. 4.3 Sample distribution per athlete profile in training and testing datasets. The left plot shows the number of longitudinal profiles by sample count for male and female athletes in the training dataset. The right plot presents the same distribution for the testing dataset, comparing the Steroid-M and Steroid-F cohorts. Most profiles have between 2 and 5 samples, reflecting the imbalance in real-world data.

Table 4.2 Description of testing datasets used for model evaluation. The table outlines four datasets consisting of male and female athlete profiles.

| Datasets | Description |
|---|---|
| Steroid-M | This dataset contains 4,214 steroid samples corresponding to 755 steroid profiles from male athletes, where each profile consists of at least 3 samples. In this dataset, around 20% of the profiles are manipulated, where the last sample in the profile is replaced by a different sample from another athlete. |
| Steroid-F | Similar to the Steroid-M dataset, this dataset contains 2,307 steroid samples corresponding to 375 steroid profiles of female athletes, where each profile consists of at least 3 samples. |
| Steroid-$M_{lim}$ | This dataset represents the case of limited samples in the profile ($\mathbf{X}_{i,lim}$), where one sample has to be compared against the other to determine the similarity. It contains 1,474 steroid samples corresponding to 737 steroid profiles of male athletes, where each profile consists of only 2 samples. |
| Steroid-$F_{lim}$ | Similar to Steroid-$M_{lim}$, this dataset also represents the case of limited samples for female athletes. It contains 586 steroid samples corresponding to 293 steroid profiles, where each profile consists of only 2 samples. |

### 4.5.2 Baseline Methods

The following SoTA models were employed to compare the performance of the proposed model.

- **Logistic Regression (LR)** [219]: A sigmoid function is used to calculate the probabilities of different classes.
- **Support Vector Machine (SVM)** [52]: Steroid samples are labeled into different classes by identifying the hyperplane that maximizes the margin between the two classes; the radial basis function is used as the kernel.
- **Extra Trees (ET)** [91]: A classifier is trained with 150 randomized decision trees on various subsamples of the longitudinal profiles for classification tasks.
- **Random Forest (RF)** [29]: A classifier is trained with 100 decision trees on the longitudinal profiles for classification.
- **Gradient Boosting (XGB)** [41]: A classifier is trained with a subsample rate of 0.8, using individual profiles as input.

- **Neural Network (NN)** [191]: A fully connected network with five layers is trained on individual profiles using a learning rate of 0.001 to minimize the binary cross-entropy loss function.
- **Bayesian Method (BM)** [270]: Calculates personalized thresholds for each parameter based on the prior distribution of the reference population. Compares a new sample against a critical range defined by a given specificity under normal conditions. Effective for detecting suspicious samples and considered state-of-the-art in anti-doping.
- **Digital Athlete Passport (DAP)** [234]: Uses principal component analysis and centroid concept to reduce correlated parameters to a smaller set of mutually independent components. Provides a complete visualization of the longitudinal profile in a three-dimensional space.

### 4.5.3 Experimental Settings

The Steroid-All dataset was used for the training and validation of the proposed SCNN model as well as all baseline models. In this dataset, longitudinal profiles were randomly selected, and in 50% of the cases, the last sample was swapped with a sample from a different athlete and labeled as swapped profiles (class 1). The remaining 50% of the profiles were considered clean and labeled accordingly (class 0). Each profile was normalized separately, i.e., all samples within a profile were normalized to unit norm. The dataset was randomly partitioned into 80% for training and 20% for validation.

**Hyperparameters** Hyperparameters are optimized on the training set. Table 4.3 shows the best hyperparameters found after optimizing the model using the Optuna [3]. In order to prevent the model from overfitting on the training data, in addition to using dropout, an early stopping criterion is used when the validation accuracy does not increase for 10 iterations.

## 4.6 Results

### 4.6.1 Performance Comparison

The performance of the SCNN model was compared with baseline models for detecting sample swapping across different testing datasets, as shown in Table 4.4. The evaluation was conducted using accuracy (AC), specificity (SP), sensitivity (SN), and area under the ROC curve (AU) as performance metrics. Due to domain-specific constraints, high-specificity

Table 4.3 Hyperparameter configuration for the SCNN model. The table shows the key architectural and training hyperparameters selected after model optimization.

| **Hyperparameter** | **Value** |
|---|---|
| Convolutional layers | 2 |
| Dense layers | 2 |
| 1st Conv (# filters) | 128 |
| 2nd Conv (# filters) | 32 |
| Filter size | $3\times1$ |
| Dropout | 0.2 |
| Optimizer | Adam |
| Learning rate | 1e-4 |
| # Epochs | 200 |

conditions are required to minimize false positive cases (cost factor). Therefore, the configuration of all models was adjusted to enable performance comparison at a specificity level of $99 \pm 2\%$.

Table 4.4 Evaluation results of SCNN and all the baseline models on different testing datasets. BM and DAP models cannot be evaluated on Steroid-$M_{lim}$ and Steroid-$F_{lim}$ datasets because these models need at least 3 steroid samples in the longitudinal profile.

| **Datasets** | **Metrics** | **LR** | **SVM** | **ET** | **RF** | **XGB** | **NN** | **BM** | **DAP** | **SCNN** |
|---|---|---|---|---|---|---|---|---|---|---|
| Steroid-M | AC | 0.754 | 0.849 | 0.804 | 0.816 | 0.895 | 0.857 | 0.760 | 0.810 | **0.915** |
| | SP | 1.000 | 0.988 | 0.986 | 0.994 | 0.991 | 0.980 | 0.920 | 0.978 | 0.978 |
| | SN | 0.000 | 0.387 | 0.210 | 0.258 | 0.618 | 0.457 | 0.730 | 0.750 | 0.721 |
| | AU | 0.500 | 0.779 | 0.706 | 0.722 | 0.871 | 0.811 | – | – | 0.915 |
| Steroid-F | AC | 0.779 | 0.832 | 0.797 | 0.797 | 0.889 | 0.850 | 0.850 | 0.869 | **0.901** |
| | SP | 1.000 | 0.993 | 1.000 | 1.000 | 0.999 | 1.000 | 0.850 | 0.890 | 0.960 |
| | SN | 0.000 | 0.265 | 0.000 | 0.084 | 0.482 | 0.360 | 0.380 | 0.610 | 0.617 |
| | AU | 0.500 | 0.766 | 0.566 | 0.534 | 0.841 | 0.734 | – | – | 0.901 |
| Steroid-$M_{lim}$ | AC | 0.723 | 0.863 | 0.737 | 0.745 | 0.871 | 0.798 | – | – | **0.871** |
| | SP | 1.000 | 0.992 | 1.000 | 1.000 | 0.991 | 0.996 | – | – | 0.959 |
| | SN | 0.000 | 0.589 | 0.000 | 0.042 | 0.625 | 0.367 | – | – | 0.532 |
| | AU | 0.500 | 0.758 | 0.524 | 0.539 | 0.774 | 0.637 | – | – | 0.871 |
| Steroid-$F_{lim}$ | AC | 0.700 | 0.823 | 0.703 | 0.710 | 0.829 | 0.700 | – | – | **0.813** |
| | SP | 1.000 | 0.992 | 1.000 | 1.000 | 0.994 | 0.999 | – | – | 0.956 |
| | SN | 0.000 | 0.443 | 0.011 | 0.034 | 0.455 | 0.000 | – | – | 0.514 |
| | AU | 0.500 | 0.714 | 0.500 | 0.517 | 0.722 | 0.500 | – | – | 0.754 |

Overall, LR could not detect any swapped profile correctly and showed poor performance. On a similar note, multiple decision tree-based models (i.e., ET and RF) also show sensitivity values $< 25\%$. This shows that the tree-based approach is not suitable for determining the

relatedness among the samples. Moreover, it can be observed that all the models (except DAP, BM, XGB) exhibit poor performance in terms of sensitivity. However, the accuracy of all the models is much better because the dataset is highly imbalanced. For the datasets consisting of only 2 samples in the longitudinal profile, only SVM and XGB achieved sensitivity values $> 50\%$ among all the baseline models.

The proposed SCNN model outperforms all the baseline models, including the current SoTA methods on all the datasets. It achieves a sensitivity of 72% and 62% on Steroid-M and Steroid-F datasets, respectively. It achieves an overall accuracy of $> 81\%$ for all the datasets. The model also performed well in the case of a limited sample dataset, where it achieved an accuracy of $> 81\%$. It is observed that this model shows better performance on the profiles of male athletes compared to female athletes. This is because the training dataset contains relatively more profiles for male athletes than female athletes.

### 4.6.2 Precision-Recall Curve

Fig. 4.4 shows the ROC and Precision-Recall (PRC) curve for SCNN against other baselines for the Steroid-M dataset at a normal setting. The requirement that BM and DAP models need at least 3 samples to perform analysis shows their inability to find fraudulent trade in limited cases datasets.

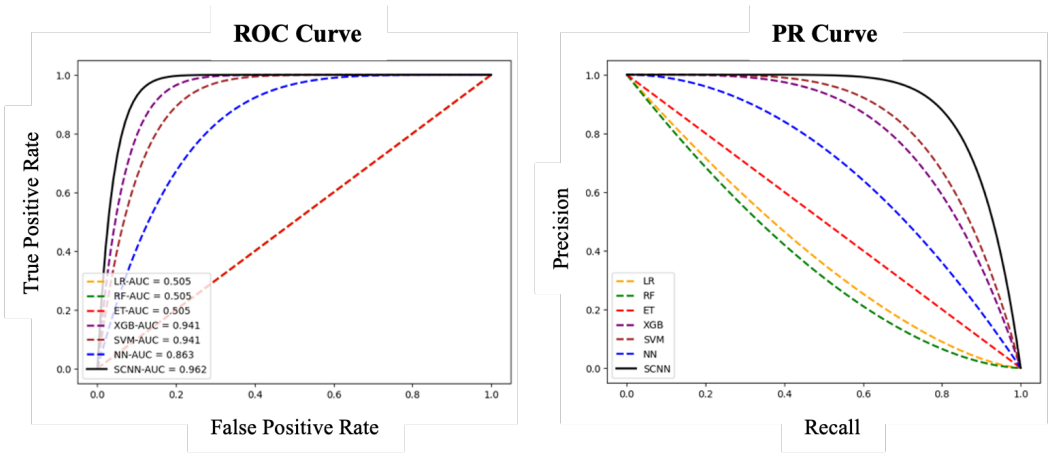


Fig. 4.4 ROC and Precision-Recall (PR) curves for SCNN and baseline models on the Steroid-M dataset under normal settings. SCNN achieves the highest AUC value, showing better discrimination and precision-recall performance compared to all the baseline methods. BM and DAP models are excluded from these plots as they do not produce continuous decision scores, being based on Bayesian and clustering techniques.

In addition, the dataset consists of the longitudinal profile of confirmed positive cases of steroid doping by the laboratory, i.e., use of prohibited substance. The model successfully

flagged these longitudinal profiles as suspicious. The limited number of confirmed sample swapping cases during this evaluation indicates the less prevalence nature of such cases in real-world sports. However, identifying sample swapping cases in the anti-doping analysis is a crucial and demanding task. The evaluations demonstrate that this method produces promising results with the potential to improve the current adaptive model of flagging sample swapping cases.

### 4.6.3 Parameter Sensitivity

The effect of different values of $F_{\text{thres}}$ and $P_{\text{thres}}$ on the performance of the SCNN model was also examined. Let us consider a 2D space spanned by $F_{\text{thres}} \times P_{\text{thres}}$, where $F_{\text{thres}}, P_{\text{thres}} \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$, and each point in the space represents a different setting of the model. The model was evaluated at every point in the space, and the sensitivity and specificity values were analyzed. $F_{\text{thres}} = 0.5$ and $P_{\text{thres}} = 0.5$ represent the normal setting, and $F_{\text{thres}} = 0.7$ and $P_{\text{thres}} = 0.8$ represent the high specificity setting, as shown in Fig. 4.5.

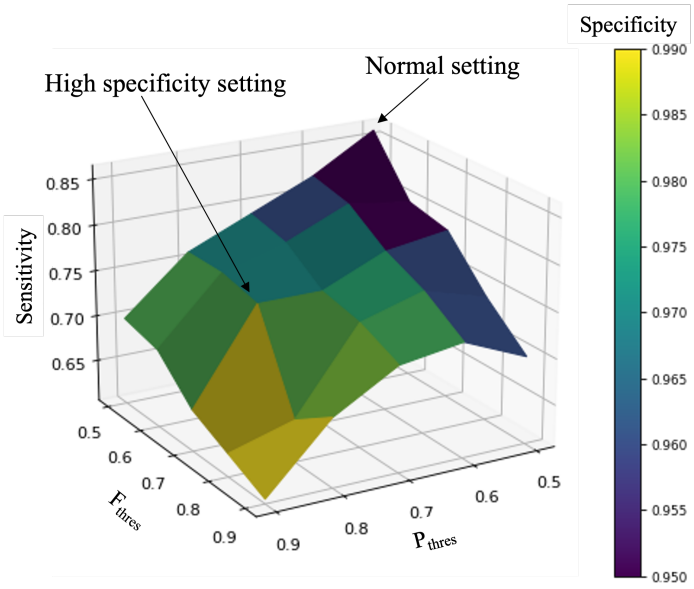


Fig. 4.5 The plot showing the performance of SCNN model under varying threshold settings. It shows the sensitivity of the model along the vertical axis, with specificity encoded via the color map. The plot highlights the trade-off between sensitivity and specificity, with the normal and high-specificity settings annotated.

### 4.6.4 Ablation Studies

To evaluate the effect of subsample length on model performance, an ablation study was conducted in which the proposed model was tested using subsamples of varying lengths. Performance was measured in terms of sensitivity under high-specificity conditions. As shown in Fig. 4.6, the results demonstrate how different subsample lengths affect the model's ability to effectively capture relevant patterns when all possible combinations of steroid samples from the longitudinal profile are considered in defining the subsamples.

The results show that 3 is an optimum subsample length for the model, which demonstrates that the model is capable of capturing well the likelihood of relatedness among the samples when the testing sample is compared with two other randomly selected samples. Choosing a larger length for the subsample can cause overfitting of the model, and therefore, the performance drops drastically.

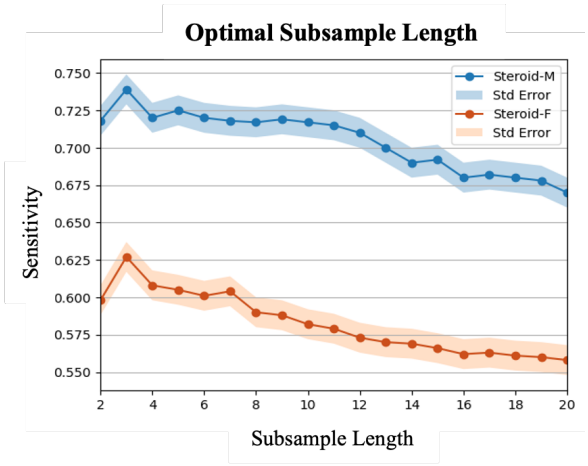


Fig. 4.6 Sensitivity of the SCNN model on the testing dataset as a function of subsample length. The plot compares performance on Steroid-M and Steroid-F datasets, showing that shorter subsample lengths yield higher sensitivity. Steroid-M consistently achieves better sensitivity across all lengths, with optimal performance around lengths 3-5. Shaded regions indicate standard error.

## 4.7 Case Study

A case study was performed in which the model was tested on longitudinal profiles provided by the laboratory. These profiles consists of two real-world sample swapping cases confirmed by DNA analysis. The term "case" refers to the longitudinal profiles of real-world athletes who were identified as having been involved in sample swapping. In these cases, samples that

were purported to belong to a particular athlete were proven by DNA analysis to belong to someone else. These two longitudinal profiles included multiple suspicious samples (testing samples), where none of them were from the same athlete. Both profiles were successfully triggered by the SCNN model and identified as sample swapping cases, as shown in Table 4.5.

Table 4.5 Performance of different models on DNA-verified longitudinal profiles. The table shows the percentage of profiles correctly identified in three case categories: sample swapping, steroid doping, and clean profiles.

| **Cases** | **Profiles** | **LR** | **SVM** | **ET** | **RF** | **XGB** | **NN** | **BM** | **DAP** | **SCNN** |
|---|---|---|---|---|---|---|---|---|---|---|
| Sample Swapping | 2 | 0.00 | 0.00 | 0.00 | 0.00 | 0.50 | 0.50 | 1.00 | 1.00 | **1.00** |
| Steroid Doping | 5 | 0.00 | 0.20 | 0.20 | 0.60 | 0.20 | 0.80 | 1.00 | 1.00 | **1.00** |
| Clean Profiles | 23 | 0.17 | 0.21 | 0.33 | 0.21 | 0.43 | 0.54 | 0.78 | 0.82 | **0.91** |

## 4.8 Summary

This chapter presented the Subsampling-based Convolutional Neural Network (SCNN), a deep network for anomaly detection in heterogeneous longitudinal clinical data, specifically targeting sample swapping in anti-doping analysis. The model addresses one of the most important challenges in real-world longitudinal data analysis: the limited number of samples available for each individual. This is particularly necessary in anti-doping contexts where athletes, especially younger or less frequently tested ones, may only have two recorded samples over long time periods. Like SACNN, SCNN addresses the four challenges highlighted in RQ1, i.e., irregular sampling, heterogeneous profile lengths, limited numbers of samples per athlete, and the scarcity of labeled anomaly data, but it approaches them through a fundamentally different strategy based on subsampling and aggregation.

To overcome irregular sampling intervals and heterogeneous profile lengths, SCNN decomposes each athlete's trajectory into multiple temporally ordered subsamples or triplets, creating consistent and comparable inputs regardless of profile length or spacing. This subsampling strategy also mitigates the problem of limited samples per athlete, since even profiles with only two or three entries can generate informative subsample units. The convolutional encoders then extract implicit differential features across these subsamples, capturing relational consistency rather than absolute values. Finally, SCNN reduces reliance on ground-truth anomaly labels by aggregating predictions across all subsamples, producing robust profile-level anomaly scores without requiring direct supervision from annotated doping cases. Through this design, SCNN is able to identify irregularities such as sample

manipulation and profile inconsistency by modeling deviations in intra-individual temporal consistency.

Empirical evaluations on both synthetic and real-world datasets, including DNA-confirmed cases of sample manipulation, demonstrated that SCNN consistently outperforms SoTA baseline models and domain-specific approaches. The SCNN achieved sensitivity values above 70% at a high specificity threshold of 99%, which is a requirement in anti-doping applications where false positives lead to costly follow-up testing and reputational risks. The model's capacity to detect anomalies even in profiles with only two samples shows a significant advancement over current methods, most of which are inapplicable under such conditions. These findings underscore the robustness and adaptability of SCNN across diverse testing conditions, athlete populations, and data sparsity regimes.

In addition to empirical success, SCNN makes a broader methodological contribution by introducing a data-efficient strategy for anomaly detection in irregular longitudinal datasets. By focusing on subsample-level consistency and aggregation, SCNN demonstrates that robust anomaly detection is possible even under extreme data limitations. This makes SCNN particularly valuable in anti-doping contexts, where many younger athletes have limited longitudinal histories. Beyond sports, the approach offers general applicability to other domains of longitudinal clinical monitoring where data sparsity and irregularity are inherent, such as rare disease tracking or personalized medicine.

# Section III: Incorporation of Metabolism Pathway Structure

# Chapter 5

# STT-LLM: Structural-Temporal Tokenization for Large Language Models

## 5.1 Introduction

[1]Large Language Models (LLMs) have demonstrated impressive generalization abilities across diverse natural language and multimodal tasks, including question answering, reasoning, and code generation [37, 190, 300]. This has led to growing interest in adapting LLMs to scientific domains, particularly those with limited supervision and complex data structures. However, most LLMs are pretrained on unstructured textual corpora and rely on discrete token sequences, making them not well-suited for directly modeling numerical, structured, and temporal data [235, 207]. Longitudinal clinical data, such as hormone levels or biomarker trajectories, are inherently multivariate, irregularly sampled, and often linked through biological pathways or physiological graphs [174]. These structured dependencies, such as enzymatic relationships in steroid metabolism, are important for anomaly detection in anti-doping, since deviations are often meaningful only when considered in relation to other biomarkers. Ignoring these dependencies risks detecting statistical outliers that lack biological plausibility or missing coordinated manipulations that occur across multiple markers. Therefore, bridging this gap between natural language-oriented LLMs and longitudinal clinical profiles remains a fundamental challenge, directly motivating RQ2.

*Challenge 1: Longitudinal clinical data contain structured dependencies that LLMs are not natively equipped to model.* Standard LLMs operate on sequences of discrete tokens optimized for text [207, 133]. In contrast, longitudinal clinical profiles consist of

[1]**Based on Publication:** Rahman, M.R., Hammouda, M., Maass, W. (2025). Structural-Temporal Tokenization for Resource-Efficient Language Models on Longitudinal Clinical Profile. *In Proceedings of the Neural Information Processing Systems (NeurIPS 2025), Main Track*. (under review)

multivariate, continuous signals recorded over irregular time intervals, often augmented with domain-specific structure such as metabolic pathway graphs [260]. In anti-doping, for example, testosterone and epitestosterone levels are tightly coupled through shared metabolic pathways, and deviations are only meaningful when modeled jointly. Prior work in graph-based modeling [92, 75, 182] and time-series transformers [285, 109, 328] addresses such complexities using specialized architectures, but these remain incompatible with generic LLM inference pipelines and fail to embed structured priors effectively [153, 51]. As a result, LLMs show poor alignment when applied to anomaly detection and forecasting tasks in high-dimensional clinical trajectories.

*Challenge 2: Resource and privacy constraints prevent large-scale LLM deployment in clinical and decentralized settings.* LLMs are typically hosted on cloud infrastructure, raising significant *privacy* and *regulatory concerns* [54]. Moreover, their large parameter counts and memory footprints limit deployment on edge devices, where real-time inference is desirable. While parameter-efficient fine-tuning strategies [108, 338] have reduced compute costs, they do not resolve the issue of token mismatch in domain adaptation [248], which requires dedicated preprocessing and embedding alignment. Further, while techniques like federated learning [339] address decentralized training, they do not solve the core issue of aligning LLM inference with structured, pathway-informed clinical inputs.

To address these challenges, this chapter introduces the Structural-Temporal Tokenization framework (STT-LLM), designed to incorporate biological pathway structure into anomaly detection for longitudinal clinical profiles. Unlike conventional LLM approaches that treat clinical measurements as flat sequences or tabular inputs [244], STT-LLM integrates a graph formulation to model structural dependencies (e.g., metabolic pathways) and applies self-attention mechanisms to capture temporal evolution. These components are combined into a unified structural-temporal embedding. To bridge the embedding space with LLM input expectations, specialized structural and temporal tokenizers are designed to transform the joint embeddings into token sequences compatible with LLMs. The resulting tokenized representations are then processed by a pre-trained LLM equipped with low-rank adaptation for efficient task-specific learning. This design allows STT-LLM to support different clinical downstream tasks under resource constraints, enabling low-latency, privacy-preserving inference on local hardware. To evaluate the framework, STT-LLM is applied to a domain-specific benchmark consisting of steroid profiles from real-world athletes in sports. Two main tasks are examined: i) sequence forecasting and ii) anomaly detection under both zero-shot and few-shot settings. Comparisons against open-source LLMs (LLaMA-2/3 [287, 99], Mistral [134], Phi-4 [1], Falcon [8], etc.) show that STT-LLM outperforms all baselines by significant margins. In the low-shot setting, STT-LLM achieves over 10% improvement in error reduction and

higher sensitivity in anomaly detection, demonstrating strong generalization for irregular time-series. These results highlight that integrating structural priors, such as metabolic pathway constraints, into LLM-based frameworks is important for biologically valid and operationally useful anomaly detection in anti-doping and beyond. The key contributions of this chapter are:

- STT-LLM is proposed as a unified structural-temporal embedding framework by incorporating domain-specific pathway structure into anomaly detection while modeling temporal dynamics in longitudinal profiles.
- The model consists of specialized tokenizers that transform structural-temporal embeddings into token sequences compatible with LLMs, enabling seamless integration with general-purpose language models.
- Experimental results show strong performance on clinical sequence prediction and anomaly detection tasks, including a case study on DNA-verified steroid profiles where STT-LLM outperforms baseline LLMs while remaining resource-efficient.

## 5.2 Related Work

**Tokenization and Embedding for Domain Adaptation**

Tokenization plays a foundational role in aligning raw inputs with the internal representations of LLMs, yet it remains a relatively underexplored area in domain adaptation compared to pretraining and fine-tuning strategies. Classical methods such as byte-pair encoding [255] and WordPiece [323] are effective for natural language but poorly suited for specialized domains where tokens may have domain-specific semantics or structural meaning. Recent efforts have explored task-aware token selection for domain generalization and efficiency [126, 170]. TAPEX [171], TabLLM [117], and TABBIE [129] adapt LLMs to tabular inputs through specialized token formats and training objectives. In graph-based domains, GraphPrompt [277] and Graph-of-Thought [22] integrate graph embeddings via soft prompts or fusion modules, enabling zero-shot reasoning over relational data. More broadly, vocabulary refinement approaches have focused on low-resource adaptation [349, 97], cross-lingual transfer [194, 166], and code modeling [45, 53], often requiring new subword vocabularies or embedding reinitialization. Embedding strategies range from mean-pooling and distance-based transfer [173] to hypernetwork-based token generation [80], but typically rely on auxiliary models or task-specific heuristics. While these methods promise, they often require retraining large models or domain-specific infrastructure.

**LLMs for Longitudinal Clinical Modeling**

Recent work has explored repurposing LLMs for general time-series tasks through prompt augmentation and embedding reprogramming strategies [233]. For example, models such as Time-LLM [136] and UniTime [172] demonstrate that pretrained LLMs can be reprogrammed to model time-indexed data by projecting temporal patches into token sequences. Despite these advances, most frameworks treat time-series data as flat or fully textified inputs, neglecting the temporal granularity and variable semantics critical in clinical monitoring. Moreover, forecasting from clinical narratives has been explored through timeline extraction [85] and event ordering [157], but often relies on fixed annotation spans and lacks fine-grained temporal resolution. Structured longitudinal clinical data models traditionally rely on physiological scores or structured features (e.g., SOFA, SAPS) [124, 212], whereas more recent models aim to build patient-specific representations from narrative texts and structured labs [132, 21]. However, the gap between LLMs and domain-specific data distributions remains a key challenge, particularly under zero- or few-shot settings.

## 5.3 Preliminaries

Let us consider longitudinal clinical profile consisting of repeated measurements of multiple parameters across time. Formally, the longitudinal clinical profile for a given individual can be represented as $\mathbf{X}_i = [\mathbf{x}_{ij}] \in \mathbb{R}^{p \times n_i}$, where $p$ is the number of parameters, $n_i$ is the number of samples in profile $\mathbf{X}_i$, and $\mathbf{x}_{ijk}$ denotes the parameter $k$ of the $j$-th sample. The clinical data may also include structural information encoded as a feature interaction graph $A \in \mathbb{R}^{p \times p}$, where $A_{k,l}$ represents the relationships between parameter $k$ and $l$. The two primary clinical tasks are aimed to be addressed:

**Sequence Prediction** Given the observed sequence up to time $t$, denoted as $\mathbf{X}_{i,1:t} = [\mathbf{x}_{ij}]_{j=1,\ldots,t}$, the future values are aimed to be predicted for $t+1$ as $\hat{\mathbf{x}}_{i,t+1} = f_\theta(\mathbf{X}_{i,1:t}, A)$, where $f_\theta$ is a predictive function parameterized by $\theta$. The function $f_\theta$ models both temporal dependencies across time and structural dependencies among parameters.

**Anomaly Detection** Irregular patterns in the longitudinal clinical profile can be identified at two levels: i) *Local anomaly detection* to identify anomalous samples within an individual clinical profile, meaning that one or more samples $\mathbf{x}_{ij}$ may show abnormal behavior relative to the individual's own trajectory. Let us consider for each sample, a local anomaly score is computed $s_{ij}^{\text{local}} = g_\phi^{\text{local}}(\mathbf{x}_{ij}, \hat{\mathbf{x}}_{ij})$, where $\mathbf{x}_{ij} = [\mathbf{x}_{ij,1}, \ldots, \mathbf{x}_{ij,p}]$, $\hat{\mathbf{x}}_{ij} = [\hat{\mathbf{x}}_{ij,1}, \ldots, \hat{\mathbf{x}}_{ij,p}]$, and $g_\phi^{\text{local}}$ is a scoring function parameterized by $\phi$. One or more samples can be flagged as

locally anomalous if their scores exceed a predefined threshold $\mathscr{A}_{\text{local}} = \{j \mid s_{ij}^{\text{local}} > \varepsilon_{\text{local}}\}$. ii) *Global anomaly detection* to determine whether the entire clinical profile of an individual is anomalous. Specifically, an individual profile is considered globally anomalous if any sample (preferably the most recent observation in the clinical domain) is identified as anomalous. The global anomaly score is defined as $s_i^{\text{global}} = s_{i,n_i}^{\text{local}}$, where $n_i$ denotes the final sample index in the profile. A profile is classified as globally anomalous if $s_i^{\text{global}} > \varepsilon_{\text{global}}$.

## 5.4 STT-LLM: Structural-Temporal Tokenization for Large Language Models

STT-LLM is proposed which integrates joint structural-temporal embeddings, along with structural and temporal tokenizers, to effectively capture and represent the intricate structural and temporal relationships inherent in longitudinal clinical and similar datasets, as shown in (Fig. 5.1).

### 5.4.1 Input Prompt

The input prompt $\mathbf{I}$ consists of two components: the task $\mathbf{P}$, which is a textual description providing instructions, and the longitudinal clinical profile $\mathbf{X}_i$, which contains multivariate time-series measurements. The task prompt $\mathbf{P}$ is processed using a pre-trained language tokenizer to produce token embeddings $Z_{\text{Pre}}$, while the longitudinal clinical data $\mathbf{X}_i$ is fed into the proposed tokenization framework that integrates structural and temporal dependencies. This dual processing strategy enables the model to align semantic task instructions with rich domain-specific data representations.

### 5.4.2 Structural-Temporal Embeddings

**Structural Component** Given an adjacency matrix $A$ and a degree matrix $D$ of feature interaction graph, the normalized graph Laplacian $\mathscr{L} = I - D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ ($I$: identity matrix, $D$: node degrees) [140]. This normalized Laplacian $\mathscr{L}$ encodes important structural properties such as connectivity and community structure. The eigen-decomposition is calculated as $\mathscr{L} = U\lambda U^{-1}$ ($U$: eigenvectors, $\lambda$: eigenvalues). To obtain the structural embedding, the eigenvectors are projected through a learnable transformation: $\mathbf{E}_S = W_{\mathbf{E}_S}U + b_{\mathbf{E}_S}$ ($W_{\mathbf{E}_S}$, $b_{\mathbf{E}_S}$: trainable parameters).

**Temporal Component** The temporal behavior in the longitudinal clinical profile is modeled using an attention mechanism as:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5.1}$$

where $Q$, $K$, $V$ are linear projections [292]. To incorporate temporal order, positional encodings are added, defined as $PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d}}\right)$ and $PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d}}\right)$ ($pos$: position, $i$: dimension index). These encodings allow the model to distinguish between positions in the input sequence. The attention output $\mathbf{A}_T$, is passed through a feed-forward network to produce $Z_{ST} = \text{ReLU}(\mathbf{A}_T W_{\mathbf{E}_{T_1}} + b_{\mathbf{E}_{T_1}})W_{\mathbf{E}_{T_2}} + b_{\mathbf{E}_{T_2}}$ and layer normalization is applied to produce $\mathbf{E}_T = \text{LayerNorm}(Z_{ST})$. This architecture stabilizes training and facilitates gradient flow. The resulting temporal embeddings $\mathbf{E}_T$ capture dynamic patterns and dependencies important for modeling longitudinal clinical profiles. Finally, the structural and temporal embeddings are concatenated to form the unified structural-temporal embedding $\mathbf{E}(\mathbf{X}_i) = \mathbf{E}_S \,||\, \mathbf{E}_T \in \mathbb{R}^{(p+n_i)\times p}$. This joint embedding ensures comprehensive integration of structural and temporal information, preparing the longitudinal clinical data for tokenization and modeling.

### 5.4.3 Tokenization

**Structural Tokenizer** $(S)$ The framework processes the structural aspects of the structural-temporal embeddings by effectively encoding a parameter graph constructed from domain knowledge in longitudinal clinical profile. The input structural representation of longitudinal clinical profile $A$ is combined with the learned structural-temporal embedding $\mathbf{E}(\mathbf{X}_i)$, yielding the concatenated input:

$$\mathbf{X}_S = A||\mathbf{E}(\mathbf{X}_i), \quad \mathbf{X}_S \in \mathbb{R}^{(2p+n_i)\times p} \tag{5.2}$$

The combined input $\mathbf{X}_S$ is then processed through a multi-layer perceptron (MLP) with two layers. The first layer applies a ReLU nonlinearity $H_S = \text{ReLU}(\mathbf{X}_S W_{S_1} + b_{S_1}), \quad H_S \in \mathbb{R}^{(2p+n_i)\times d_{\text{hidden}}}$, followed by a linear transformation $Z_S^{MLP} = H_S W_{S_2} + b_{S_2}, \quad Z_S^{MLP} \in \mathbb{R}^{(2p+n_i)\times d_{LLM}}$, where $W_{S_1} \in \mathbb{R}^{p\times d_{\text{hidden}}}$, $W_{S_2} \in \mathbb{R}^{d_{\text{hidden}}\times d_{LLM}}$, $b_{S_1}$, $b_{S_2}$ are trainable parameters. To ensure stable training and consistent scaling of the token embeddings, layer normalization is applied $Z_S = \text{LayerNorm}(Z_S^{MLP}), \quad Z_S \in \mathbb{R}^{(2p+n_i)\times d_{LLM}}$, where $d_{LLM}$ is the target embedding dimension compatible with the downstream LLM. The resulting structural token embeddings $Z_S$ encode both the structural relationships captured by the graph and the dynamic patterns captured by the structural-temporal embeddings.

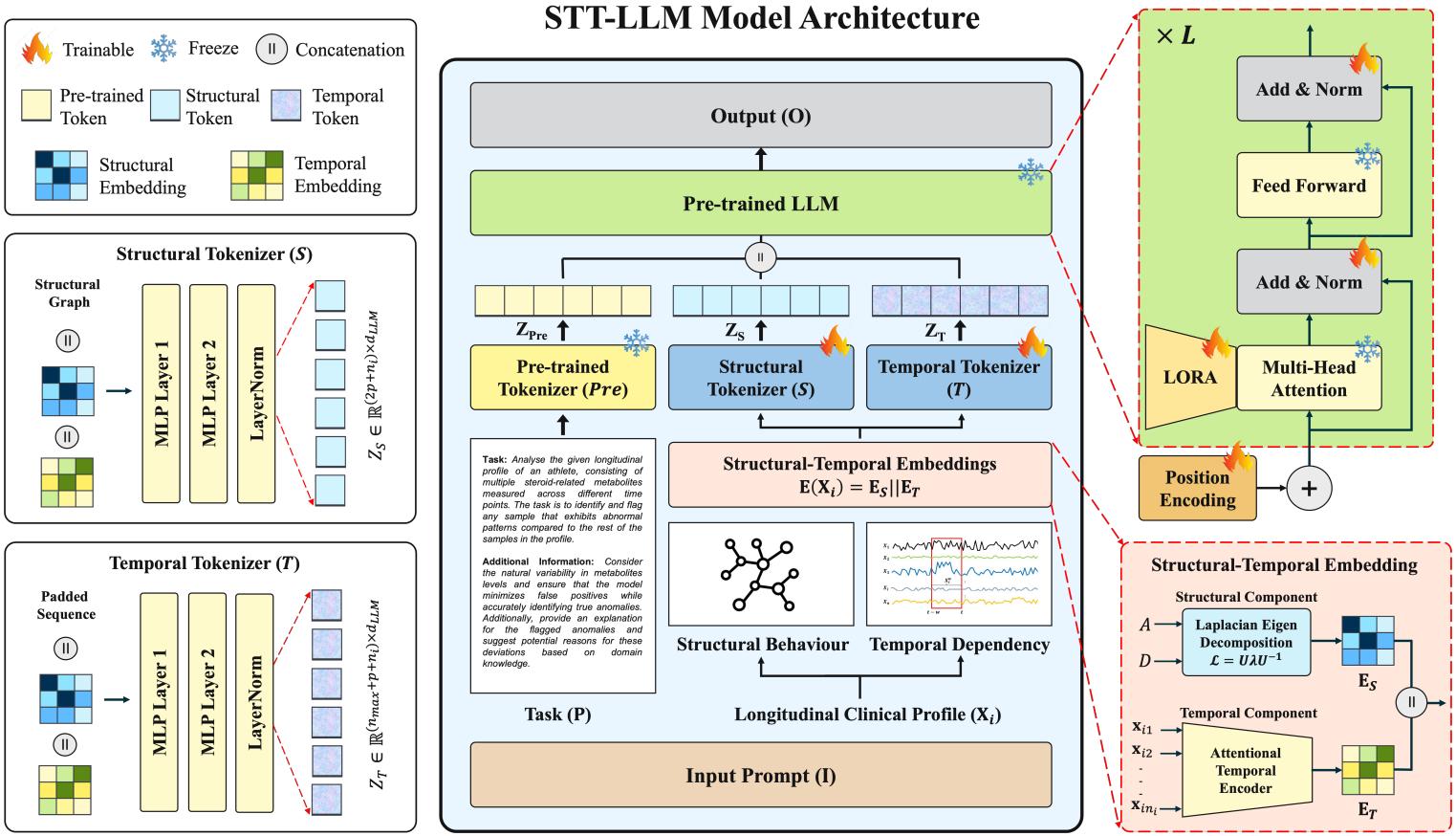


Fig. 5.1 Model architecture of the STT-LLM framework for longitudinal clinical data. The model integrates structural and temporal information from longitudinal clinical profiles using two dedicated tokenizers. The structural tokenizer ($S$) processes metabolism pathway information via structural graphs and MLP layers, while the temporal tokenizer ($T$) captures temporal dynamics using padded sequences and temporal embeddings. Their outputs ($Z_S$, $Z_T$) are concatenated with pre-trained token embeddings ($Z_{\text{Pre}}$) to form the final input embeddings to a frozen pre-trained LLM. The LLM is adapted using Low-Rank Adaptation (LoRA) at its attention layers for efficient fine-tuning.

**Temporal Tokenizer** $(T)$ To handle sequences of varying lengths (heterogeneity), the following steps are applied: i) *Padding:* Sequences shorter than $n_{\max}$ are zero-padded to ensure uniform input dimensions and ii) *Masking:* Mask $M \in \mathbb{R}^{n_{\max}}$ indicates valid time steps, with 1 marking real measurements and 0 marking padded positions. This ensures the model focuses computations on valid temporal entries. The padded temporal sequence $\mathbf{X}_T^{\text{padded}}$ is combined with the structural-temporal embedding $\mathbf{E}(\mathbf{X}_i)$:

$$\mathbf{X}_T = \mathbf{X}_T^{\text{padded}} || \mathbf{E}(\mathbf{X}_i), \quad \mathbf{X}_T \in \mathbb{R}^{(n_{\max}+p+n_i)\times p} \tag{5.3}$$

where $n_{\max}$ is the maximum sequence length of longitudinal clinical profile in the batch. The concatenated temporal input is passed through a two-layer MLP. The first layer applies a nonlinear transformation $H_T = \text{ReLU}(X_T W_{T_1} + b_{T_1}), \quad H_T \in \mathbb{R}^{(n_{\max}+p+n_i)\times d_{\text{hidden}}}$, followed by a second linear layer $Z_T^{\text{MLP}} = H_T W_{T_2} + b_{T_2}, \quad Z_T^{\text{MLP}} \in \mathbb{R}^{(n_{\max}+p+n_i)\times d_{LLM}}$. While the MLP processes the entire longitudinal profile, the mask $M$ ensures only valid time steps influence the learned embeddings. Finally, layer normalization is applied to stabilize learning and ensure consistent scaling $Z_T = \text{LayerNorm}(Z_T^{\text{MLP}}), \quad Z_T \in \mathbb{R}^{(n_{\max}+p+n_i)\times d_{LLM}}$. The resulting temporal token embeddings $Z_T$ effectively capture both the temporal evolution and structural context, which allows the model to handle variable-length sequences robustly.

### 5.4.4 Model Training

The output token embeddings $Z_{\text{Pre}}$, $Z_S$, and $Z_T$ are concatenated using $\mathbf{Z} = Z_{\text{Pre}} || Z_S || Z_T$, where $\mathbf{Z} \in \mathbb{R}^{L\times d_{LLM}}$, with $L$ denoting the token sequence length and $d_{LLM}$ the embedding dimension compatible with the LLM backbone. This combined representation is passed to a pre-trained LLM, which has been augmented with LoRA adapter $\mathbf{O} = \text{Adapter(LLM)}$, where $\mathbf{O}$ represents the model output for different downstream tasks, such as sequence prediction and anomaly detection over longitudinal clinical profiles. During training, the tokenizers ($S$, $T$) are trained jointly with the LoRA adapter, while the core LLM weights remain frozen. This setup allows efficient adaptation to specific downstream tasks with minimal computational overhead, leveraging the generalization capabilities of the pre-trained LLM while enabling domain-specific adaptation through the tokenizers and LoRA layers. The training objective functions can be defined according to the downstream task.

## 5.5 Experiments

STT-LLM is applied for doping analytics in sports, where detecting abnormal steroid patterns over time is important for identifying potential prohibited drug abuse by athletes.

### 5.5.1 Datasets

The models are evaluated on real-world athlete datasets consisting of longitudinal clinical profile of steroid data profiled from human urine samples: Steroid-M (male), Steroid-F (female), Steroid-$M_{lim}$ (male, limited), and Steroid-$F_{lim}$ (female, limited) [234, 232]. The dataset includes measurements of six key steroid metabolites: testosterone (T), epitestosterone (E), etiocholanolone (Etio), androsterone (A), 5$\alpha$-androstanediol (5$\alpha$Adiol), and 5$\beta$-androstanediol (5$\beta$Adiol) following the steroid metabolism pathway to synthesize [223]. The profile lengths range from 2 to 20 samples per athlete, reflecting realistic variability in longitudinal monitoring. These datasets cover diverse population groups and temporal resolutions, allowing us to comprehensively evaluate STT-LLM under realistic conditions. The detailed summary of the datasets are shown in Table 5.1.

Table 5.1 Data statistics used to evaluate the STT-LLM framework.

| **Datasets** | **Gender** | **# Profiles** | **# Samples** | $len(n)$ |
|---|---|---|---|---|
| Steroid-M | Male | 755 | 4214 | 3-20 |
| Steroid-F | Female | 375 | 2307 | 3-20 |
| Steroid-$M_{\text{lim}}$ | Male | 737 | 1474 | 2 |
| Steroid-$F_{\text{lim}}$ | Female | 293 | 586 | 2 |

### 5.5.2 Baseline Methods

This model is compared against different small LLMs that can be deployed on resource-efficient environment. The selected baselines include Qwen-2.5 (7B) [330], Falcon-3 (7B) [8], Mistral (7B) [134], LLaMA-2 (7B) [287], LLaMA-3.1 (8B) [99], Phi-4 (7B) [1], and DeepSeek-R1 (7B) [57]. Each model is fine-tuned on different downstream tasks using its native tokenization strategy. These models typically fall within the 7-8 billion parameter range, making them well-suited for efficient inference on local workstations without the need for large-scale GPU infrastructure.

### 5.5.3 Experimental Settings

All experiments were conducted on a workstation equipped with an NVIDIA TITAN RTX GPU (24GB), Intel i9 processor, and 31GB total RAM. The same computational setup was used for both STT-LLM and all baseline models to ensure fair and consistent comparisons. The evaluation was performed under two settings: *zero-shot*, and *few-shot* (2-20 labeled

examples as in-context prompts). The evaluation metrics used are RMSE, MAE, and MAPE for sequence prediction, and accuracy, sensitivity, precision, F1-score, and AUC for anomaly detection. A high specificity value (0.99) was set due to domain requirements. To account for variability, all reported results are averaged over three independent runs with standard deviations reported where applicable.

## 5.6 Results

### 5.6.1 Anomaly Detection

**Zero-shot setting** Table 5.2 and Fig. 5.2 shows that STT-LLM significantly outperforms baseline models in both local and global anomaly detection under zero-shot conditions. For local anomaly detection, STT-LLM achieves sensitivity of 15.0% on Steroid-M and 17.0% on Steroid-$F_{lim}$, while most baselines show near-zero sensitivity. This is because these models default to classifying all samples as normal, resulting in artificially inflated accuracy values around 95-96% but completely failing to identify any anomalous samples. In contrast, STT-LLM trades a small drop in accuracy (87-88%) for substantial gains in sensitivity and precision, reflecting its ability to detect true anomalies. In global anomaly detection, all models achieve better accuracy, as the classification task is inherently less sparse and the signal-to-noise ratio is higher. STT-LLM achieves the highest F1-scores (0.26 on Steroid-M, 0.29 on Steroid-F) and AUC values (0.57 on Steroid-M, 0.59 on Steroid-F), outperforming baselines by up to $\sim$10%. These results highlight STT-LLM's ability to handle both sparse (local) and dense (global) anomaly tasks, demonstrating increased robustness and generalization compared to standard LLMs, especially in anomaly detection scenarios where sensitivity is critical.

**Few-shot setting** Fig. 5.3 shows that STT-LLM achieves substantial gains in global anomaly detection as the number of shots increases. Unlike the baselines, which often exhibit unstable or noisy trends across shot sizes, STT-LLM shows consistent improvements across most metrics. For sensitivity, STT-LLM increases from 0.15 (2 shots) to 0.6 (20 shots) on Steroid-M, representing more than a threefold improvement in detecting true anomalies. Precision improves steadily as well, reaching near-perfect levels on Steroid-F and Steroid-$F_{lim}$, indicating that the model sharply reduces false positives as supervision increases. F1-score trends further highlight the balanced gains of STT-LLM, with performance rising sharply between 2 and 20 shots, e.g., 0.15 to 0.7 (Steroid-M), demonstrating the model's ability to jointly improve sensitivity and precision. Overall, STT-LLM's performance curves

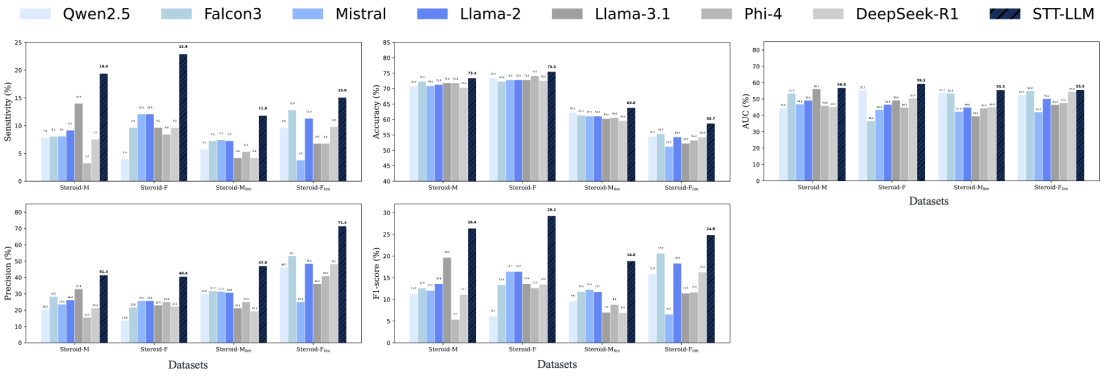


Fig. 5.2 Zero-shot global anomaly detection performance comparison across four datasets using different pre-trained LLMs. STT-LLM consistently outperforms all baseline models, particularly in sensitivity and F1-score, indicating its better capability to detect subtle anomalies in longitudinal profiles.

remain smooth, while baselines frequently show oscillating or deteriorating patterns as shots increase, reflecting their difficulty in integrating few-shot supervision effectively.

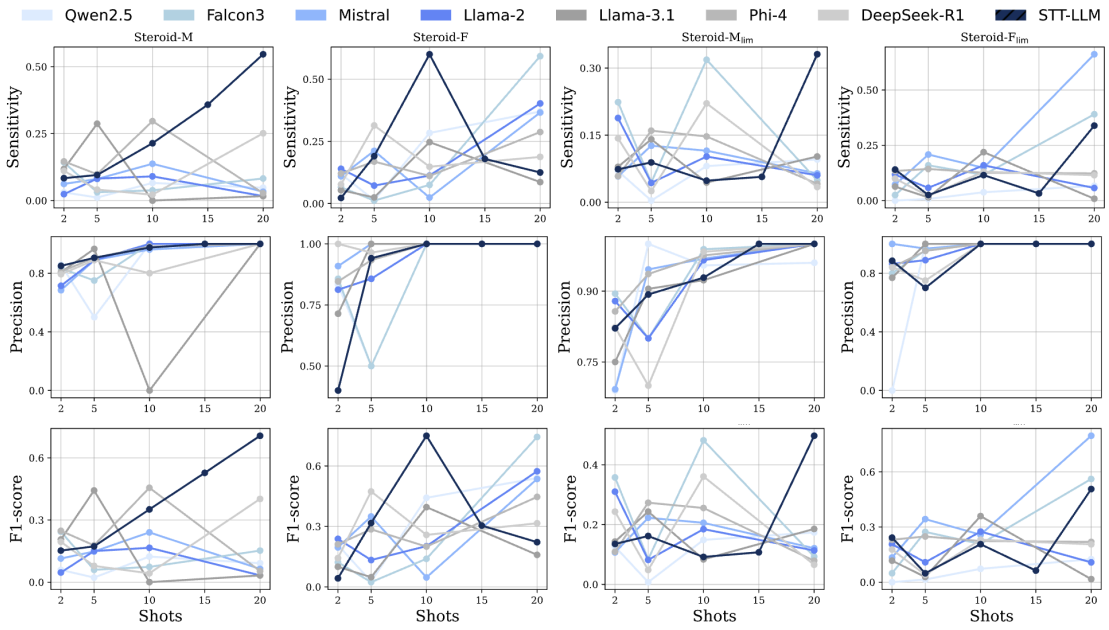


Fig. 5.3 Few-shot global anomaly detection performance across different metrics. Performance comparison of STT-LLM and baselines for global anomaly detection tasks on different datasets. Metrics are evaluated at different shot settings (2, 5, 10, 15, 20). STT-LLM consistently achieves higher F1-scores and improved sensitivity, indicating better anomaly detection capability from limited examples.

### 5.6.2 Sequence Prediction

**Zero-shot setting** Fig. 5.4 shows that STT-LLM consistently outperforms all LLM baselines by achieving the lowest error scores. For Steroid-M and Steroid-F, STT-LLM reduces RMSE value (%100) to 79.3 and 68.4, respectively, while all baselines remain above 83, indicating its improved ability to model metabolic patterns even without supervision. The gains are even more pronounced in the limited datasets, where STT-LLM achieves strikingly low RMSE value (%100) of 30.0 and 1.2, respectively, outperforming the next-best models by large margins. For MAE value (%10), STT-LLM consistently achieves the lowest errors across datasets, with values dropping to near 5-6 on the limited datasets, reflecting accurate point-wise predictions.

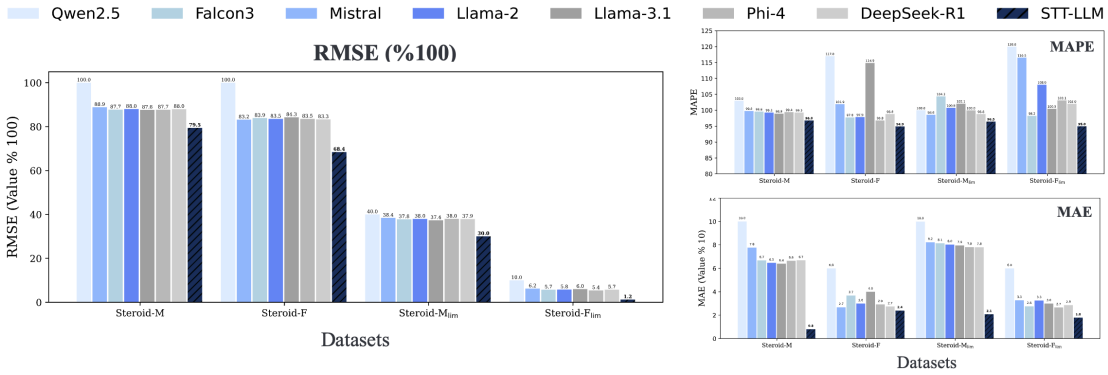


Fig. 5.4 Zero-shot sequence prediction performance of STT-LLM across different datasets. The figure compares STT-LLM with several open-source LLMs on four steroid datasets. STT-LLM consistently achieves the lowest error across all datasets and metrics, showing its robustness in modeling longitudinal steroid profiles in a zero-shot setting.

**Few-shot setting** Table 5.3 and Fig. 5.5 shows several important findings. Contrary to expectations, the error metrics increase as the number of shots increases from 5 to 20 across all the models. This indicates that simply increasing the number of in-context examples does not necessarily improve performance. The rise in error is likely due to including heterogeneous and potentially noisy profiles as prompts, which may confuse the model instead of guiding it, especially in a domain like longitudinal clinical monitoring, where inter-individual variation is high. Despite this, STT-LLM consistently achieves the best RMSE across all datasets and shot counts, demonstrating robust temporal generalization. For example, at 5-shot, STT-LLM achieves the lowest RMSE on Steroid-M$_{lim}$ (1730.11), Steroid-F$_{lim}$ (1276.32), and maintains higher performance across more shots as well. Similarly, in terms of MAE, this model outperforms baselines on Steroid-F$_{lim}$ with a score of 643.71 (10-shot) and 642.90 (20-shot). STT-LLM maintains high overall stability and minimal

fluctuation in MAPE compared to LLM baselines. These findings suggest that STT-LLM outperforms baselines consistently in absolute error terms and demonstrates better resilience to prompt variability and shot-induced drift.

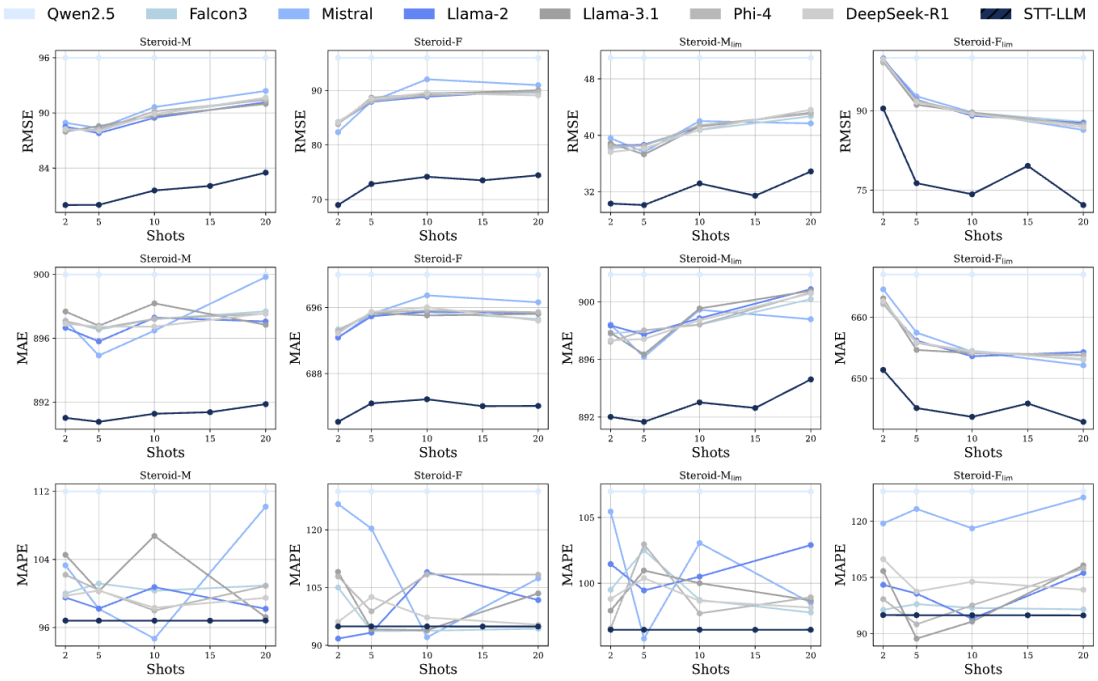


Fig. 5.5 Few-shot sequence prediction performance across multiple datasets. The figure shows the performance of STT-LLM and several baseline LLMs across four datasets. Performance is measured at varying shot levels (2, 5, 10, 15, 20). STT-LLM consistently outperforms the baselines, particularly at low-shot settings, indicating stronger generalization capability from limited examples.

Table 5.2 Local and global anomaly detection performance at the zero-shot setting for both local and global anomaly detection. STT-LLM consistently shows higher sensitivity and F1-score, particularly in local anomaly detection tasks, showing its strength in identifying subtle anomalous patterns without any prior exposure.

| **Datasets** | **Model** | **Local Anomaly** | | | | | **Global Anomaly** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Acc↑** | **Sens↑** | **Prec↑** | **F1↑** | **AUC↑** | **Acc↑** | **Sens↑** | **Prec↑** | **F1↑** | **AUC↑** |
| Steroid-M | Qwen-2.5 | **0.96±.01** | 0.00±.00 | 0.00±.00 | 0.00±.00 | 0.47±.02 | 0.71±.02 | 0.08±.03 | 0.20±.02 | 0.11±.02 | 0.45±.02 |
| | Mistral | 0.87±.02 | 0.05±.01 | 0.02±.01 | 0.03±.01 | 0.43±.02 | 0.71±.02 | 0.08±.02 | 0.23±.03 | 0.12±.02 | 0.47±.02 |
| | Falcon-3 | 0.94±.02 | 0.01±.00 | 0.02±.01 | 0.01±.01 | 0.46±.02 | 0.72±.02 | 0.08±.02 | 0.28±.03 | 0.13±.02 | 0.53±.02 |
| | LLaMA-2 | 0.90±.02 | 0.05±.01 | 0.03±.01 | 0.04±.01 | 0.42±.02 | 0.71±.02 | 0.09±.02 | 0.26±.02 | 0.14±.03 | 0.49±.02 |
| | LLaMA-3.1 | 0.87±.02 | 0.07±.01 | 0.03±.01 | 0.05±.01 | 0.51±.02 | 0.72±.02 | 0.14±.02 | 0.33±.03 | 0.19±.03 | 0.56±.02 |
| | Phi-4 | 0.87±.02 | 0.08±.01 | 0.04±.01 | 0.05±.01 | 0.50±.02 | 0.72±.02 | 0.03±.01 | 0.15±.02 | 0.05±.01 | 0.46±.02 |
| | DeepSeek-R1 | 0.95±.01 | 0.02±.01 | 0.01±.01 | 0.01±.00 | 0.39±.02 | 0.70±.02 | 0.08±.02 | 0.21±.02 | 0.11±.02 | 0.45±.02 |
| | **STT-LLM** | 0.87±.02 | **0.15±.02** | **0.07±.01** | **0.09±.02** | **0.57±.02** | **0.73±.02** | **0.19±.03** | **0.41±.03** | **0.26±.03** | **0.57±.02** |
| Steroid-F | Qwen-2.5 | 0.87±.02 | 0.04±.01 | 0.02±.01 | 0.02±.01 | 0.46±.02 | 0.73±.02 | 0.04±.01 | 0.14±.02 | 0.06±.01 | 0.55±.02 |
| | Mistral | **0.96±.01** | 0.00±.00 | 0.00±.00 | 0.00±.00 | 0.62±.02 | 0.73±.02 | 0.12±.02 | 0.26±.03 | 0.16±.02 | 0.43±.02 |
| | Falcon-3 | 0.95±.01 | 0.00±.00 | 0.00±.00 | 0.00±.00 | 0.60±.02 | 0.72±.02 | 0.09±.02 | 0.22±.02 | 0.13±.02 | 0.37±.02 |
| | LLaMA-2 | 0.87±.02 | 0.06±.01 | 0.02±.01 | 0.03±.01 | 0.55±.02 | 0.73±.02 | 0.12±.02 | 0.26±.02 | 0.16±.02 | 0.47±.02 |
| | LLaMA-3.1 | 0.95±.01 | 0.01±.00 | 0.03±.01 | 0.02±.01 | 0.57±.02 | 0.73±.02 | 0.10±.02 | 0.23±.03 | 0.14±.02 | 0.49±.02 |
| | Phi-4 | 0.88±.02 | 0.06±.01 | 0.02±.01 | 0.03±.01 | 0.50±.02 | 0.74±.02 | 0.08±.02 | 0.25±.02 | 0.13±.02 | 0.45±.02 |
| | DeepSeek-R1 | 0.87±.02 | 0.06±.01 | 0.02±.01 | 0.03±.01 | 0.42±.02 | 0.73±.02 | 0.10±.02 | 0.22±.03 | 0.13±.02 | 0.50±.02 |
| | **STT-LLM** | 0.87±.02 | **0.08±.01** | **0.03±.01** | **0.05±.01** | 0.47±.02 | **0.75±.02** | **0.23±.03** | **0.40±.03** | **0.29±.03** | **0.59±.02** |
| Steroid-M$_l im$ | Qwen-2.5 | 0.86±.02 | 0.00±.00 | 0.00±.00 | 0.00±.00 | 0.18±.01 | 0.62±.02 | 0.06±.02 | 0.30±.03 | 0.10±.02 | 0.54±.02 |
| | Mistral | **0.96±.01** | 0.00±.00 | 0.00±.00 | 0.00±.00 | 0.37±.02 | 0.61±.02 | 0.07±.02 | 0.31±.02 | 0.12±.02 | 0.42±.02 |
| | Falcon-3 | 0.88±.02 | 0.08±.01 | 0.03±.01 | 0.05±.01 | 0.44±.02 | 0.61±.02 | 0.07±.02 | 0.32±.02 | 0.12±.02 | 0.53±.02 |
| | LLaMA-2 | 0.88±.02 | 0.03±.01 | 0.01±.01 | 0.02±.01 | 0.22±.01 | 0.61±.02 | 0.07±.02 | 0.31±.02 | 0.12±.02 | 0.45±.02 |
| | LLaMA-3.1 | 0.88±.02 | 0.04±.01 | 0.02±.01 | 0.02±.01 | 0.39±.02 | 0.60±.02 | 0.04±.01 | 0.21±.02 | 0.07±.02 | 0.39±.02 |
| | Phi-4 | 0.89±.02 | 0.21±.02 | 0.09±.01 | 0.12±.02 | 0.65±.02 | 0.61±.02 | 0.05±.01 | 0.25±.02 | 0.09±.01 | 0.44±.02 |
| | DeepSeek-R1 | 0.87±.02 | 0.06±.01 | 0.02±.01 | 0.03±.01 | 0.43±.02 | 0.60±.02 | 0.04±.01 | 0.19±.02 | 0.07±.01 | 0.45±.02 |
| | **STT-LLM** | 0.88±.02 | **0.36±.02** | **0.12±.02** | **0.18±.02** | **0.75±.02** | **0.64±.02** | **0.12±.02** | **0.47±.03** | **0.19±.02** | **0.55±.02** |
| Steroid-F$_l im$ | Qwen-2.5 | 0.88±.02 | 0.06±.01 | 0.06±.01 | 0.06±.01 | 0.14±.01 | 0.54±.02 | 0.10±.02 | 0.46±.03 | 0.16±.02 | 0.53±.02 |
| | Mistral | 0.95±.01 | 0.01±.00 | 0.03±.00 | 0.02±.00 | **0.64±.02** | 0.51±.02 | 0.04±.01 | 0.25±.02 | 0.07±.01 | 0.42±.02 |
| | Falcon-3 | **0.96±.01** | 0.00±.00 | 0.00±.00 | 0.00±.00 | 0.27±.02 | 0.55±.02 | 0.13±.02 | 0.53±.03 | 0.21±.03 | 0.55±.02 |
| | LLaMA-2 | 0.86±.02 | 0.03±.01 | 0.01±.01 | 0.02±.01 | 0.32±.02 | 0.54±.02 | 0.11±.02 | 0.48±.03 | 0.18±.02 | 0.50±.02 |
| | LLaMA-3.1 | 0.87±.02 | 0.00±.00 | 0.00±.00 | 0.00±.00 | 0.08±.01 | 0.52±.02 | 0.07±.01 | 0.36±.02 | 0.11±.01 | 0.46±.02 |
| | Phi-4 | 0.87±.02 | 0.07±.01 | 0.03±.01 | 0.04±.01 | 0.48±.02 | 0.53±.02 | 0.07±.01 | 0.41±.03 | 0.12±.02 | 0.48±.02 |
| | DeepSeek-R1 | 0.86±.02 | 0.10±.01 | 0.04±.01 | 0.06±.01 | 0.51±.02 | 0.54±.02 | 0.10±.02 | 0.48±.03 | 0.16±.02 | 0.54±.02 |
| | **STT-LLM** | 0.87±.02 | **0.17±.02** | **0.08±.01** | **0.11±.02** | **0.54±.02** | **0.59±.02** | **0.15±.03** | **0.71±.03** | **0.25±.03** | **0.56±.02** |

Table 5.3 Few-shot sequence prediction performance of different models across four datasets under 5, 10, 15, and 20-shot settings. STT-LLM consistently achieves the lowest errors, showing its robustness and better predictive capabilities than baselines.

| **Datasets** | **Model** | **@5** | | | **@10** | | | **@15** | | | **@20** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **RMSE↓** | **MAE↓** | **MAPE↓** | **RMSE↓** | **MAE↓** | **MAPE↓** | **RMSE↓** | **MAE↓** | **MAPE↓** | **RMSE↓** | **MAE↓** | **MAPE↓** |
| Steroid-M | Qwen-2.5 | 1695.99 | 899.99 | 111.99 | 1695.99 | 899.99 | 111.99 | 1695.99 | 899.99 | 111.99 | 1695.99 | 899.99 | 111.99 |
| | Mistral | 1688.34 | 894.92 | 98.19 | 1690.63 | 896.48 | **94.69** | 1688.90 | 896.54 | 101.04 | 1692.39 | 899.84 | 110.19 |
| | Falcon-3 | 1688.02 | 896.54 | 101.17 | 1689.88 | 897.20 | 100.31 | 1690.39 | 897.48 | 100.01 | 1691.48 | 897.69 | 100.93 |
| | LLaMA-2 | 1687.80 | 895.81 | 98.21 | 1689.47 | 897.29 | 100.74 | 1690.01 | 896.86 | 100.47 | 1691.16 | 897.06 | 98.19 |
| | LLaMA-3.1 | 1688.57 | 896.78 | 100.27 | 1689.67 | 898.19 | 106.75 | 1690.56 | 897.17 | 101.46 | 1690.98 | 896.84 | 97.21 |
| | Phi-4 | 1688.20 | 896.62 | 100.38 | 1690.17 | 897.21 | 97.98 | 1690.04 | 897.36 | 102.87 | 1691.41 | 897.54 | 100.88 |
| | DeepSeek-R1 | 1688.05 | 896.73 | 100.33 | 1689.88 | 896.73 | 98.31 | 1690.31 | 897.17 | 98.81 | 1691.65 | 897.56 | 99.48 |
| | **STT-LLM** | **1680.00** | **890.77** | **96.80** | **1681.57** | **891.27** | 96.79 | **1682.06** | **891.37** | **96.79** | **1683.51** | **891.87** | **96.81** |
| Steroid-F | Qwen-2.5 | 1395.99 | 699.99 | 129.99 | 1395.99 | 699.99 | 129.99 | 1395.99 | 699.99 | 129.99 | 1395.99 | 699.99 | 129.99 |
| | Mistral | 1387.98 | 695.23 | 120.35 | 1392.05 | 697.48 | **92.08** | 1388.75 | 695.68 | 108.68 | 1390.98 | 696.63 | 107.37 |
| | Falcon-3 | 1388.12 | 695.07 | 93.71 | 1389.62 | 695.41 | 93.80 | 1388.77 | 695.67 | 115.24 | 1389.53 | 694.61 | **94.31** |
| | LLaMA-2 | 1387.93 | 694.93 | **93.30** | 1388.86 | 695.52 | 109.01 | 1388.92 | 694.91 | 100.61 | 1389.93 | 695.33 | 101.76 |
| | LLaMA-3.1 | 1388.67 | 695.34 | 94.12 | 1389.38 | 695.03 | 93.93 | 1388.72 | 695.19 | 106.55 | 1390.03 | 695.23 | 103.50 |
| | Phi-4 | 1388.07 | 695.39 | 98.83 | 1389.09 | 695.64 | 108.44 | 1389.50 | 694.70 | 99.72 | 1389.75 | 695.43 | 108.37 |
| | DeepSeek-R1 | 1388.48 | 695.47 | 102.55 | 1389.54 | 696.04 | 97.25 | 1389.05 | 695.14 | 98.93 | 1389.09 | 694.41 | 95.36 |
| | **STT-LLM** | **1372.85** | **684.39** | 94.94 | **1374.17** | **684.89** | 94.92 | **1373.51** | **684.05** | **94.91** | **1374.45** | **684.09** | 94.91 |
| Steroid-M$_l$*im* | Qwen-2.5 | 1750.99 | 901.99 | 106.99 | 1750.99 | 901.99 | 106.99 | 1750.99 | 901.99 | 106.99 | 1750.99 | 901.99 | 106.99 |
| | Mistral | 1737.63 | 896.17 | **95.80** | 1742.02 | 899.45 | 103.07 | 1738.92 | 898.42 | 103.42 | 1741.69 | 898.80 | 98.51 |
| | Falcon-3 | 1738.66 | 898.03 | 102.52 | 1740.75 | 898.42 | 98.74 | 1738.93 | 898.31 | 102.14 | 1742.69 | 900.19 | 97.78 |
| | LLaMA-2 | 1738.65 | 897.73 | 99.46 | 1741.24 | 898.86 | 100.51 | 1738.90 | 898.48 | 103.02 | 1743.15 | 900.89 | 102.91 |
| | LLaMA-3.1 | 1737.29 | 896.35 | 100.98 | 1741.25 | 899.54 | 100.01 | 1739.00 | 898.11 | 98.56 | 1743.21 | 900.77 | 98.70 |
| | Phi-4 | 1738.51 | 898.01 | 102.96 | 1741.42 | 898.43 | 97.71 | 1738.87 | 898.07 | 99.79 | 1743.05 | 900.66 | 98.94 |
| | DeepSeek-R1 | 1738.12 | 897.42 | 100.40 | 1740.81 | 898.72 | 98.67 | 1739.72 | 898.67 | 99.97 | 1743.62 | 900.59 | 98.15 |
| | **STT-LLM** | **1730.11** | **891.67** | 96.47 | **1733.18** | **893.01** | **96.47** | **1731.43** | **892.63** | **96.47** | **1734.87** | **894.61** | **96.47** |
| Steroid-F$_l$*im* | Qwen-2.5 | 1309.99 | 666.99 | 127.99 | 1309.99 | 666.99 | 127.99 | 1309.99 | 666.99 | 127.99 | 1309.99 | 666.99 | 127.99 |
| | Mistral | 1292.73 | 657.49 | 123.29 | 1289.67 | 654.38 | 118.12 | 1294.05 | 657.05 | 107.06 | 1286.36 | 652.15 | 126.36 |
| | Falcon-3 | 1291.65 | 655.82 | 97.87 | 1289.63 | 654.51 | 96.85 | 1294.77 | 656.69 | 102.03 | 1287.89 | 653.04 | 96.47 |
| | LLaMA-2 | 1292.06 | 656.18 | 100.69 | 1289.05 | 653.63 | 93.96 | 1295.08 | 656.97 | 101.13 | 1287.68 | 654.31 | 106.19 |
| | LLaMA-3.1 | 1291.13 | 654.67 | **88.66** | 1289.66 | 654.13 | **93.20** | 1293.98 | 656.87 | 115.68 | 1287.32 | 653.76 | 108.22 |
| | Phi-4 | 1291.92 | 655.84 | 92.49 | 1289.48 | 654.17 | 97.54 | 1294.68 | 656.39 | 102.23 | 1287.37 | 653.85 | 107.41 |
| | DeepSeek-R1 | 1291.64 | 655.94 | 101.25 | 1289.33 | 654.37 | 103.85 | 1294.89 | 656.59 | 97.21 | 1286.90 | 653.18 | 101.71 |
| | **STT-LLM** | **1276.32** | **645.16** | 94.92 | **1274.23** | **643.71** | 94.89 | **1279.59** | **645.90** | **94.89** | **1272.19** | **642.90** | **94.86** |

### 5.6.3 Ablation Studies

The ablations include removing all components (*w/o all*), structural tokenizer (*w/o structural*), temporal tokenizer (*w/o temporal*), embedding layer (*w/o embeddings*), and pairs of components. Table 5.4 shows that STT-LLM achieves the lowest sequence prediction errors (RMSE: 1664.59, MAPE: 96.80). Removing all components increases RMSE: +1.4%, MAE: +1.7%, MAPE: +2.2% relative to STT-LLM. Removing embeddings alone increases MAPE to 100.56 (+3.9%) and drops AUC to 0.5352 (-5.7%), highlighting the embedding layer's key role in aligning multimodal representations.

Table 5.4 Ablation study evaluating the contribution of structural and temporal components and embeddings in STT-LLM. Removing any components results in degraded performance, confirming their complementary roles. The STT-LLM model outperforms all ablated variants, especially in sensitivity and F1-score for anomaly detection, showing the synergy of integrating structural and temporal embeddings with pre-trained LLM representations.

| **Model Variants** | **Sequence Prediction** | | | **Anomaly Detection (Global)** | | | | |
|---|---|---|---|---|---|---|---|---|
| | **RMSE↓** | **MAE↓** | **MAPE↓** | **Acc↑** | **Sens↑** | **Prec↑** | **F1↑** | **AUC↑** |
| *w/o* all | 1687.71 | 896.39 | 98.93 | 0.7179 | 0.1398 | 0.3291 | 0.1962 | 0.5609 |
| *w/o* structural | 1687.49 | 896.61 | 100.65 | 0.7152 | 0.0968 | 0.2769 | 0.1434 | 0.4964 |
| *w/o* temporal | 1682.45 | 892.85 | 98.38 | 0.7126 | 0.1237 | 0.2987 | 0.1749 | 0.5500 |
| *w/o* embeddings | 1682.75 | 893.40 | 100.56 | 0.7139 | 0.1344 | 0.3125 | 0.1880 | 0.5352 |
| *w/o* structural + temporal | 1682.70 | 893.20 | 98.89 | 0.6967 | 0.0645 | 0.1791 | 0.0949 | 0.4877 |
| *w/o* embeddings + temporal | 1677.56 | 889.29 | 97.07 | 0.7245 | 0.1290 | 0.3429 | 0.1875 | 0.5474 |
| *w/o* embeddings + structural | 1679.16 | 891.78 | 97.35 | 0.7113 | 0.0914 | 0.2576 | 0.1349 | 0.4887 |
| **STT-LLM** | **1664.59** | **881.20** | **96.80** | **0.7338** | **0.1935** | **0.4138** | **0.2637** | **0.5675** |

For anomaly detection, STT-LLM achieves a good balance across different metrics. Removing all components lowers sensitivity by -27.8%, and precision by -20.5% compared to STT-LLM. Removing either the structural or temporal tokenizer reduces sensitivity by -50% (0.0968 - 0.1237) and precision by -33% (0.2769 - 0.2987), showing that both structural and temporal components are important for anomaly detection. When two components are removed, the degradation is even sharper, e.g., *w/o embeddings + structural* drops AUC by -14% (0.4887) relative to STT-LLM. Overall, the results demonstrate that all three components act synergistically to deliver the robust generalization and performance gains of STT-LLM across different tasks.

## 5.7 Case Study

To evaluate the real-world applicability of the method, a case study was conducted on 29 longitudinal steroid profiles from real-world athletes, which were verified through DNA analysis by a biomedical laboratory. Among these, 7 profiles were confirmed as anomalous due to doping-related abnormalities, with domain experts providing detailed explanations, and the remaining 22 were classified as clean. The clean profiles were used for sequence prediction, and all 29 profiles were used for anomaly detection. The model achieved better forecasting performance with RMSE: 1673.13, MAE: 868.93, and MAPE: 95.51. For anomaly detection, the model perfectly identified all 7 anomalous cases with 100% sensitivity, while only 2 clean profiles were misclassified (accuracy: 93.10%).

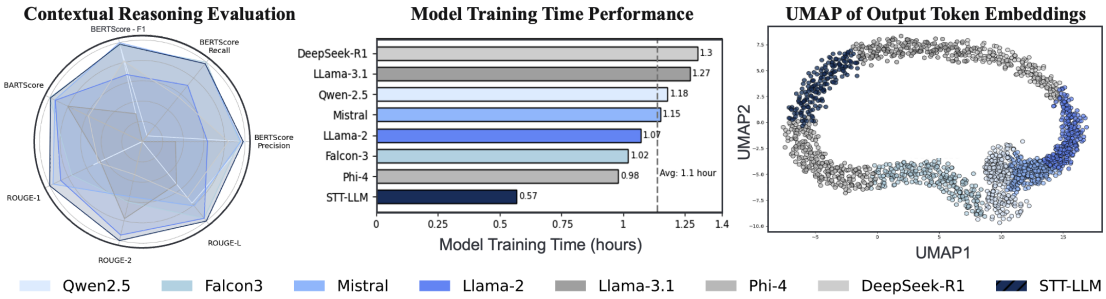


Fig. 5.6 Evaluation of contextual reasoning quality (left), model training efficiency (center), and output token embedding in UMAP representation (right). The radar chart on the left shows the contextual reasoning performance across different metrics. The center bar chart compares training time efficiency, highlighting STT-LLM's faster convergence. The right UMAP plot visualizes the clustering and separation of output token embeddings, with STT-LLM demonstrating distinct embedding structure compared to baseline LLMs.

To evaluate the contextual reasoning ability of STT-LLM, a few-shot setup was adopted in which 7 expert-annotated doping profiles were used to generate explanations for 500 additional profiles. These explanations were then used to train all models under identical training conditions. Model performance was subsequently assessed on the original 7 profiles using expert-provided ground-truth explanations. As shown in Fig. 5.6, STT-LLM was found to outperform all competing LLM baselines across multiple evaluation metrics, demonstrating higher alignment with expert interpretations. This result highlights the model's capability to capture clinically meaningful reasoning patterns from limited supervision. Additionally, training efficiency was evaluated by measuring the time required for 10 epochs (convergence) of fine-tuning. STT-LLM achieved the lowest training time (0.57 hours), substantially faster than all other baselines (average: 1.1 hours), owing to its compact structural-temporal tokenization strategy, which reduces sequence length and computational overhead. Finally,

the output token embedding spaces of different models were visualized using UMAP representation. Unlike tightly clustered distributions, the embeddings formed a continuous ring-like topology, with STT-LLM occupying a transitional zone between LLaMA-3.1 and Phi-4. This positioning suggests that STT-LLM maintains representational alignment with general-purpose LLMs while introducing localized structure unique to its domain-aware training.

## 5.8 Summary

This chapter introduced STT-LLM, a structural-temporal tokenization framework designed to adapt large language models (LLMs) for use with longitudinal clinical data. Unlike conventional LLMs, which operate on unstructured textual inputs, STT-LLM transforms structured biomedical data into token sequences compatible with standard LLM architectures. The framework constructs joint embeddings that encode both domain-specific structural relationships, such as those defined by metabolic pathways and temporal dynamics across time points. These embeddings are processed through two specialized tokenizers: one for structural dependencies and one for temporal progression. The resulting tokens are used directly by the LLM, enabling it to model longitudinal clinical data without modifying its architecture.

STT-LLM directly addresses RQ2, which explores why incorporating domain knowledge is important for improving anomaly detection in longitudinal data. Traditional time-series models often treat features independently or rely on sequence alignment methods that do not account for domain-specific feature interactions. STT-LLM incorporates this structured prior knowledge explicitly into the model input through structural tokenization, which encodes the biochemical relationships between biomarkers as node-level dependencies in a graph. This enables the LLM to reason over biologically meaningful structures and to detect coordinated changes across related biomarkers, rather than isolated fluctuations. The temporal tokenizer complements this by preserving sequence order and allowing the model to learn deviation patterns over time. By integrating structural and temporal behavior in a unified representation, STT-LLM provides a mechanism for inductive bias that improves both generalization and robustness in downstream tasks. This design allows anomaly detection to move beyond simple statistical outliers to detecting biologically implausible trajectories based on known pathway constraints, contributing to more precise and explainable decision support.

The model was evaluated on real-world steroid profiles in the context of anti-doping analytics. Tasks included anomaly detection and sequence forecasting, both in zero-shot and few-shot settings. STT-LLM outperformed several pre-trained and fine-tuned LLM baselines

in these tasks, demonstrating that embedding-guided tokenization significantly improves performance without requiring extensive retraining or access to large labeled datasets. The approach enables efficient deployment in settings with limited computational resources and strict privacy requirements, such as clinical laboratories or mobile health systems.

In summary, STT-LLM offers a general method for extending LLM capabilities to structured clinical domains by integrating biological knowledge into the input representation. It supports flexible and privacy-conscious inference while maintaining compatibility with standard LLM backbones. The framework is a step toward making large language models usable in clinical decision support systems where structured, time-dependent, and biologically coherent representations are essential.

# Chapter 6

# GRAMP: GRAph-based modeling for Metabolism Pathway

## 6.1 Introduction

[1]Modeling biological pathways is an important aspect of 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 [215]. 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 modeling biological pathways, ranging from qualitative to quantitative methods [264, 154, 123]. However, these methods have challenges like parameter estimation, model complexity, dynamic behavior, etc. Therefore, using these methods for modeling biological pathways leads to inaccurate predictions and limited applicability. However, these methods have challenges like parameter estimation, model complexity, dynamic behavior, etc. Therefore, using these methods for modeling biological pathways leads to inaccurate predictions and limited applicability.

Many forensic investigations primarily focus on analyzing these biological pathways to identify the fraudulent behavior of the individual, especially doping activities in sports [31]. Recent investigation at the 2014 Olympics Games in Sochi discovered a new form of fraudulent behavior by athletes [193]. Athletes were found attempting to replace their doping samples with clean samples obtained from other individuals to avoid positive test results.

[1]**Based on Publication:** Rahman, M.R., Hussain, M., Piper, T., Geyer, H., Equey, T., Baume, N., Aikin, R., Maass, W. (2023). modeling Metabolism Pathways using Graph Representation Learning for Fraud Detection in Sports. *In Proceedings of the IEEE International Conference on Digital Health, (ICDH 2023), Main Track.*

This fraudulent activity of sample swapping poses a substantial challenge in the forensic investigations of the World Anti-Doping organization (WADA) and other organizations.

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 [189], 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 [234, 223, 271]. In addition, machine learning has attracted considerable attention for detecting doping activities [271, 230]. Nevertheless, these approaches neglect an important factor, i.e., steroid metabolism pathways [250]. In other words, the structural relationship of different metabolites in the steroid metabolism pathways of the athlete is important 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 science or everyday life have benefited from formulating a relationship between entities as a graph. Therefore, graph networks have become increasingly popular in modeling complex systems due to their ability to capture intricate relationships [286]. They can be used to model complex real-world networks like biological pathways, where vertices represent biological entities, and edges indicate underlying connectivity [119]. Employing graph networks to model domain knowledge facilitates comprehensive coverage of important 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. The key contributions of this work are summarized as follows:

- A graph-based modeling for metabolism pathway is presented, which is capable of integrating domain knowledge of biological pathways into a machine learning model. It comprises an attention mechanism designed to capture direct relationships between different metabolites within the metabolic pathway to improve decision-making.
- Unlike previous methods, this model leverages both the structural and temporal relationships in steroid metabolism to obtain a more informative representation.
- The proposed method is evaluated on a real-world dataset collected by anti-doping organizations and laboratories. Experimental results demonstrate the effectiveness of

the model, which detects fraudulent athletes with relatively higher specificity compared to SoTA method.

## 6.2 Related Work

### Graph Representation Learning

Graph representation learning (GRL) [341] 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: i) network embedding models [220, 100, 68], which preserve the proximities among contextual nodes to capture graph structure information; ii) graph neural networks (GNNs) [140, 294, 337], which aggregate neighbor feature information to learn node embeddings; and iii) knowledge graph embedding methods [27, 267, 59], 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 SoTA in GRL. Recent advancements in GNNs, such as Graph Convolutional Networks (GCNs) [140], Graph Attention Networks (GATs) [294], and GraphSAGE [107], have further improved their expressive power and scalability. GATs incorporate an attention mechanism to calculate the weights of node neighborhoods 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 neighboring nodes and aggregating their features using functions like mean or max pooling.

In this chapter, a novel approach is proposed 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 organizations for detecting fraudulent activities in sports.

## 6.3 Preliminaries

**Sample** A urine sample collected from a given athlete for performing a doping test can be denoted as $\mathbf{x}_{ij} = \{x_{ij1}, x_{ij2}, \cdots, x_{ijp}\} \in \mathbb{R}^p$, where $p$ represents the total number of parameters, $i$ represents the athlete, and $j$ represents the sample index within the longitudinal profile. Each sample contains a set of parameters that reflect the concentration levels of different steroid metabolites in the human metabolism, as listed in Table 6.1.

Table 6.1 List of metabolites present in each steroid sample along with their corresponding 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 | $C_{19}H_{32}O_2$ |

**Longitudinal Profile** The athlete's longitudinal profile is defined as a sequence of samples collected over time and is represented by $\mathbf{X}_i = \{\mathbf{x}_{i1}, \mathbf{x}_{i2}, \cdots, \mathbf{x}_{in_i}\} \in \mathbb{R}^{n_i \times p}$, where $n_i$ is the total number of samples collected for athlete $i$. 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.

**Anomalous Behavior** The focus is placed on sample swapping, in which a contaminated sample from an athlete is exchanged with a clean sample from another individual. This results in a discrepancy between the sample under consideration ($\mathbf{x}_T$), 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 low compared to the clean athletic population. Therefore, this task can be formulated as anomaly detection problem.

**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 important for a variety of physiological processes, including the regulation of metabolism [250]. Steroid hormones, such as testosterone and estrogen, are synthesized 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. 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, while etiocholanolone and androsterone represent the end-products of androgen metabolism, and their urinary concentrations are therefore definitely elevated after exogenous testosterone administrations. Fig. 6.1 represents a 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 their performance [26]. 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 [26].
- *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 [242].

The significance of steroid metabolism in athlete doping lies in the detection and prevention of illicit usage. Anti-doping organizations employ different 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.

## 6.4 GRAph-based modeling for Metabolism Pathway (GRAMP)

A GRAph-based modeling approach for Metabolism Pathway (GRAMP) is proposed (Fig. 6.2), which incorporates domain knowledge for the identification of sample swapping in sports. The model consists of two main steps: i) embedding steroid metabolism into graph structure, and ii) model architecture for graph classification.

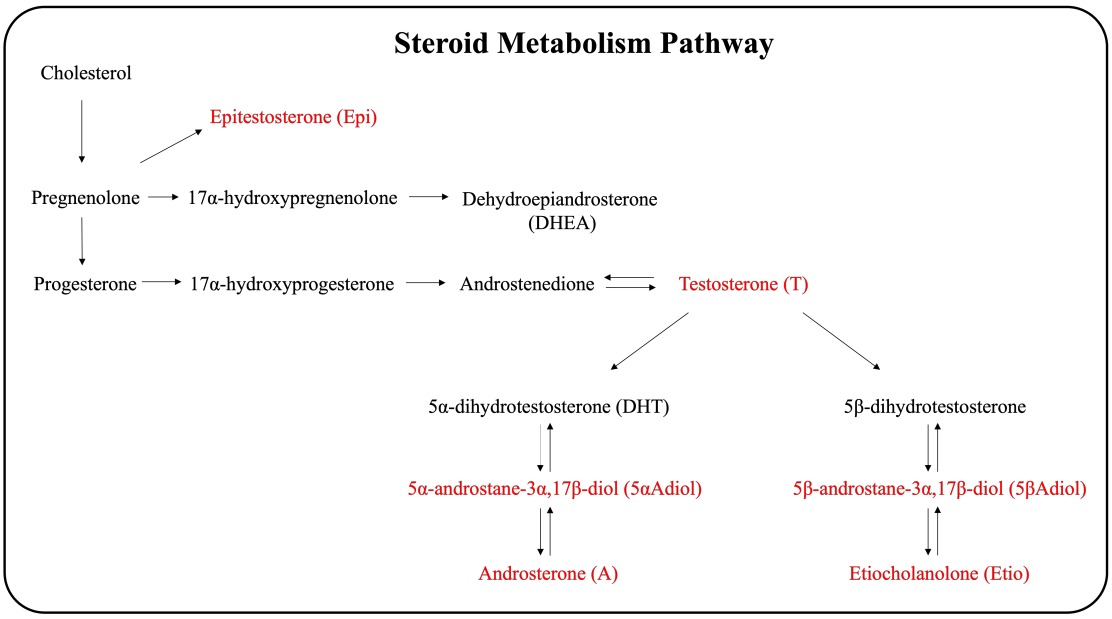


Fig. 6.1 Simplified human steroid metabolism pathway based on measured urinary steroids. The diagram shows the interconnected sequence of biochemical conversions involved in the synthesis and degradation of key steroid hormones.

### 6.4.1 Embedding Steroid Metabolism into Graph Structure

To embed the steroid metabolism pathway into a graph structure, each metabolite is considered as an individual node in the graph. The edges between these nodes are used to represent the connections and interactions between metabolites and reactions. For example, an edge may denote the conversion of testosterone to androsterone catalyzed by a specific enzyme. By representing the pathway as a graph, the structural relationships and dependencies between metabolites can be captured, enabling the discovery of important patterns and interactions within the steroid metabolism process.

**Graph Construction** A graph $G_i = (V_i, E_i)$ with directed edges consists of nodes $V_i = \{\mathbf{x}_{i1}, \mathbf{x}_{i2}, \cdots, \mathbf{x}_{in_i}\}$ and edges $E_i \subseteq V_i \times V_i$. The graph is constructed for each longitudinal profile of the athlete and the graph representation of the steroid metabolism pathway is defined as follows:

$$V_i = \Big\|_{j=1}^{n_i} V_{i,j} = \{V_{i,1} \,\|\, V_{i,2} \,\|\, \cdots \,\|\, V_{i,n_i}\} \tag{6.1}$$

$$E_i = \Big\|_{j=1}^{n_i} \Big\|_{k=1}^{n_i} E_{i,j,k} \subseteq V_i \times V_i \tag{6.2}$$

where $n_i$ denotes the total number of samples in the longitudinal profile for athlete $i$, and $\|$ represents the concatenation operation. Each sample can be represented as:

$$V_{i,j} = \{x_{ij0}, x_{ij1}, x_{ij2}, ..., x_{ijp}\} \tag{6.3}$$

$$E_{i,j,k} \subseteq V_{i,j} \times V_{i,j} \tag{6.4}$$

where $x_{ij0}$ represents the master node for each sample and $x_{ij1}$ to $x_{ijp}$ nodes represent each metabolites.

**Master Node** A master node is defined 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, the master node is defined as the cumulative representation of all metabolites within a given sample.

$$x_{ij0} = \sum_{k=1}^{p} x_{ijk} \tag{6.5}$$

where $x_{ij0}$ is the master node for sample $j$ of athlete $i$, and $x_{ijk}$ represents the metabolite $k$ in the sample. The master node serves as a central point of reference for each sample, capturing the overall characteristics of the metabolites present in that sample.

### 6.4.2 Model Architecture for Graph Classification

Once the steroid metabolism pathway has been transformed into a graph structure, a suitable model architecture is required for graph classification. The goal is to effectively utilize 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 neighboring 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 [294] 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 neighboring 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 the graph representation of the steroid metabolism pathway, the relevant features and interactions between metabolites can be effectively captured. The GAT

model is trained using labeled data, optimizing the attention weights and model parameters to achieve high-performance graph classification on the steroid metabolism pathway data. Table 6.2 show the detailed model architecture of the GRAMP model, including the graph attention mechanism acting on different nodes of a graph structure.

**Graph Attention Layer** It 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 in graph $G_i$, i.e., $m = n_i \times p$. Since each node is represented with a single metabolism parameter, $h_i \in \mathbb{R}$.

Self-attention is performed on the nodes using a shared attentional mechanism that computes attention coefficients $a_{ij}$ for each pair of nodes $i$ and $j$ in the graph. The attention coefficients are computed as follows:

$$a_{ij} = a^T [Wh_i || Wh_j] \tag{6.6}$$

where $W \in \mathbb{R}$ 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. Next, a non-linear activation function $e_{ij} = \sigma(a_{ij})$ is applied to the attention coefficients.

To ensure that the coefficients are easily comparable across different nodes, they are normalized over all neighbors of node $i$ using the softmax function:

$$\alpha_{ij} = softmax(e_{ij}) = \frac{exp(e_{ij})}{\sum_{k \in Ngb_i} exp(e_{ik})} \tag{6.7}$$

where $\alpha_{ij}$ represents the normalized attention coefficient from node $i$ to node $j$, and $Ngb_i$ denotes the set of neighbors of node $i$ in the metabolism pathway graph.

Finally, the output of the attention layer is computed by aggregating the features of neighboring nodes weighted by their attention coefficients:

$$h'_i = \sigma(\sum_{j \in Ngb_i} \alpha_{ij} W h_j) \tag{6.8}$$

where $\sigma$ is a non-linear activation function, and $W$ is a learnable weight matrix applied to the features of neighboring nodes. This process allows the model to focus on the most relevant nodes in the graph, effectively capturing the relationships between metabolites in the steroid metabolism pathway.

**Loss Function** For the graph classification task of distinguishing anomalous and normal longitudinal profiles, the binary cross-entropy (BCE) loss function was used as follow:

Table 6.2 Detailed architecture of GRAMP model, showing the input output dimensions and the number of attention heads used in each layer.

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

$$\mathscr{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\Big(\mathbf{y}_i \cdot \log \hat{\mathbf{y}}_i + \big(1-\mathbf{y}_i\big) \cdot \log\big(1-\hat{\mathbf{y}}_i\big)\Big) \tag{6.9}$$

where $N$ is the total number of longitudinal profiles, $\mathbf{y}_i$ is the true label for profile $i$, and $\hat{\mathbf{y}}_i$ is the predicted probability for profile $i$. The model is trained to minimize this loss function, thereby improving its ability to accurately classify graphs as either anomalous or clean.

## 6.5 Experiments

### 6.5.1 Datasets

WADA and other anti-doping organizations 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. The dataset represents the longitudinal profiles of real-world male and female athletes [234]. It consists of 1,432 longitudinal profiles corresponding to 7,545 samples, where each athlete may have between 3 and 20 samples in their profile. The dataset was randomly partitioned, with 80% used for training and 20% for testing the algorithm. A summary of the number of samples belonging to male and female athletes is provided in Table 6.3. Each sample comprises a set of biomarkers,

referred to as steroid metabolism parameters, which exhibit significant changes upon steroid administration, as listed in Table 6.1.

Table 6.3 Data statistics used for training and testing the GRAMP model.

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

### 6.5.2 Baseline Methods

A set of baseline models was selected to serve as a performance benchmark for comparing the proposed GRAMP model. These baselines consist of both non-graph and graph-based models that do not incorporate domain knowledge 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 optimized using the training dataset.

- **Bayesian Method (SoTA)** [271]: Determines personalized thresholds for each steroid parameter by modeling prior distributions derived from a reference population. These thresholds are then used to assess new samples.
- **Random Forest (RF)** [29]: An ensemble learning method that combines multiple decision trees to improve classification accuracy and improve interpretability.
- **XGBoost (XGB)** [41]: Utilizes an optimized, distributed gradient boosting algorithm to achieve high predictive performance on structured data.
- **Graph Convolutional Network (GCN)** [140]: Learns representations of nodes in a graph structure, where each node corresponds to a sample in the longitudinal profile.
- **Graph Isomorphism Network (GIN)** [326]: Learns expressive node embeddings by aggregating both local and global substructural information, with each node representing a sample.
- **Graph Attention Network (GAT)** [294]: Applies attention mechanisms to graph nodes for learning informative embeddings, achieving SoTA performance on various graph-based tasks. Each node represents a sample.

### 6.5.3 Experimental Settings

Given that the fraud detection problem has been framed as a supervised graph classification task, it is important that a labeled dataset be used, comprising samples for each class, i.e., normal and anomalous profiles. A random selection was performed on the dataset, with 50% of the profiles chosen. In each selected profile, one sample was manually replaced with a sample from a different profile. These modified profiles were labeled as anomalous (labeled as '1'). The remaining 50% of the profiles were considered normal (labeled as '0'). To ensure consistency, each profile was normalized separately to unit norm.

All the models are implemented based on the SCIKIT-LEARN [218], XGBOOST [41], and PYTORCH-GEOMETRIC [82] packages. One significant challenge during model training was overfitting, which limits the model's generalization capability. Since the training dataset is small, addressing overfitting was identified as a critical concern in this analysis. Therefore, the $k$-fold cross-validation method [238] was performed to train the 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. The hyperparameter optimization framework is used to efficiently explore a substantial grid space while promptly eliminating unpromising trials and implemented this framework using OPTUNA [3]. The optimized trained model is deployed on the testing set, enabling predictions for previously unseen profiles. The performance evaluation of each model was conducted using accuracy, sensitivity, specificity, and area under the ROC curve.

## 6.6 Results

### 6.6.1 Performance Comparison

The performance of the GRAMP model was compared with all baseline models for detecting fraudulent behaviour, i.e., sample swapping, on both male and female datasets, as presented in Table 6.4 and Table 6.5, 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. The proposed GRAMP model outperformed all baselines, showing that adding domain knowledge by defining a graph structure based on the metabolism pathway is effective. Sensitivity values exceeding 80% and AUC values exceeding 90% were achieved on both male and female athletes.

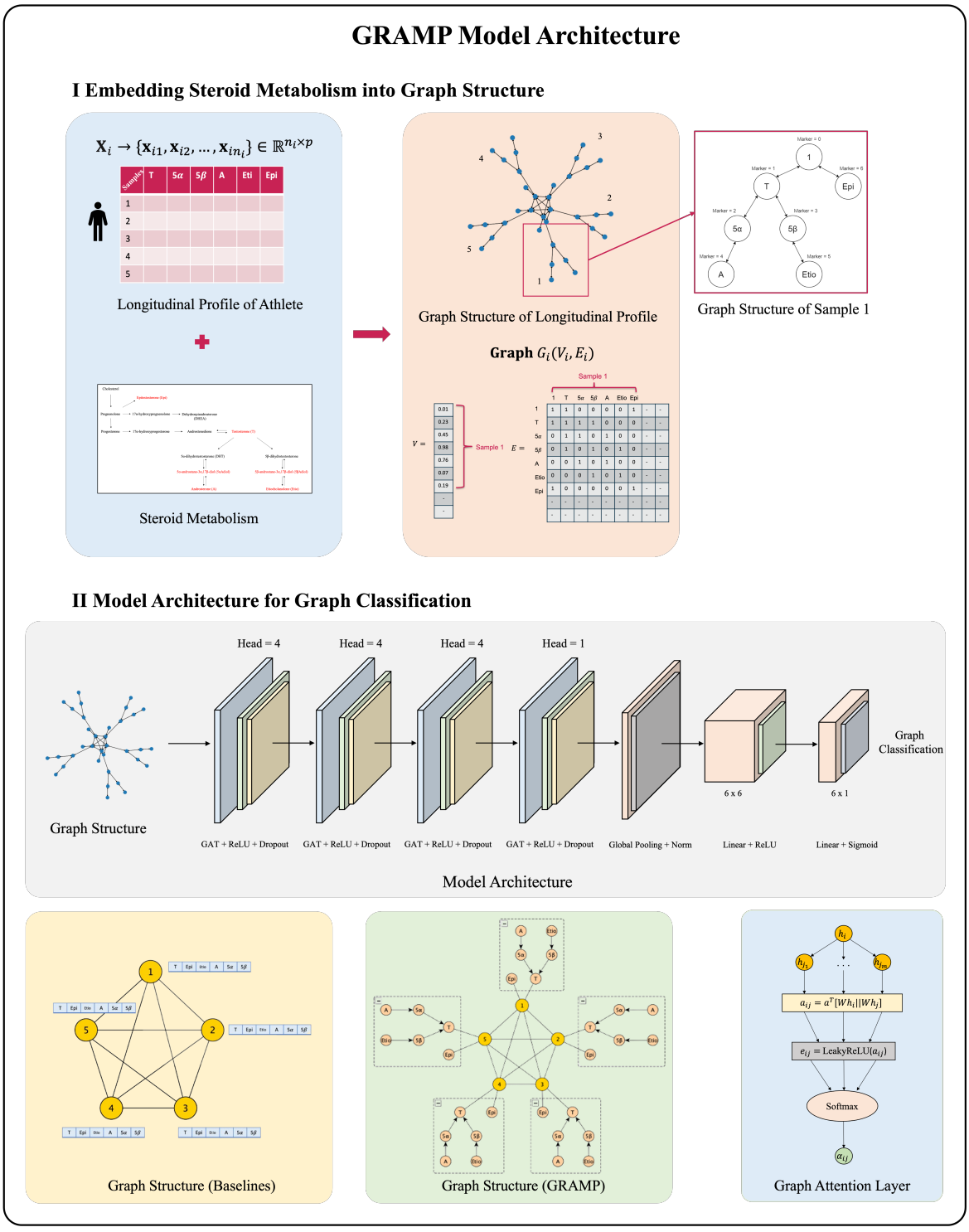


Fig. 6.2 Overview of the GRAMP model: i) The steroid metabolism pathway is embedded into a graph structure by treating metabolites as nodes and defining their biochemical relationships as edges. This is combined with the longitudinal steroid profile of an athlete to form a time-resolved graph representation for each sample. ii) The graph classification architecture leverages multiple Graph Attention Network (GAT) layers, each followed by ReLU activation and dropout for regularization. Baseline and GRAMP-specific graph construction strategies are shown, along with the internal mechanism of the attention computation.

Table 6.4 Performance comparison of the proposed GRAMP model and baseline methods on male athletes. The GRAMP model consistently achieves the better scores across all metrics, showing its improved generalization capability.

| **Metrics** | **SoTA** | | **RF** | | **XGB** | | **GCN** | | **GIN** | | **GAT** | | **GRAMP** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Train | Test | Train | Test | Train | Test | Train | Test | Train | Test | Train | Test | Train | Test |
| AC | - | 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** |
| AU | - | - | 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 6.5 Performance comparison of the proposed GRAMP model and baseline methods on female athletes. The GRAMP model consistently achieves the better scores across all metrics, showing its improved generalization capability.

| **Metrics** | **SoTA** | | **RF** | | **XGB** | | **GCN** | | **GIN** | | **GAT** | | **GRAMP** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Train | Test | Train | Test | Train | Test | Train | Test | Train | Test | Train | Test | Train | Test |
| AC | - | 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** |
| AU | - | - | 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** |

### 6.6.2 Precision-Recall Curve

The ROC and PRC curves for all models evaluated on male and female datasets are presented in Fig. 6.3. As depicted, the proposed model outperforms all the baseline models in both curves. The results for graph models 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.

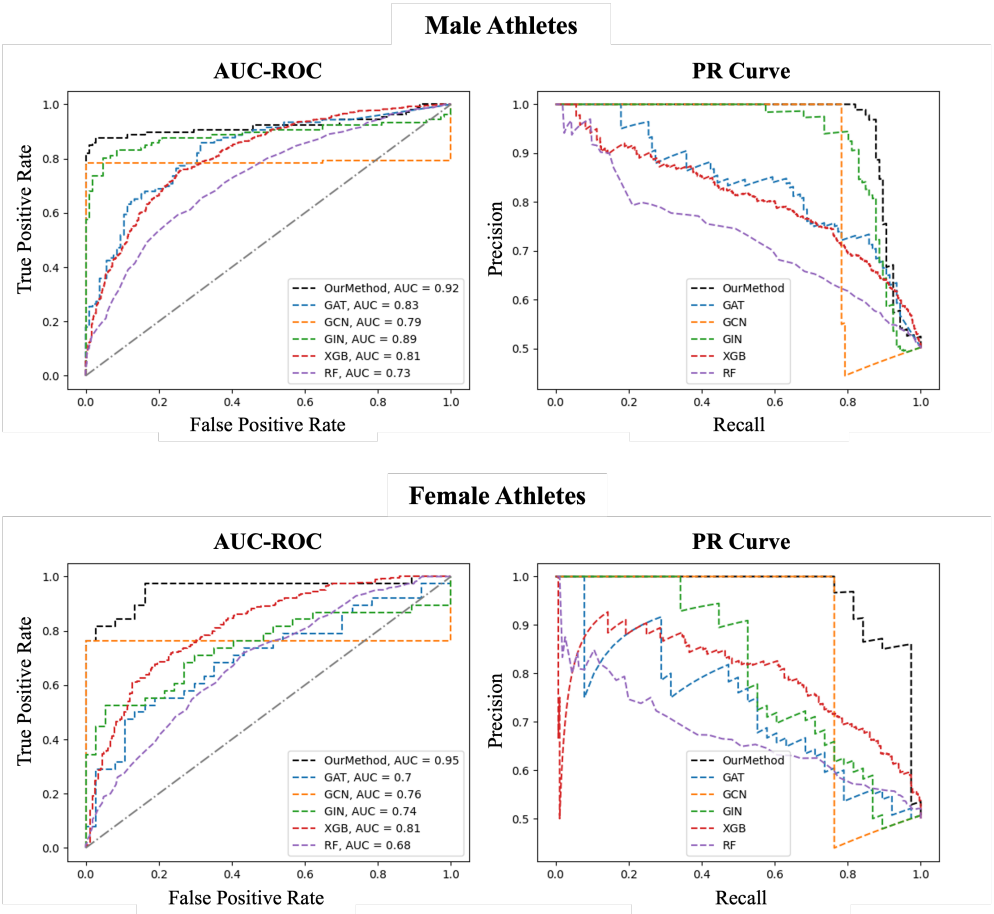


Fig. 6.3 ROC and PR curves comparing the performance of the proposed GRAMP model and baseline models for male and female athletes. The proposed model consistently achieves the highest AUC across both ROC and PR curves, showing better detection capability across genders.

A longitudinal profile of both a male and a female athlete was randomly selected from the testing dataset, and the pairwise attention coefficients for one of the samples were computed. Fig. 6.4 illustrates the weighted contribution of the neighborhood to each node. It can be observed that testosterone and epitestosterone exhibit the highest attention coefficients to

the master node in comparison to other metabolism parameters. This indicates that higher importance is assigned by the model to the message passing between the master node and testosterone, as well as between the master node and epitestosterone.

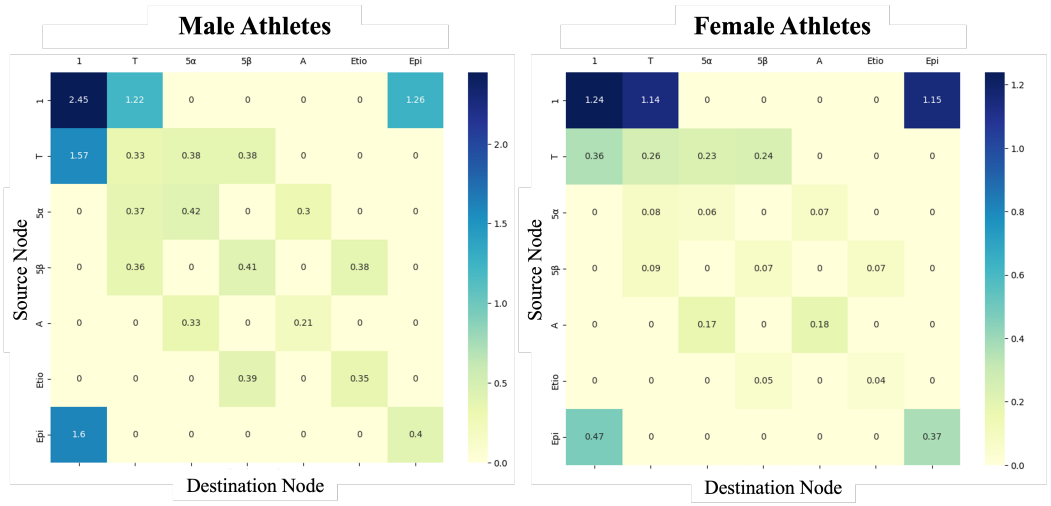


Fig. 6.4 Pairwise attention coefficients for a randomly selected sample from the randomly selected longitudinal profile of male and female athletes. The attention heatmaps show the significance and direction of information propagation between different steroid metabolites within the GRAMP model. Higher attention weights indicate stronger influence from source to destination node, thereby revealing the underlying dependencies and pathway-specific interactions that contribute to the model's interpretability.

Since the data for male athletes are generally more sparse than for female athletes, i.e., greater variation is observed in the concentration values of metabolism parameters in the male body. Therefore, higher attention coefficient values are recorded for male athletes. Additionally, two cases of information propagation are considered: i) when testosterone is the source and the master node is the destination, indicating message passing from testosterone to the master nodem and ii) when the master node is the source and testosterone is the destination. For male athletes, relatively similar attention coefficient values are observed in both cases, suggesting that bidirectional message passing is significant. In contrast, for female athletes, relatively higher attention coefficients are observed in the latter case. A similar pattern is also observed with epitestosterone.

Overall, the proposed GRAMP model consistently outperforms other SoTA baseline models due to two factors. First, this model effectively captures the structural behavior 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, the model explicitly considers their structural characteristics. Next, the model incorporates an attention mechanism that generates high-level embeddings, facilitating improved pattern

learning. Consequently, GRAMP 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.

### 6.6.3 Ablation Studies

The ablation studies were performed 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, different variations in the master node were tried by selecting different functions, including *SUM*(nodes) (the sum of values of all nodes), T/E (the ratio of T and E), and *Avg*(nodes) (the mean of values of all nodes). Fig. 6.5 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, it was observed that T/E shows low performance for male athletes because it only contains information about testosterone and epitestosterone. On the other hand, *Avg*(nodes) shows low performance for female athletes because the concentration values of all the metabolism parameters have different scales, especially for female athletes due to data sparsity. Therefore, averaging all the values would not be a feasible solution and selecting the sum of all the values of all the metabolism parameters outperforms the other two variations for both male and female athletes.

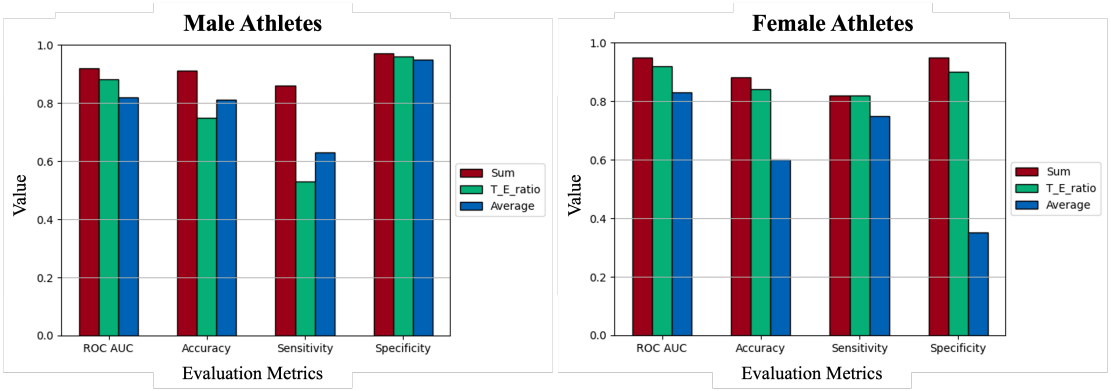


Fig. 6.5 Performance of the GRAMP model under different configurations of the master node for male and female athletes. The comparison shows how the choice of master node affects the model's capability in detecting and classifying sample swapping cases.

Next, to assess the importance of the attention layer, the number of GAT layers in the model network was varied. Fig. 6.6 presents the model's performance for both male and female athletes. It was observed that the performance increased with the addition of GAT layers up to a certain point, after which it began to decline. Four layers were found to be

optimal for this problem, indicating that at least four hops are required for effective message passing within the graph network.

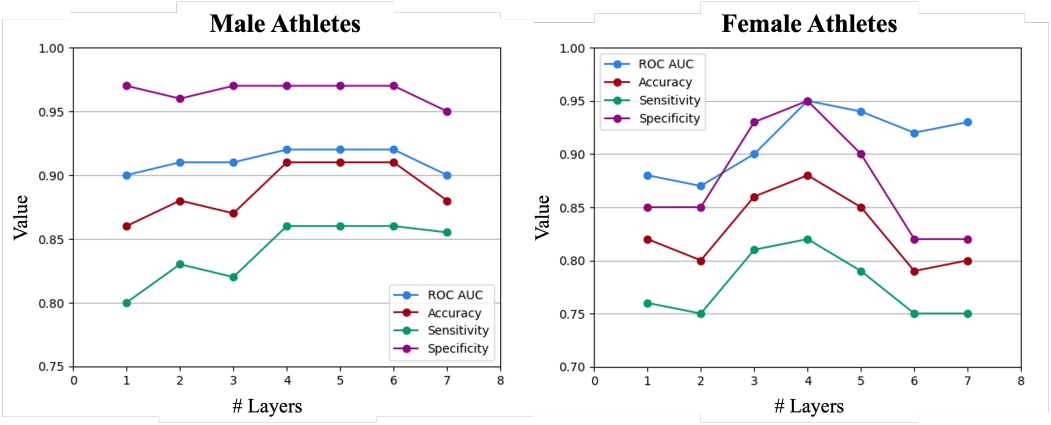


Fig. 6.6 Performance of the GRAMP method across different numbers of graph attention layers for male and female athletes. The plots depict how different metrics vary with the number of GAT layers.

## 6.7 Summary

Modeling biological pathways plays a crucial role in supporting decision-making in clinical and forensic applications, including anti-doping investigations in sports. The issue of sample swapping has emerged as a serious form of fraudulent behaviour, allowing athletes to avoid positive doping test results. Many existing detection methods treat biomarkers as independent variables and do not consider the underlying biological structure that connects these variables through known metabolic pathways. To address this limitation, this chapter introduced GRAMP, a graph-based anomaly detection model that incorporates domain knowledge through graph representation learning. GRAMP models the steroid metabolism pathway as a structured graph and uses graph attention mechanisms to learn meaningful embeddings that reflect both direct and indirect relationships between steroid biomarkers. The model was evaluated on real-world longitudinal datasets and demonstrated improved accuracy and reduced false positives compared to existing SoTA approaches, validating its utility for decision-making in anti-doping contexts.

GRAMP directly addresses RQ2, which concerns the role of domain knowledge in improving the performance of anomaly detection systems. The existing models rely on simplifying assumptions such as feature independence or uniform feature importance. GRAMP overcomes these constraints by leveraging Graph Attention Networks, which enable it to

learn context-sensitive node embeddings where the influence of each biomarker is modulated by its neighbours in the metabolic pathway. This mechanism ensures that anomalies are detected not just as statistical outliers in individual features but as pathway-level deviations that manifest across biochemically related nodes. As such, GRAMP is particularly effective in identifying manipulation patterns, such as hormonal suppression or synthetic enhancement, that propagate through the network in coordinated and physiologically implausible ways. The proposed model introduces a biologically informed computational framework that utilizes attention mechanisms to provide not only performance gains but also interpretability. By assigning edge-wise attention coefficients during training, GRAMP identifies the most influential nodes and substructures responsible for anomalous predictions. Furthermore, the use of metabolic graphs enables the model to generalize across athletes with varying profile lengths and biological baselines, offering robustness in noisy data conditions, which is a persistent challenge in real-world longitudinal datasets.

The evaluation showed that GRAMP achieves higher sensitivity and specificity than existing methods, including the Bayesian model currently used by WADA. It correctly identified more sample swapping cases and can reduce unnecessary DNA testing by up to 15%, showing clear operational advantages. GRAMP also outperformed non-graph models by 17-25% in decision accuracy, and its robustness under sparse and noisy data conditions further supports its use in real-world anti-doping workflows.

In conclusion, GRAMP provides a methodologically grounded approach for integrating biological knowledge into anomaly detection systems using graph neural networks. By capturing structured interactions among biomarkers, the model delivers biologically consistent and interpretable outputs suitable for regulatory decision-making. This contributes to both improved performance in detecting fraudulent behaviour and the broader goal of integrating structured domain knowledge into machine learning models for clinical and forensic applications.

# Section IV: Interpretability and Domain-Informed Reasoning

# Chapter 7

# MPP: Metabolism Pathway-driven Prompting

## 7.1 Introduction

[1]Longitudinal clinical profiles represent repeated measurements of biological samples such as blood, urine, or other biological specimens collected over time [254, 6]. These profiles are important in capturing the dynamic nature of biological processes, as they provide a time-evolving perspective of various physiological processes. The biomarkers measured within these samples often reflect underlying metabolic pathways [201]. Detecting anomalies in such data requires not only accurate statistical modeling but also interpretability, as clinicians and regulators need to understand the rationale behind why a profile is considered suspicious.

In clinical settings, anomaly detection in these longitudinal profiles is an important task [127]. Identifying abnormal behavior in such data can reveal critical insights, ranging from disease diagnosis to sample tampering as potential doping activity in sports [203, 330]. It mainly helps clinicians to monitor biological and physiological changes over time and detect suspicious behavior. Several studies have highlighted the potential and limitations of Large Language Models (LLMs) in clinical domain-specific tasks [8, 134, 287]. Despite their success in text generation, completion tasks, etc., their ability to process and analyze longitudinal clinical data, particularly in the context of metabolic pathways and biological changes, remains underexplored [99, 1]. Understanding how these models can leverage metabolic information to make informed decisions is critical for improving their performance in anomaly detection tasks.

[1]**Based on Publication:** Rahman, M.R., Liu, R., Maass, W. (2024). Incorporating Metabolic Information into LLMs for Anomaly Detection in Clinical Time-Series. *In Workshop on Time Series in the Age of Large Models: Neural Information Processing Systems (NeurIPS 2024).*

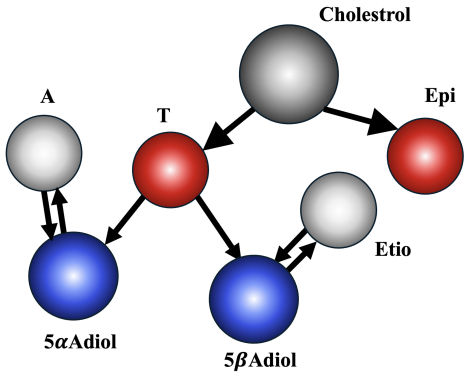


Fig. 7.1 Simplified representation of the steroid metabolism pathway, showing the biochemical relationships between key urinary steroid metabolites. Nodes represent individual metabolites, while directed edges indicate the metabolic conversions.

This chapter addresses RQ3, which concerns how anomaly detection systems can provide interpretable and domain-informed reasoning. Specifically, a targeted prompting method is proposed by integrating metabolic pathway structures into LLMs to improve their ability to detect anomalies based on contextual understanding. The effectiveness of this approach is demonstrated in the context of doping detection in sports, where it is applied to identify suspicious urine samples within athletes' longitudinal profiles. These profiles include the concentrations of different metabolites, reflecting the steroid metabolism as shown in Fig. 7.1, and are important for identifying potential doping activities [223, 231]. The key contributions of this work can be summarized as follows:

- Metabolism Pathway-driven Prompting (MPP) is proposed, which incorporates information about metabolic pathway structures and the temporal evolution of different metabolites into LLMs for anomaly detection. This approach improves explainability by enabling contextual reasoning grounded in domain-specific biochemical knowledge.
- The effectiveness of this method is demonstrated in the context of doping detection in sports and compared with the baseline prompting methods like zero-shot learning, in-context learning and chain-of-thought.

## 7.2 Related Work

**Anomaly Detection in Longitudinal Profiles**

Detecting anomalies in longitudinal clinical data is a key task in biomedical data analysis, often used for disease monitoring, treatment assessment, and fraud detection in sports.

Traditional methods such as Isolation Forests [168] and variational autoencoders like $\beta$-VAE [182] are widely used for unsupervised detection of outliers in high-dimensional data. While effective in capturing statistical deviations, these models typically lack biological interpretability. To address this, recent works in systems biology and metabolomics have focused on integrating domain knowledge, such as biochemical pathway structures, into data-driven models [99]. Graph-based representations of metabolic networks have shown particular promise in modeling steroid metabolism for anomaly detection, especially in anti-doping efforts [231], where the temporal and relational aspects of metabolite transitions are critical for detecting suspicious physiological patterns.

#### Language Models for Biomedical Reasoning

In biomedical settings, models like BioGPT [181] and ClinicalBERT [10] have been fine-tuned for domain-specific applications including clinical note summarization, diagnosis classification, and question answering. However, these models are predominantly trained on unstructured textual data and are not inherently designed to reason over structured time-series or graph-based clinical inputs. While prompting strategies such as zero-shot learning [260], in-context learning [217], and chain-of-thought prompting [92] have been proposed to improve LLM generalization, they remain limited in clinical anomaly detection due to their lack of temporal and physiological reasoning capabilities. Moreover, studies such as [1] have highlighted that generic LLMs often fail to recognize meaningful biomedical patterns without structured domain adaptation. This work introduces a targeted prompting approach that combines the temporal dynamics of clinical profiles with metabolic pathway graphs to improve LLM reasoning and anomaly detection in structured biomedical time-series data.

## 7.3 Preliminaries

Let us consider the longitudinal profile of athletes $\mathbf{X}_i = \{\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{in_i}\}$, where $\mathbf{x}_{ij} \in \mathbb{R}^p$ with total $p$ metabolite and $\mathbf{x}_{ij,k}$ represents the measurement of metabolite $k$ at time $j$. The temporal difference is defined as $\Delta\mathbf{x}^T_{ij,k} = \mathbf{x}_{ij,k} - \mathbf{x}_{i(j-1),k}$ representing the change in metabolite $k$ over time. The anomaly detection task is to learn a function $f(\mathbf{x}_{ij})$ that gives an anomaly score to each sample $\mathbf{x}_{ij}$ in the longitudinal profile $\mathbf{X}_i$. The function flags the anomalous sample if the magnitude of the sum of $\Delta\mathbf{x}^T_{ij,k}$ exceeds a predefined threshold $\delta$,

indicating significant deviation from the expected change:

$$f(\mathbf{x}_{ij}) = \begin{cases} 1, & \left| \sum_{k=1}^{p} \Delta \mathbf{x}_{ij,k}^{T} \right| > \delta \\ 0, & \text{else} \end{cases} \tag{7.1}$$

The metabolic structural difference is defined as $\Delta \mathbf{x}_{ij,k}^{M} = \mathbf{x}_{ij,k} - \mathbf{x}_{ij,(k+1)}$ which needs to be considered.

## 7.4 Metabolism Pathway-driven Prompting (MPP)

A targeted prompting method is proposed by integrating metabolic pathway structures and their temporal evolution as shown in Fig. 7.2. First, LLM (*Pre-Prompt I*) is tasked to analyze the longitudinal profile and detect anomalies using zero-shot learning. Here, the LLM usually considers the temporal changes between consecutive samples. If these changes exceed the statistically significant threshold, it flags the corresponding sample as anomalous with an explanation. In a separate session, the LLM was provided with *Pre-Prompt II*, which included the temporal and metabolic graph representation of the given longitudinal profile, along with a task to extract domain-specific contextual information from these graph structures. The LLM generates a detailed textual explanation by assessing whether the temporal changes are consistent with the expected metabolic behavior based on known pathways. Next, the textual representation of domain knowledge is provided to the previous LLM, which is then tasked to rethink (*Prompt*) by incorporating this domain-specific information. The LLM refines the initial prediction by combining the domain-specific information and provide more accurate, and contextually aware prediction.

### 7.4.1 Temporal Graph

The graph $G_T = (V_T, E_T)$ represents the change in concentration levels of different steroids over time. Nodes are defined as $V_T = \{\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{in_i}\}$, where each node corresponds to the sample in $\mathbf{X}_i$ and the node feature represents the measurements for the $p$ steroids. The edges $E_T = \{w_T(\mathbf{x}_{i1} \rightarrow \mathbf{x}_{i2}), w_T(\mathbf{x}_{i2} \rightarrow \mathbf{x}_{i3}), \ldots, w_T(\mathbf{x}_{i(n_i-1)} \rightarrow \mathbf{x}_{in_i})\}$ represent transitions between nodes over time, connecting the $\mathbf{x}_{ij}$ between successive time points and the edge weights as the Euclidean distance between the steroid levels at two time points and normalized to the range [0, 1], incorporating the changes in all $p$ steroids. For the edge connecting $\mathbf{x}_{i(n_i-1)}$ and $\mathbf{x}_{in_i}$, the weight could be calculated as:

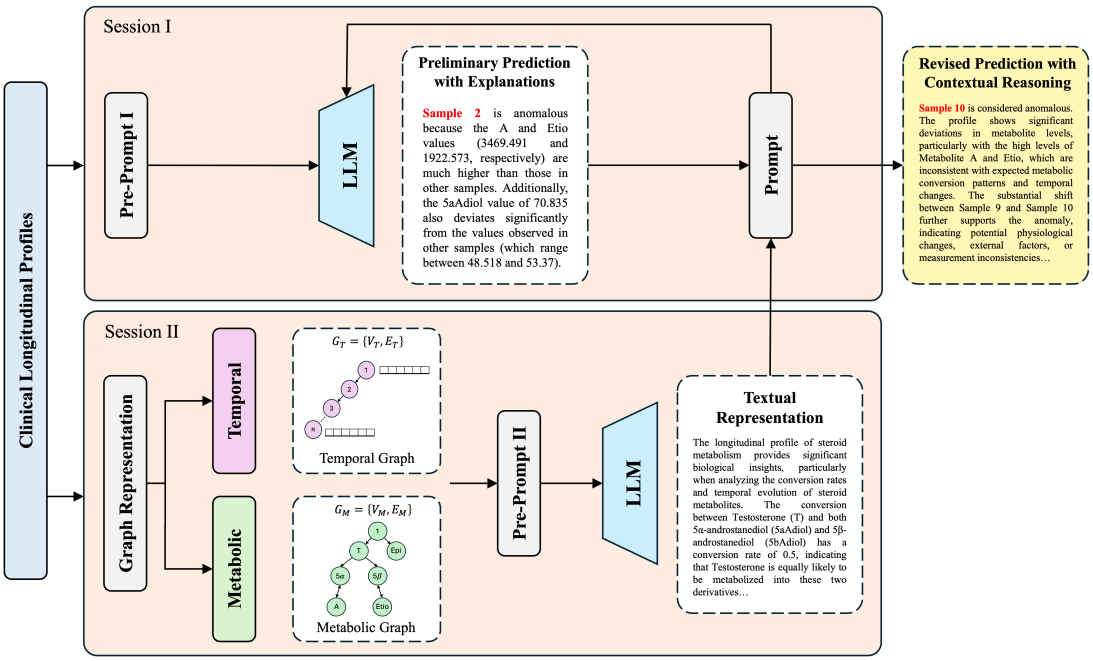


Fig. 7.2 Schematic overview of the Metabolism Pathway-driven Prompting (MPP) method. The method operates in two sessions: Session I leverages the LLM with a pre-prompt derived directly from clinical longitudinal profiles to generate preliminary predictions and explanations. Session II constructs both temporal and metabolic graphs to form a graph-based textual representation, which is then used to refine the prompt through contextual reasoning. The revised prediction integrates pathway knowledge and temporal trends, enabling more biologically plausible and explainable outputs from the LLM.

$$w_T(\mathbf{x}_{i(n_i-1)} \to \mathbf{x}_{in_i}) = \sqrt{\sum_{k=1}^{p} (\mathbf{x}_{i(n_i-1),k} - \mathbf{x}_{in_i,k})^2} \tag{7.2}$$

The temporal graph is represented as an adjacency matrix $A_T$ where each entry $A_T(\mathbf{x}_{i(j-1)}, \mathbf{x}_{ij})$ represents the weight of the edge from node $\mathbf{x}_{i(j-1)}$ to node $\mathbf{x}_{ij}$:

$$A_T = \begin{pmatrix} 0 & w_T(\mathbf{x}_{i1} \to \mathbf{x}_{i2}) & w_T(\mathbf{x}_{i1} \to \mathbf{x}_{i3}) & \dots & w_T(\mathbf{x}_{i1} \to \mathbf{x}_{in_i}) \\ 0 & 0 & w_T(\mathbf{x}_{i2} \to \mathbf{x}_{i3}) & \dots & w_T(\mathbf{x}_{i2} \to \mathbf{x}_{in_i}) \\ 0 & 0 & 0 & \dots & w_T(\mathbf{x}_{i3} \to \mathbf{x}_{in_i}) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix} \tag{7.3}$$

### 7.4.2 Metabolic Graph

The graph $G_M = (V_M, E_M)$ represents the directional flow of $p$ different metabolites in the pathway. Nodes are defined as $V_M = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_p\}$, where each node $\mathbf{s}_i$ represents a steroid. The edges represent the interactions or metabolic conversions between these steroids. The weight of an edge $w_M(\mathbf{s}_i \rightarrow \mathbf{s}_j)$ represents the conversion rate from steroid $\mathbf{s}_i$ to steroid $\mathbf{s}_j$, where $i, j = 1, 2, \ldots, p$ and $i \neq j$. If there is no conversion between two steroids, the corresponding entry is zero. The metabolic graph is represented as an adjacency matrix $A_M$ where each entry $A_M(\mathbf{s}_i, \mathbf{s}_j)$ represents the weight of the edge from node $\mathbf{s}_i$ to node $\mathbf{s}_j$:

$$A_M = \begin{pmatrix} 0 & w_M(\mathbf{s}_1 \rightarrow \mathbf{s}_2) & w_M(\mathbf{s}_1 \rightarrow \mathbf{s}_3) & \ldots & w_M(\mathbf{s}_1 \rightarrow \mathbf{s}_p) \\ w_M(\mathbf{s}_2 \rightarrow \mathbf{s}_1) & 0 & w_M(\mathbf{s}_2 \rightarrow \mathbf{s}_3) & \ldots & w_M(\mathbf{s}_2 \rightarrow \mathbf{s}_p) \\ w_M(\mathbf{s}_3 \rightarrow \mathbf{s}_1) & w_M(\mathbf{s}_3 \rightarrow \mathbf{s}_2) & 0 & \ldots & w_M(\mathbf{s}_3 \rightarrow \mathbf{s}_p) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w_M(\mathbf{s}_p \rightarrow \mathbf{s}_1) & w_M(\mathbf{s}_p \rightarrow \mathbf{s}_2) & w_M(\mathbf{s}_p \rightarrow \mathbf{s}_3) & \ldots & 0 \end{pmatrix} \tag{7.4}$$

## 7.5 Experiments

### 7.5.1 Datasets

Two real-world datasets (Steroid-M and Steroid-F) were used, consisting of longitudinal steroid profiles collected from male and female athletes, respectively [234, 232]. The Steroid-M dataset contains 755 profiles with 4214 samples and Steroid-F dataset contains 375 profiles with 2307 samples. The data contains less than 20% anomalous longitudinal profile.

### 7.5.2 Baseline Methods

Experiments are conducted using different open-source LLMs: (i) LLaMa 2-7B [235], (ii) Mistral-7B [207], (iii) Falcon-7B [174], and (iv) GPT2 [133]. These models are selected due to their efficiency in providing quicker results, which is particularly suitable for the size of the dataset. The performance of the proposed method is compared with various baseline prompting methods, including Zero-Shot prompting (ZS) [260], In-Context Learning (ICL) [217], and Chain-of-Thought (CoT) [92], as well as two non-LLM-based models, IsoForest [75] and $\beta$-VAE [182].

### 7.5.3 Experimental Settings

All experiments were conducted on a workstation equipped with an NVIDIA TITAN RTX GPU (24GB), Intel i9 processor, and 31GB total RAM. The same computational setup was used for all prompting methods and language models to ensure fair and consistent comparisons. All models were run in inference mode without fine-tuning, using float16 precision where supported. The implementation was carried out using PyTorch (v2.1.0), and the Hugging Face Transformers library (v4.39.1), with GPU acceleration enabled through CUDA 12.1 and cuDNN 8.9.

Graph structures were constructed using the NETWORKX (V3.2) package and transformed into structured textual prompts. Classification metrics such as accuracy, sensitivity, specificity, and F1-score are used for the anomaly detection task. To reduce randomness, all experiments were repeated over three independent runs with fixed random seeds, and results were averaged. The full prompt designs used in the MPP framework are shown in Fig. 7.3, 7.4, 7.5. The prompts were designed to be concise yet informative, ensuring that the LLMs could effectively leverage the metabolic pathway information while maintaining clarity in the task description.

**Pre-Prompt I:**

***Task:*** *Given the following longitudinal steroid profile of an athlete, where multiple urine samples have been collected at different time points, each containing values for various steroid parameters such as A, Etio, E, T, 5aAdiol, and 5bAdiol, identify if any sample is anomalous compared to the others. If an anomalous sample exists, return the sample number and provide an explanation for why it is considered anomalous. If no anomalies are found, state that the profile is clean.*

*Longitudinal profile data:*

| Sample No. | A | Etio | E | T | 5aAdiol | 5bAdiol |
|---|---|---|---|---|---|---|
| 1 | 2208.064 | 1237.146 | 16.851 | 29.344 | 48.518 | 122.957 |
| 2 | 2428.87 | 1360.864 | 18.534 | 26.406 | 48.518 | 135.251 |
| 3 | 2428.87 | 1113.434 | 16.851 | 29.344 | 48.518 | 122.957 |
| 4 | 2208.064 | 1113.434 | 16.851 | 29.344 | 48.518 | 122.957 |
| 5 | 1987.258 | 1360.864 | 15.162 | 29.344 | 48.518 | 135.251 |
| 6 | 2208.064 | 1113.434 | 16.851 | 26.406 | 53.37 | 110.662 |
| 7 | 2208.064 | 1237.152 | 18.534 | 32.275 | 43.667 | 110.662 |
| 8 | 1987.258 | 1113.434 | 16.851 | 29.344 | 53.37 | 122.957 |
| 9 | 1987.258 | 1360.864 | 18.534 | 29.344 | 43.667 | 135.251 |
| 10 | 3469.491 | 1922.573 | 17.402 | 27.341 | 70.835 | 132.211 |

*Please perform anomaly detection and explain the findings.*

Fig. 7.3 Pre-Prompt I for Metabolism Pathway-driven Prompting. The task instructs the language model to identify anomalous samples based on deviations from the rest of the samples in the longitudinal profile and provide explanations for detected anomalies.

## 7.6 Results

### 7.6.1 Performance Comparison

Table 7.1 shows that by incorporating domain-specific knowledge of metabolic pathways, MPP improves the LLMs' understanding of clinical data, leading to better performance. For the LLaMA 2-7B model, MPP achieves an accuracy of 71.4% and an F1 score of 57.0%, outperforming ZS's 65.2% accuracy and 40.3% F1 score on Steroid-M. MPP improves both sensitivity and specificity, which is important in clinical settings to balance correctly identifying actual anomalies while minimizing false positives. In contrast, ICL and CoT generally underperform due to their lack of domain-specific guidance, i.e., ICL with GPT2 on Steroid-M yields only 28.2% accuracy and a negligible 0.2% F1 score. This underperformance highlights the importance of incorporating domain knowledge, as MPP does, to improve model performance for specialized tasks like clinical anomaly detection.

### 7.6.2 t-SNE Representation of Embeddings

Fig. 7.6 shows the cluster formation in the embedding space of the LLM output which represents the distinct latent patterns captured by each prompting method. Across all models, the MPP forms well-defined clusters, indicating that it consistently produces more structured and distinct embeddings compared to the other prompting methods. This suggests that MPP effectively captures relevant patterns for anomaly detection in clinical data, outperforming the more dispersed clustering seen in ZS and ICL. The CoT also produces structured clusters, but MPP shows greater distinction and compactness, especially in LLaMA 2-7B and Mistral-7B, highlighting the efficacy of pathway-driven prompting.

**Pre-Prompt II:**

***Information:*** *Given the domain knowledge representing in form of structural and temporal graph explaining the steroid metabolism and the temporal evolution of metabolites respectively. In the structural graph, each node represents a steroid metabolite, and edge weights represent conversion rates between from one metabolite to another. In the temporal graph, each node represents a sample, and the edge weights represent the distances between the samples, showing how the steroid profiles change over time.*

***Task:*** *Generate a textual representation of this domain knowledge for the given longitudinal profile by extracting the biological information. Do not explain the graph structure.*

- *Include the detailed insights from the edge weights of the structural graph which represent the conversion rate between the two metabolites.*
- *Include the detailed insights from the node matrix of the temporal graph which represent the concentration level of different metabolites in every sample (node).*
- *Include the detailed insights from the edge weights of the temporal graph which represent the euclidean distance.*

*Structural Graph*

$E_S =$

| Node | 1 | A | Etio | Epi | T | 5aAdiol | 5bAdiol |
|---|---|---|---|---|---|---|---|
| 1 | 0.0 | 0.0 | 0.0 | 0.5 | 0.5 | 0.0 | 0.0 |
| A | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Etio | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Epi | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| T | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.5 | 0.5 |
| 5aAdiol | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 5bAdiol | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 |

*Temporal Graph*

$V_T =$

| Node | A | Etio | E | T | 5aAdiol | 5bAdiol |
|---|---|---|---|---|---|---|
| 1 | 2208.064 | 1237.146 | 16.851 | 29.344 | 48.518 | 122.957 |
| 2 | 2428.87 | 1360.864 | 18.534 | 26.406 | 48.518 | 135.251 |
| 3 | 2428.87 | 1113.434 | 16.851 | 29.344 | 48.518 | 122.957 |
| 4 | 2208.064 | 1113.434 | 16.851 | 29.344 | 48.518 | 122.957 |
| 5 | 1987.258 | 1360.864 | 15.162 | 29.344 | 48.518 | 135.251 |
| 6 | 2208.064 | 1113.434 | 16.851 | 26.406 | 53.37 | 110.662 |
| 7 | 2208.064 | 1237.152 | 18.534 | 32.275 | 43.667 | 110.662 |
| 8 | 1987.258 | 1113.434 | 16.851 | 29.344 | 53.37 | 122.957 |
| 9 | 1987.258 | 1360.864 | 18.534 | 29.344 | 43.667 | 135.251 |
| 10 | 3469.491 | 1922.573 | 17.402 | 27.341 | 70.835 | 132.211 |

$E_T =$

| Node | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.0 | 0.088 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 2 | 0.088 | 0.0 | 0.084 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 3 | 0.0 | 0.084 | 0.0 | 0.066 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 4 | 0.0 | 0.0 | 0.066 | 0.0 | 0.142 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 5 | 0.0 | 0.0 | 0.0 | 0.142 | 0.0 | 0.142 | 0.0 | 0.0 | 0.0 | 0.0 |
| 6 | 0.0 | 0.0 | 0.0 | 0.0 | 0.142 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 7 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.088 | 0.0 | 0.0 |
| 8 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.088 | 0.0 | 0.084 | 0.0 |
| 9 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.084 | 0.0 | 1.0 |
| 10 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 |

Fig. 7.4 Pre-Prompt II for Metabolism Pathway-driven Prompting. This prompt incorporates domain knowledge from structural and temporal graphs to improve contextual reasoning.

**Prompt:**

***Task:*** *Below is the information about the domain knowledge of the structural and temporal behaviour of the longitduinal steroid profile. Use this information and think again about the anomolous sample in the longitudinal profile. If an anomalous sample exists, return the sample number and provide an explanation for why it is considered anomalous. If no anomalies are found, state that the profile is clean. These insights provide a detailed biological understanding of the temporal changes and metabolic conversion pathways, offering clues about physiological states, potential anomalies, or patterns within the steroid profile over time.*

***Information:***
*The longitudinal profile of steroid metabolism provides significant biological insights, particularly when analyzing the conversion rates and temporal evolution of steroid metabolites.*

***Metabolism Pathway:***
*The structural graph offers a detailed view of the steroid metabolism, focusing on how metabolites are interconverted. The edge weights between metabolites represent the conversion rates, reflecting the intensity or likelihood of one metabolite converting into another. For example:*

- *The conversion between Testosterone (T) and both 5α-androstanediol (5aAdiol) and 5β-androstanediol (5bAdiol) has a conversion rate of 0.5, indicating that Testosterone is equally likely to be metabolized into these two derivatives.*
- *A notable conversion occurs between 5aAdiol and metabolite A with a conversion rate of 1.0, suggesting a very high or complete conversion of 5aAdiol into metabolite A.*
- *Additionally, 5bAdiol shows a weaker connection to Etiocholanolone (Etio) with a conversion rate of 0.1, implying a less significant metabolic conversion between these metabolites.*
- *No direct conversions are observed between other metabolites like Etio, Epiandrosterone (Epi), or metabolite A, indicating that these might represent metabolic endpoints or have indirect interactions not captured by the direct edge weights.*

***Temporal Evolution:***
*The temporal graph shows the evolution of steroid profiles across various samples, with Euclidean distances between nodes representing changes in metabolite concentrations over time. Larger distances between nodes indicate more substantial changes in steroid profiles between samples:*

- *The smallest distances are seen between samples 2 and 3 (0.0845) and samples 3 and 4 (0.0661), indicating minimal changes in the steroid profiles across these time points, reflecting stable or slightly fluctuating metabolism.*
- *In contrast, a significant change is observed between samples 9 and 10, with a distance of 1.0, pointing to a substantial shift in the steroid profile at this point in time, suggesting either a physiological response, external intervention, or an anomaly in metabolism.*
- *The temporal progression from sample 5 to 6 also shows a moderate shift with a distance of 0.1426, highlighting evolving metabolite levels.*
- *Metabolite A shows significant variation, starting at 2208.064 in sample 1 and rising to a peak of 3469.491 in sample 10. This indicates a substantial accumulation of metabolite A over time, suggesting it plays a key role in metabolic progression or reflects a stress response.*
- *Etio levels fluctuate as well, starting at 1237.146 in sample 1, dropping to 1113.434 across several samples, and rising to 1922.573 in sample 10, further indicating significant metabolic variation.*
- *5aAdiol levels remain relatively stable, around 48.518 across most samples, except for sample 6 (53.37) and sample 10 (70.835), suggesting minor but important variations in 5α-reduction activity.*
- *The concentration of Testosterone (T) exhibits some variation, particularly peaking at 32.275 in sample 7, but generally stabilizing around 29.344, implying consistent androgenic activity throughout most of the samples.*

Fig. 7.5 Prompt for Metabolism Pathway-driven Prompting. The prompt integrates domain knowledge from both the structural graph (representing metabolic conversion relationships) and the temporal graph (capturing progression and fluctuation of metabolites over time) to refine LLM reasoning. It guides the model to re-evaluate anomalous samples based on biologically grounded insights, such as conversion rates and temporal shifts in metabolite levels, enabling contextual understanding of potential anomalies in longitudinal steroid profiles.

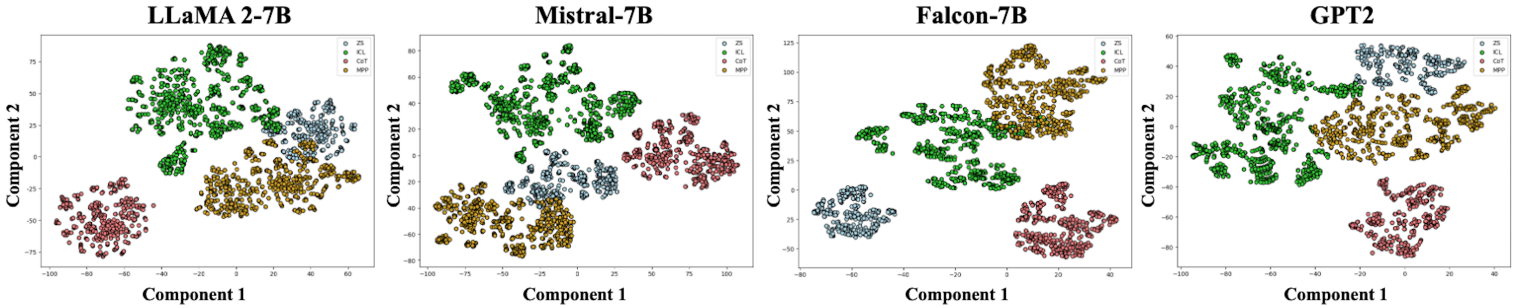


Fig. 7.6 t-SNE visualization of output embeddings from different prompting strategies across different LLMs. Each method produces a distinct cluster in the latent space, showing the influence of the prompting approach on the semantic structure of the generated representations. The MPP strategy shows better separation and coherence, indicating more structured and task-aligned output embeddings.

Table 7.1 Performance comparison of the proposed MPP method with baseline prompting strategies across different LLMs on Steroid-M and Steroid-F datasets. Non-LLM baselines, including IsoForest and $\beta$-VAE, are also reported. The MPP method consistently outperforms other methods across different metrics and models.

| **Model** | **Method** | **Steroid-M** | | | | **Steroid-F** | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **AC** | **SN** | **SP** | **F1** | **AC** | **SN** | **SP** | **F1** |
| LLaMA 2-7B | ZS | 0.652 | 0.912 | 0.563 | 0.403 | 0.402 | 0.567 | 0.382 | 0.250 |
| | ICL | 0.563 | 0.012 | 0.710 | 0.005 | 0.458 | 0.008 | 0.506 | 0.002 |
| | CoT | 0.228 | 0.526 | 0.130 | 0.208 | 0.426 | 0.506 | 0.381 | 0.250 |
| | **MPP** | **0.714** | **0.966** | 0.630 | **0.570** | **0.634** | **0.922** | 0.464 | **0.592** |
| Mistral-7B | ZS | 0.763 | 0.931 | 0.632 | 0.578 | 0.724 | 0.012 | 0.905 | 0.028 |
| | ICL | 0.834 | 0.920 | 0.753 | 0.677 | 0.506 | 0.026 | 0.636 | 0.009 |
| | CoT | 0.501 | 0.894 | 0.598 | 0.517 | 0.626 | 0.012 | 0.752 | 0.002 |
| | **MPP** | **0.895** | 0.928 | **0.882** | **0.808** | **0.758** | **0.356** | 0.893 | **0.198** |
| Falcon-7B | ZS | 0.352 | 0.960 | 0.125 | 0.364 | 0.395 | 0.308 | 0.474 | 0.406 |
| | ICL | 0.560 | 0.014 | 0.710 | 0.005 | 0.527 | 0.472 | 0.536 | 0.338 |
| | CoT | 0.524 | 0.673 | 0.432 | 0.440 | 0.388 | 0.024 | 0.383 | 0.008 |
| | **MPP** | **0.767** | 0.950 | **0.632** | **0.578** | **0.684** | **0.820** | 0.522 | **0.605** |
| GPT2 | ZS | 0.326 | 0.456 | 0.282 | 0.202 | 0.201 | 0.284 | 0.191 | 0.125 |
| | ICL | 0.282 | 0.006 | 0.355 | 0.002 | 0.229 | 0.004 | 0.253 | 0.001 |
| | CoT | 0.114 | 0.263 | 0.065 | 0.104 | 0.213 | 0.253 | 0.190 | 0.125 |
| | **MPP** | **0.357** | **0.483** | 0.315 | **0.285** | **0.317** | **0.461** | 0.232 | **0.296** |
| Non-LLM | IsoForest | 0.786 | 0.296 | 0.985 | 0.451 | 0.719 | 0.364 | 0.986 | 0.528 |
| | $\beta$-VAE | 0.752 | 0.006 | **0.992** | 0.012 | 0.681 | 0.002 | **0.994** | 0.004 |

## 7.7 Summary

This chapter introduced Metabolism Pathway-driven Prompting (MPP), a method designed to improve the reasoning capabilities of language models for detecting anomalies in structured biomedical data. Unlike standard prompting strategies, MPP incorporates prior domain knowledge by encoding metabolic and temporal dependencies into a prompt design, guiding the LLM's inference toward biologically plausible explanations. The method integrates both the metabolic pathway topology and temporal progression of biomarkers, enabling the LLM to generate outputs that are not only accurate in identifying anomalies but also aligned with biological interpretations. The approach was applied to real-world anti-doping datasets focused on steroid metabolism, where MPP demonstrated improved sensitivity and interpretability compared to standard prompting strategies.

MPP directly addresses RQ3, which concerns how anomaly detection models can be designed to support domain-informed reasoning and interpretability. The interpretability in LLM-based systems is achieved when the model's output is both accurate and grounded in structured knowledge that is understandable by domain experts. MPP operationalizes this principle by translating biological pathway graphs and biomarker trajectories into a structured natural language prompt format. Each prompt captures relevant interactions between biomarkers, such as enzyme-driven conversions, as well as temporal dynamics that may indicate manipulation or physiological anomalies. This design shifts the LLM's role from generic question answering to context-sensitive reasoning over structured domain content, enabling it to generate narrative explanations that reference specific metabolites, their interdependencies, and the time-based context in which anomalies arise.

MPP leverages structured input embeddings to inject graph-based representations into a natural language format that LLMs can process without modifying their architecture. This enables modularity and model-agnostic deployment, allowing it to work with different LLMs and adapt to various biological pathways. The reasoning generated by MPP reflects causal or correlative links across biomarkers, providing explanations such as, *"Testosterone and Epitestosterone decreased simultaneously, which is inconsistent with the expected ratio defined by the steroid biosynthesis pathway."* These structured prompts support users in verifying the model's rationale, increasing the reliability and practical applicability of the detection process.

In conclusion, MPP provides a prompting-based solution that extends the utility of LLMs for longitudinal clinical data analysis. By embedding domain knowledge into the prompt structure, it improves the reasoning quality of LLM outputs, addressing a key challenge in anomaly detection for high-stakes settings. This contributes toward building trustworthy AI

systems that are not only accurate but also usable by domain experts in forensic and clinical workflows.

# Chapter 8

# DAP: Digital Athlete Passport

## 8.1 Introduction

[1]Large sports events attract the attention of billions of people. Illegal performance enhancement by substances or methods can be traced back to the Olympic Games of Ancient Greece [18]. In modern times, the case of Lance Armstrong revealed massive doping in cycling and triggered investigations across multiple sports [64]. Consequently, the World Anti-Doping Agency (WADA) was founded to identify and prosecute athletes found guilty of doping [319]. From its beginning, anti-doping analytics has relied on methods inherited from biology and biochemistry, analyzing urine and blood samples collected during and beyond competition [178]. More recently, the success of machine learning has prompted investigations into its applicability for doping analytics [270]. Yet, a persistent challenge remains: decision-making in anti-doping is rarely grounded in absolute truth but instead in evidence-based assessments qualified by expert judgement. This makes transparency and interpretability central requirements for any computational system designed to support anti-doping investigations. During the Olympic Winter Games at Sochi, a subsequent report found that at least two female ice hockey players' samples were swapped with a urine sample containing male DNA, and others were found guilty of tampering with the original samples [193]. Urine swapping is the act of exchanging urine with another individual's or the athlete's stored clean urine to evade a positive test (WADA, 2020). More than 1000 athletes across 30 sports were involved in large-scale sample swapping at Sochi 2014. It was a massive program of cheating and cover-ups that has been running on an unprecedented scale since 2011 and will increase in future events [193]. This simple but new form of

[1]**Based on Publication:** Rahman, M.R., Piper, T., Geyer, H., Equey, T., Baume, N., Aikin, R., Maass, W. (2022). Data Analytics for Uncovering Fraudulent Behaviour in Elite Sports. *In Proceedings of the International Conference on Information Systems (ICIS 2022), Main Track*.

doping became a threat to the whole anti-doping decision-making organization. Current statistical methods are unable to reliably uncover such manipulations, raising questions about the potential of data-driven approaches to complement existing workflows.

As described in WADA's Technical Document TD2021EAAS [316], testing laboratories follow standardized procedures to measure steroid profiles using Gas Chromatography-Mass Spectrometry (GC-MS) [188]. Results are stored in the Anti-Doping Administration Management System (ADAMS), where an adaptive Bayesian model flags profiles for further review [270]. While effective in detecting substance-based doping, this approach is less reliable for detecting sample swapping, since deviations caused by substitution do not necessarily exceed statistical thresholds derived from population distributions. Suspicious cases require confirmatory procedures, such as GC/C/IRMS validation or DNA analysis [56], both of which are resource-intensive. In large events like the Olympic Games, thousands of samples are processed, and hundreds may be flagged as suspicious. Conducting DNA analysis on all of them is prohibitively costly and time-consuming, underscoring the need for complementary tools that can triage cases more efficiently. In recent years, data-driven methods have shown promise in healthcare and forensic sciences [262, 33]. This motivates their application to anti-doping, where models should not only achieve high performance but also provide interpretable outputs that experts can trust. Unlike purely predictive models, approaches in this domain should emphasize transparency, as inaccurate or poorly justified decisions can have severe consequences for both athletes and institutions.

This chapter introduces the Digital Athlete Passport (DAP), a methodology designed to detect suspicious cases of sample swapping while emphasizing interpretability. DAP combines statistical and machine learning methods with visualization techniques to flag potential anomalies and present intra-athlete profile similarities transparently. By doing so, DAP bridges the gap between automated detection and human-centered decision-making, providing expert users with visual evidence to validate predictions. The key contributions of this work are:

- A data-driven methodology is introduced for detecting sample swapping, highlighting interpretability and addressing limitations in existing detection methods.
- A model is developed that not only flags potential sample swapping incidents but also visualizes intra-athlete steroid profile similarity for better transparency and understanding.
- Comparative performance evaluation against baseline models demonstrates the effectiveness of the proposed approach in real-world anti-doping datasets.

## 8.2 Related Work

**Fraud Detection in Anti-Doping Analytics**

Doping activities can be classified into blood doping, steroid doping and sample swapping. Most of the data-driven research done until now mainly focuses on blood doping and steroid doping. For example, a study used different machine learning algorithms to detect the presence of doping substance erythropoietin in athletes' blood samples [230]. Some studies applied different machine learning algorithms with resampling techniques to find athletes at the highest risk of doping based on their performance data [137]. The Bayesian approach was also used for the detection of blood doping by using the interindividual performance data [200]. These studies mainly focused on blood doping. On the other hand, the literature on steroid doping includes the use of Support Vector Machine on the athlete's steroid profile to find how much a profile deviates from the normal population profiles [239]. Another study used machine learning algorithms such as Random Forest and XGBoost to predict abnormalities in steroid profiles [310]. The statistical method like Hotelling's $T_2$ test and Principal Component Analysis was also used to detect anomalous steroid profile [7]. All these works consider reference population data to define a clean profile and use different algorithms to find the deviation of anomalous steroid profile. However, this method cannot be used to detect sample swapping where only the samples collected from the same athlete should be considered for defining the clean profile. Therefore, there is a need to explore data-driven methods for the sample swapping problem.

However, there is no study performed so far on addressing the problem of sample swapping using a data-driven approach. The current SoTA method for finding the sample swapping is still the Bayesian method of the Adaptive Model, followed by laboratory testing [270]. Once triggered by the Adaptive Model, the confirmation tests like IRMS tests or subsequent DNA analysis of the athlete are performed to verify whether the sample is from the same athlete or is substituted by the athlete [223, 282]. However, these laboratory-based approaches are too expensive and cannot be implemented on all the athletes' samples during large athletic events like Olympic Games. Moreover, it is also a time-consuming process due to the fact that conducting each confirmation test requires a significant amount of time and resources. This is the reason why in most cases, unscrupulous athletes are caught after several months of the athletic event. This shows why there is a need to explore a new and more efficient method in this direction.

In this work, the proposed model provides a solution to this problem by creating a digital profile of an athlete. In this profile, the relatedness of all the athlete's samples can be visualized, and changes can be tracked as new samples are added to identify potential

cases of sample swapping. This method is very cost-effective and detects sample swapping in real-time. Therefore, this model can help the decision makers to flag the sample swapping cases during the athletic event and explain their decisions. In this way, it will reduce the number of laboratory-based testing needed and hence, cost and time beneficial.

## 8.3 Preliminaries

The aim of this study is to develop a method that can detect the sample swapping activity performed by the athlete. In addition, the model should trigger whether a new steroid profile matches with the steroid profile of previous samples of the athlete collected over time. So, a visualization tool or a quantification measure is needed that can show the relatedness among the steroid profiles of the same athlete. Therefore, a comprehensive analysis of steroid profiles has to be conducted to understand the underlying principles of different biomarkers.

## 8.4 Digital Athlete Passport (DAP)

Data visualization is an important concept in the data-driven approach [295]. It helps to explore data structure, detect outliers, identify trends/patterns or even interpret the result to gain information. Therefore, it is important to visualize the steroid samples in either two- or three-dimensional space. However, since the steroid profile consists of 11 parameters representing different biomarkers, the elevated values of any of these biomarkers can significantly impact the other biomarkers. Therefore, the steroid profile should be visualized in a space that takes into account all the parameters at the same time.

Let us consider a 3D space spanned by any three arbitrarily chosen parameters. A total of 165 different spaces will be required to fully visualize a steroid profile and capture all relevant aspects. Fig. 8.1 shows an example of the longitudinal steroid profile of an athlete with a testing sample $\mathbf{x}_T$ in 8 different spaces (out of 165 spaces) spanned by three different arbitrary chosen steroid parameters. Based on these plots, it is difficult to state whether the testing sample $\mathbf{x}_T$ belongs to the same athlete profile or from another athlete since there is no evidence of which space should be considered for decision making. Therefore, there is a need to find a visualization aid that incorporates the behavior of all the parameters together.

In this work, Digital Athlete Passport (DAP) is proposed as an effective approach for understanding the relatedness among steroid profiles and for providing a comprehensive visualization concept for these profiles. DAP incorporates Principal Component Analysis (PCA) and the concept of centroids to illustrate the similarity between steroid samples from an athlete. Since there is a linear relationship between the steroid parameters, PCA helps

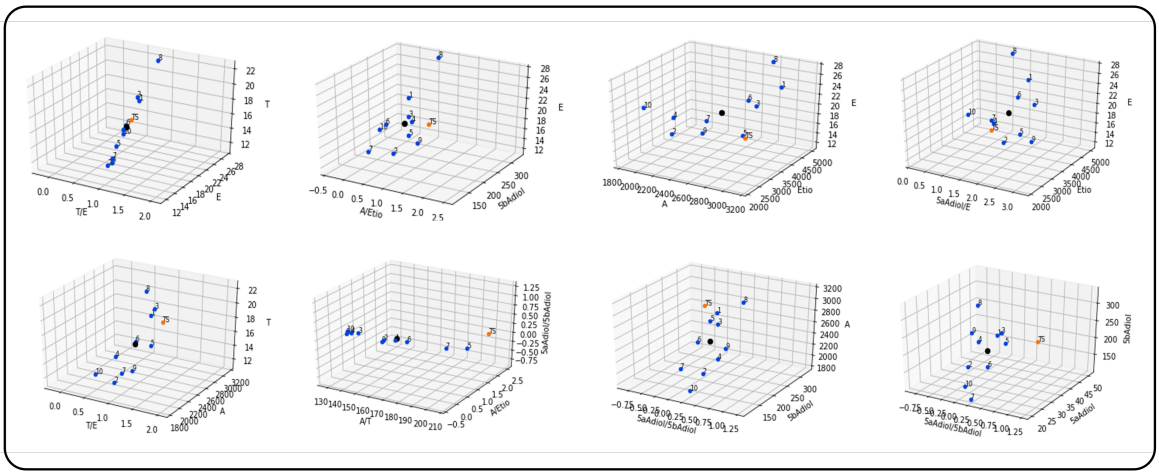

Fig. 8.1 Longitudinal steroid profile of an athlete visualized across 8 different 3D projections, each formed by selecting different combinations of steroid parameters. These plots helps to uncover patterns and identify anomalies in the steroid profiles over time, showing how different parameter combinations affect the structural representation of the athlete's metabolic trajectory.

to reduce correlated parameters to a smaller set of mutually-independent components that explain a large percentage of the covariance in the original steroid parameter space. Other dimensionality reduction algorithms, such as autoencoders, require large training datasets. Given that each athlete's longitudinal profile contains only 2 to 20 samples, they are not a suitable choice for this context. Moreover, PCA with the centroid approach also helps to solve the visualization problem by mapping the steroid sample from a multi-dimensional space to three-dimensional space and tracking the changes in the overall profile of the athlete when a new sample is added.

### 8.4.1 Principal Component Analysis

The PCA is an unsupervised learning technique [161] that projects the data into a new space spanned by a set of basis vectors such that the maximum amount of information is preserved in a lower number of basis vectors of the new space. The data is projected on these basis vectors called principal components, which are orthogonal unit vectors that maximise the variance in the data. The weights of each principal component represented by $w(k)$ is calculated by the following expression:

$$w(k) = \left\{ \frac{w^T \mathbf{X}_i(k)^T \mathbf{X}_i(k) w}{w^T w} \right\} \tag{8.1}$$

where $k = \{1,2,3\}$ and $\mathbf{X}_i$ refers to the longitudinal profile. The transformed profile $\mathbf{X}'_i$ consists of $\mathbf{x}'_{ij}(k) = \mathbf{x}_{ij} \cdot w(k)$. The profile is transformed in such a way that it contains the maximum variance in the first component, the second maximum variance in the second component and so on. In DAP, PCA is applied to transform $\mathbf{X}_i$ (consisting of 11 parameters) into a set of 3 principal components.

### 8.4.2 Centroid

The concept of Centroid or Center-of-Mass (CoM) is common in classical mechanics [141], which has a useful application in many domains. It refers to a unique point in the space where the weighted relative position of the distributed points sums to zero. This means that if different points are spanned in the space, the CoM is represented as the approximate center of all these points and can be calculated using the following expression:

$$\mathbf{x}_{i,CoM}(k) = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{x}'_{ij}(k), \quad k = \{1,2,3\} \tag{8.2}$$

where $\mathbf{x}_{i,CoM}(k)$ represents the centroid of all the transformed samples in the longitudinal profile with $k$ representing the three principal components of the transformed sample and $n_i$ represents the number of samples in longitudinal profile.

Whenever a new steroid sample is added to an athlete's profile, it is important that the relatedness of this sample to the previous samples be measured. This problem can be addressed by tracking the position of the CoM. If the new steroid sample is located far from the previous samples, a significant deviation in the position of the CoM will occur. Therefore, the variation in the position of the CoM can serve as a useful measure for monitoring consistency among steroid samples within a longitudinal profile.

### 8.4.3 DAP Algorithm

The longitudinal profile $\mathbf{X}_i$ of the athlete comprising all prior samples and the testing sample $\mathbf{x}_T$ to be checked for sample swapping is considered.

- *Step 1*: The three principal components for each steroid sample were calculated to visualize the longitudinal profile of the athlete in 3D space. Since at least three samples are required for the profile. For this analysis, only longitudinal profiles consisting of a minimum of three samples were considered. The PCA is performed on the first three steroid samples of the athlete's longitudinal profile after randomizing the order of the samples to remove any kind of bias. The calculated weights for each principal component are then used to transform the next steroid sample in the profile.

- *Step 2*: The CoM point is calculated based on the transformed samples by the principal components. This process is iterated until all the samples of the profile are considered, including $\mathbf{x}_T$. This excludes the collinearity between the original parameters and hence provides better values.

- *Step 3*: The three components of all the transformed samples are plotted in a 3D space, as this representation captures most of the variance in the athlete's longitudinal profile.

Fig. 8.2 presents a randomly selected athlete's longitudinal profile visualized in three arbitrarily chosen steroid parameter spaces (SPS), along with the corresponding transformed samples in the principal component space (PCS) after applying the DAP algorithm. The CoM (shown in black) represents the arithmetic center of all transformed samples. The position of the TS (in orange), relative to the CoM and the other samples, indicates the likelihood of whether $\mathbf{x}_T$ belongs to the same athlete profile. In this example, it can be observed in the PCS that $\mathbf{x}_T$ does not align with the same athlete profile, which is a distinction that was not evident in the SPS.

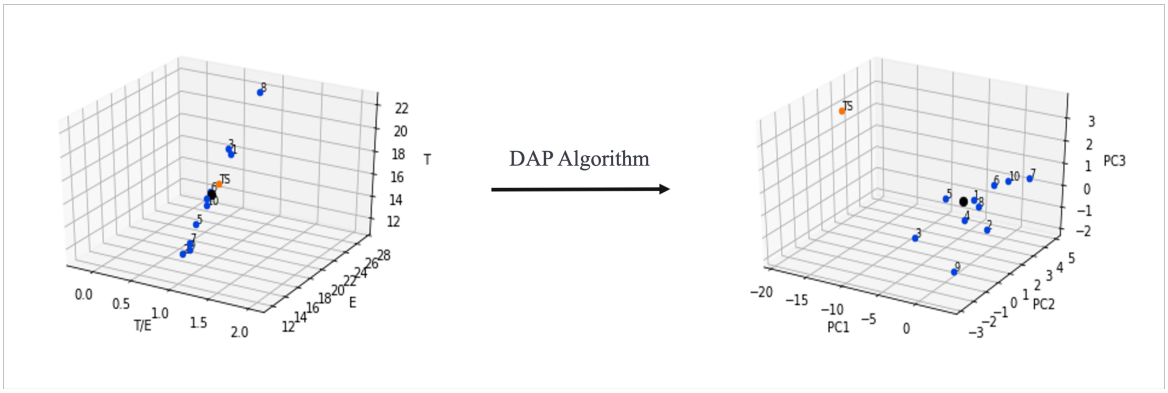


Fig. 8.2 Comparison of an athlete's longitudinal steroid profile before and after applying DAP algorithm. The left plot represents the steroid parameter space, where the profile is visualized using selected steroid biomarkers. The right plot shows the same profile transformed into principal component space after applying the DAP algorithm.

**Consecutive Distance** Understanding the change in the characteristics of an athlete's longitudinal profile upon the addition of a new steroid sample is important to note. In DAP, this is achieved by tracking the position of the CoM each time a new steroid sample is added. The underlying intuition is that greater similarity between the new sample and the existing samples will result in smaller deviations of the CoM. Therefore, the consecutive distances between CoM positions calculated after the addition of each new sample are computed. Euclidean geometry is applied for the distance computation using the following expression:

**Algorithm 1** Digital Athlete Passport algorithm

**Require:** $n_i \geq 3$
$\mathbf{X}_i \leftarrow \mathbf{x}_{i1}, \mathbf{x}_{i2}, \mathbf{x}_{i3}$
**for** $k \leq 3$ **do**
    $w(k) \leftarrow \text{PCA}(\mathbf{X}_i)$ ▷ calculate weights
    $k \leftarrow k+1$
**while** $j \neq n_i$ **do**
    **for** $k \leq 3$ **do**
        $\mathbf{x}'_{ij}(k) \leftarrow \mathbf{x}_{ij}(k) \cdot w(k)$
        $k \leftarrow k+1$
    **if** $j \geq 3$ **then**
        $x^j_{CoM}(k) \leftarrow \frac{1}{j}\sum_{m=1}^{j} \mathbf{x}'_{im}(k)$ ▷ centroid
        $d_j \leftarrow |x^j_{CoM}(k) - x^{j-1}_{CoM}(k)|$ ▷ consecutive distance
        $D_j \leftarrow D_j + d_j$ ▷ cumulative distance
        $j \leftarrow j+1$
    **else**
        $j \leftarrow j+1$

$$d_j = \sqrt{\sum_{k=1}^{3} \left(\mathbf{x}^j_{i,CoM}(k) - \mathbf{x}^{j-1}_{i,CoM}(k)\right)^2} \tag{8.3}$$

where $d_j$ represents the distance shifted by CoM when the $j^{th}$ sample is added to the profile, and $k$ represents the three components of the CoM in PCS. The value of $d_j$ is expected to be small when the new sample is similar to the previous samples in the longitudinal profile. However, if the new sample is from a different athlete, a significant spike in the value of $d_j$ is observed, indicating that the new sample does not belong to the same athlete profile. This distance is calculated for each sample added to the longitudinal profile and plotted against the number of samples in the profile.

**Cumulative Distance** The total distance deviated by the CoM after all samples are added to the profile is also calculated. This allows for tracking the extent to which the characteristics of the profile are affected by the addition of new samples. It is observed that as more samples are included in the longitudinal profile $\mathbf{X}_i$, the impact on the position of the CoM diminishes, and the value of $d_j$ begins to decrease, unless a suspicious sample from a different athlete is introduced. In such cases, a sudden spike is observed in the plot. The cumulative distance is computed using the following expression:

$$D_j = \sum_{j=3}^{n_i} d_j \tag{8.4}$$

Here, $D_j$ represents the cumulative distance calculated up to the addition of the $j^{th}$ sample. Since the CoM computation begins with the first three samples, the index $j$ starts at 3. This requirement arises because three components of the transformed sample are needed for the DAP algorithm.

**Contribution of Each Steroid Parameter** Understanding the contribution of each steroid parameter to the principal components is essential, as it helps determine the relevance of specific parameters in the decision-making process. Therefore, the feature importance of each steroid parameter within an athlete's longitudinal profile was calculated in PCS. The importance of each feature is reflected by the magnitude of its corresponding absolute values in the eigenvectors of $\mathbf{X}_i^T \mathbf{X}_i$, i.e., the larger the absolute value, the greater the feature's contribution to that principal component. In DAP, this calculation is performed separately for each longitudinal profile.

**Variance Captured by Each Component** Each principal component captures a specific proportion of variance in the longitudinal profile data. The proportion of total variance captured by the three principal components was calculated.

## 8.5 Experiments

### 8.5.1 Datasets

The data is extracted from the ADAMS database, which consists of real-world athlete data collected from 1 September 2018 until 31 March 2021 and called Steroid-All. This data contains 254,478 urine samples corresponding to 65,039 athletes, and each athlete could have between 2 and 20 samples in their profile. Table 8.1 shows the summary of the number of samples belonging to male and female athletes. For each athlete, only the raw steroid profile values, gender, competition type (i.e., whether tested during the competition (INC) or out of competition (OOC)), specific gravity of the sample (SG), and an anonymized athlete ID were extracted into an anonymized dataset, in accordance with the WADA International Standard for the Protection of Privacy and Personal Information (ISPPPI) [315].

The steroid profile of the urine samples consists of a set of biomarkers called steroid parameters that show significant changes in the administration of steroids. These parameters

Table 8.1 Summary of the number of longitudinal profiles and associated steroid profiles for male and female athletes in Steroid-All dataset.

| Athletes | Profiles | Samples |
|---|---|---|
| Male | 52,152 | 166,237 |
| Female | 12,887 | 88,241 |
| **Total** | **65,039** | **254,478** |

are testosterone (T), epitestosterone (E), etiocholanolone (Etio), androsterone (A), 5$\alpha$-androstanediol (5$\alpha$Adiol), and 5$\beta$-androstanediol (5$\beta$Adiol), and their ratios T/E, A/Etio, A/T, 5$\alpha$Adiol/5$\beta$Adiol, 5$\alpha$Adiol/E as described in TD2021EAAS [316].

### Data Pre-processing

**Missing Values** In the dataset, some samples were found to contain missing values, marked with 0 for all parameters. This issue could primarily be attributed to errors during data extraction from the ADAMS database. However, it is also possible that the test results were not reported or updated in ADAMS due to analytical issues encountered by the testing laboratory. Since the objective is to assess similarity among samples within an athlete's longitudinal profile, imputing missing values using data from other profiles or athletes is not a viable solution, as it could introduce bias into the profile. Therefore, all samples containing missing values were removed. Fig. 8.3 presents the data distribution of the raw dataset and the corresponding statistics after the removal of samples with missing values for both male and female athletes.

**Reference Ranges, LOQ and LOD** Each steroid parameter is expected to lie within a certain range. However, some samples were observed to contain abnormally high values for specific parameters. To address this, the observed values were compared against the maximum values reported by [291], and all samples containing off-values were removed.

Another issue is the Limit of Quantification (LOQ) refers to when the laboratory cannot quantify the concentration of the steroid parameter by GC-MS, and therefore, is reported as *-1*, whereas Limit of Detection (LOD) refers to when the chromatography peak signal of the parameter cannot be detected (i.e., is below the detection capability of the assay) and reported as *-2*. These values were replaced with the lowest concentration values measurable with an uncertainty not exceeding 30%, as specified in WADA Technical Document TD2021EAAS [316].

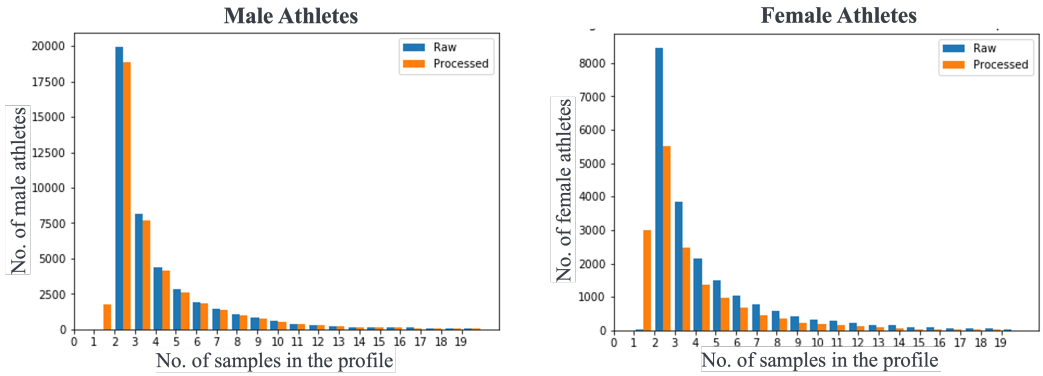


Fig. 8.3 Distribution of the number of samples per longitudinal profile for male (left) and female (right) athletes. The blue bars represent the raw dataset containing all recorded samples, while the orange bars represent the processed dataset after removing samples with missing values. This comparison shows the extent of data loss due to missing entries.

**Correction Due to Urinary Concentration** Not all the collected samples have the same concentration since some are more diluted than others. To compare the measured concentrations between different samples, the urinary concentrations need to be normalized using the urinary density. The concentration value of testosterone parameter of all the samples was corrected to a specific gravity of 1.020 as given by TD2021DL [314]:

$$T = \frac{1.020 - 1}{SG - 1} \times T_{\text{raw}} \tag{8.5}$$

where $T_{\text{raw}}$ represents the concentration value before the correction is applied and $SG$ represents the specific gravity of the sample. Similarly, the correction for A, Etio, E, 5$\alpha$Adiol and 5$\beta$Adiol was also applied. The steroid ratios are unaffected by the urinary specific gravity.

## 8.5.2 Descriptive Analysis

The distribution of all the steroid parameters of the samples was tested by using the 2-sample Kolmogorov-Smirnov test (K-S test) [65]. It is a standard test for deciding whether the two distributions are consistent with each other. The K-S test was performed to compare the distribution of samples collected during competition and out of competition for both male and female athletes. The $p$-value for each steroid parameter shows that there is a significant influence on the steroid profile of the samples due to the testing during the competition. In a recent laboratory study, is was found that there is a confounding factor due to the physical and mental stress on athletes, which causes a significant amount of elevated values in the

profile [223]. The statistical results show a similar change and thus are consistent with their findings.

Fig. 8.4 presents the statistical distribution of the testosterone parameter for both male and female athletes. It can be observed that testosterone values are more dispersed among male athletes compared to female athletes, resulting in lower inter-individual variance among female samples. The linear relationship was observed between the six steorid parameters and their ratio-based parameters. For example, T is directly proportional to the T/E ratio. As a result, collinearity among the steroid parameters was observed. The distributions of the steroid parameters were statistically described using the mean, median, and the first (Q1 = 0.25) and third (Q3 = 0.75) quartiles. Table 8.2 and 8.3 present the detailed descriptive statistics of the steroid parameters for male and female athletes, respectively.

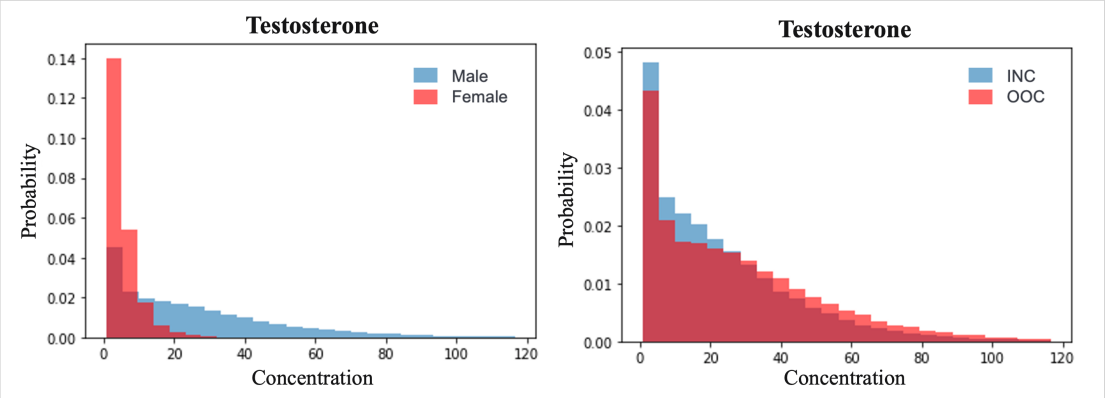


Fig. 8.4 Distribution of testosterone concentrations across different cohorts. The left plot compares the distribution between male (blue) and female (red) athletes, highlighting the generally higher values in male profiles. The right plot compares in-competition (INC) and out-of-competition (OOC) samples of male athletes, indicating the variations in testosterone levels potentially influenced by physiological or contextual factors.

### 8.5.3 Baseline Models

A set of baseline models was selected to compare the performance of the proposed model. These models are trained and optimized on the training dataset. The models include Logistic Regression (LR) [219], Support Vector Machine (SVM) [52], Random Forest [29], Gradient Boosting (XGB) [41] and Bayesian Method of Adaptive Model (SoTA) [270].

Table 8.2 Descriptive statistics of different steroid parameters for in-competition (INC) and out-competition (OOC) steroid samples from male athletes. The table reports the mean and standard deviation, minimum, interquartile range (IQ1 and IQ3), median, and maximum values for each parameter, along with the *p*-values from the K-S test assessing distributional differences between INC and OOC samples.

| **Parameter** | **INC (n = 79,653)** | | | | | | **OOC (n = 88,098)** | | | | | | ***p*-value** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | mean±sd | min | IQ1 | median | IQ3 | max | mean±sd | min | IQ1 | median | IQ3 | max | |
| A | 3188±1186 | 117 | 1925 | 2786 | 3798 | 27052 | 3203±1130 | 119 | 1624 | 2483 | 3247 | 24688 | 2.8e-14 |
| Etio | 2108±1101 | 41 | 1165 | 1785 | 3176 | 17700 | 2306±1275 | 31 | 1261 | 1740 | 3244 | 25949 | 2.0e-99 |
| E | 201±88.5 | 0.5 | 125 | 218 | 261 | 913 | 200±87.5 | 0.0 | 128 | 213 | 258 | 924 | 2.0e-01 |
| T | 28.4±18.6 | 0.5 | 12.5 | 21.3 | 38.1 | 148 | 30.0±22.5 | 0.0 | 12.1 | 20.2 | 38.2 | 172 | 3.8e-99 |
| 5$\alpha$Adiol | 26.1±16.7 | 0.0 | 12.1 | 20.7 | 33.8 | 148 | 27.0±18.2 | 0.0 | 11.2 | 19.9 | 34.2 | 172 | 3.8e-99 |
| 5$\beta$Adiol | 29.4±17.2 | 0.0 | 14.2 | 23.2 | 37.2 | 148 | 30.4±18.2 | 0.0 | 13.1 | 21.9 | 37.8 | 172 | 3.8e-99 |
| T/E | 1.41±1.18 | 0.0 | 0.7 | 1.1 | 1.6 | 9.6 | 1.47±1.23 | 0.0 | 0.7 | 1.1 | 1.7 | 10.0 | 6.0e-60 |
| A/Etio | 1.47±0.85 | 0.1 | 0.8 | 1.1 | 1.6 | 8.6 | 1.41±0.81 | 0.0 | 0.8 | 1.1 | 1.6 | 8.7 | 2.0e-99 |
| A/T | 0.17±0.13 | 0.0 | 0.1 | 0.1 | 0.2 | 1.2 | 0.16±0.12 | 0.0 | 0.1 | 0.1 | 0.2 | 1.2 | 3.8e-99 |
| 5$\alpha$Adiol/5$\beta$Adiol | 0.71±0.5 | 0.0 | 0.3 | 0.5 | 0.9 | 3.6 | 0.72±0.5 | 0.0 | 0.3 | 0.5 | 0.9 | 3.6 | 3.8e-99 |
| 5$\alpha$Adiol/E | 0.12±0.7 | 0.0 | 0.1 | 0.3 | 0.6 | 24.9 | 0.15±2.3 | 0.0 | 0.1 | 0.3 | 1.7 | 17.2 | 0.0 |

Table 8.3 Descriptive statistics of different steroid parameters for in-competition (INC) and out-competition (OOC) steroid samples from female athletes. The table reports the mean and standard deviation, minimum, interquartile range (IQ1 and IQ3), median, and maximum values for each parameter, along with the *p*-values from the K-S test assessing distributional differences between INC and OOC samples.

| **Parameter** | **INC (n = 34,820)** | | | | | | **OOC (n = 49,040)** | | | | | | ***p*-value** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | mean±sd | min | IQ1 | median | IQ3 | max | mean±sd | min | IQ1 | median | IQ3 | max | |
| A | 3233±1255 | 7.1 | 1408 | 1983 | 2476 | 27102 | 1977±1240 | 9.0 | 960 | 1179 | 1810 | 24922 | 2.8e-154 |
| Etio | 2178±1151 | 0.7 | 1240 | 1960 | 3174 | 21180 | 1387±1074 | 0.0 | 718 | 1077 | 1818 | 20940 | 2.0e-99 |
| E | 94.7±51.5 | 0.4 | 54.4 | 74.4 | 121 | 409 | 93.8±51.7 | 0.0 | 54.7 | 74.1 | 120 | 409 | 2.0e-01 |
| T | 9.3±7.5 | 0.1 | 4.4 | 7.4 | 12.9 | 48.0 | 10.0±8.2 | 0.0 | 4.2 | 7.2 | 13.2 | 47.2 | 3.8e-99 |
| 5$\alpha$Adiol | 7.6±6.1 | 0.0 | 3.2 | 5.9 | 10.2 | 48.0 | 8.1±6.7 | 0.0 | 3.1 | 5.7 | 10.7 | 47.2 | 3.8e-99 |
| 5$\beta$Adiol | 8.2±6.5 | 0.0 | 3.7 | 6.4 | 10.9 | 48.0 | 8.7±7.0 | 0.0 | 3.6 | 6.2 | 11.2 | 47.2 | 3.8e-99 |
| T/E | 1.08±1.01 | 0.0 | 0.4 | 0.8 | 1.1 | 6.1 | 1.06±1.01 | 0.0 | 0.4 | 0.8 | 1.1 | 6.1 | 6.0e-60 |
| A/Etio | 1.48±0.7 | 0.1 | 0.8 | 1.1 | 1.6 | 6.1 | 1.41±0.7 | 0.0 | 0.8 | 1.1 | 1.6 | 6.1 | 2.0e-99 |
| A/T | 0.17±0.13 | 0.0 | 0.1 | 0.1 | 0.2 | 1.2 | 0.16±0.12 | 0.0 | 0.1 | 0.1 | 0.2 | 1.2 | 3.8e-99 |
| 5$\alpha$Adiol/5$\beta$Adiol | 0.71±0.5 | 0.0 | 0.3 | 0.5 | 0.9 | 3.6 | 0.72±0.5 | 0.0 | 0.3 | 0.5 | 0.9 | 3.6 | 3.8e-99 |
| 5$\alpha$Adiol/E | 3.7±3.5 | 0.0 | 1.7 | 2.8 | 4.6 | 16.1 | 3.9±4.4 | 0.0 | 1.8 | 2.8 | 4.8 | 17.2 | 7.7e-11 |

### 8.5.4 Experimental Settings

The model provides a visualization aid to understand the similarity of the samples. Thus, the primary way to evaluate whether a sample belongs to the same athlete can be done by domain experts (human evaluation) after the application of the DAP algorithm on the longitudinal profile of the athlete.

An additional evaluation metric was proposed to quantify sample swapping using the DAP algorithm. Let the distances between the centroid $\mathbf{x}_{i,CoM}(k)$ and each sample in the longitudinal profile $\mathbf{X}_i$ be denoted as $d_1, d_2, \ldots, d_{n_i}$, and let the distance between the centroid and the testing sample $\mathbf{x}_T$ be $d_{\mathbf{x}_T}$. The mean ($\mu_{d_i}$) and standard deviation ($\sigma_{d_i}$) of all distances were calculated. The idea is to compare $d_{\mathbf{x}_T}$ with the distribution of $d_i$ to classify the testing sample $\mathbf{x}_T$ as an anomaly according to the following expression:

$$\text{Decision} = \begin{cases} \text{Anomalous}, & d_{\mathbf{x}_T} > \mu_{d_i} + 3\sigma_{d_i} \\ \\ \text{Clean}, & d_{\mathbf{x}_T} \leq \mu_{d_i} + 3\sigma_{d_i} \end{cases} \tag{8.6}$$

The decision rule of $3\sigma_{d_i}$ was selected after conducting a sensitivity analysis on the training dataset. The Steroid-All dataset was divided into a training set (80%) and a testing set (20%), as shown in Table 8.4. From the dataset, 50% of the profiles were randomly selected, and in each selected profile, the last sample was manually swapped with a sample from a different athlete. These modified profiles were labelled as swapped profiles (class 1), while the remaining 50% of the profiles were labelled as clean profiles (class 0). This procedure was implemented to simulate a scenario involving both swapped and clean cases for classification purposes. A sensitivity analysis was performed on the decision rule to assess its impact on various evaluation metrics. Table 8.5 presents the model's performance on the training dataset using different values for the decision rule: $1\sigma_{d_i}$, $1.5\sigma_{d_i}$, $2\sigma_{d_i}$, $3\sigma_{d_i}$, and $4\sigma_{d_i}$. The results indicate that the model performs best with a decision rule of $3\sigma_{d_i}$, achieving a high accuracy of 0.83 for male athletes and 0.74 for female athletes. The sensitivity and specificity values are also high, indicating that the model is effective in detecting swapped samples while maintaining a low false positive rate. The model's performance slightly decreases with higher thresholds, suggesting that the decision rule should be carefully selected to balance sensitivity and specificity.

For example, a sample swapping scenario was created in which the longitudinal profile of an athlete was arbitrarily selected, and $\mathbf{x}_T$ (taken from another athlete's longitudinal profile) was added to the profile. The DAP algorithm was then applied to this longitudinal profile. Fig. 8.5 presents the complete output of the DAP algorithm, including plots of the consecutive

Table 8.4 Data statistics of the training and testing sets used in the study, showing the number of longitudinal profiles and corresponding steroid samples for male and female athletes.

| **Athletes** | **Training** | | **Testing** | |
|---|---|---|---|---|
| | Profiles | Samples | Profiles | Samples |
| **Male** | 33,618 | 128,807 | 8,405 | 32,342 |
| **Female** | 12,572 | 67,498 | 3,144 | 16,762 |
| **Total** | 42,023 | 161,149 | 15,716 | 84,260 |

Table 8.5 Sensitivity analysis on the decision rule threshold using different standard deviation multiples added to the mean distance, showing its effect on different metrics for male and female athletes. This analysis helps evaluate the robustness of the model's classification performance with respect to threshold variations.

| **Athletes** | **Metrics** | $\mu_{d_i}+1\sigma_{d_i}$ | $\mu_{d_i}+1.5\sigma_{d_i}$ | $\mu_{d_i}+2\sigma_{d_i}$ | $\mu_{d_i}+3\sigma_{d_i}$ | $\mu_{d_i}+4\sigma_{d_i}$ |
|---|---|---|---|---|---|---|
| **Male** | AC | 0.78 | 0.78 | 0.79 | 0.83 | 0.75 |
| | SN | 0.82 | 0.75 | 0.72 | 0.72 | 0.56 |
| | SP | 0.75 | 0.87 | 0.85 | 0.89 | 0.94 |
| **Female** | AC | 0.72 | 0.73 | 0.72 | 0.74 | 0.67 |
| | SN | 0.71 | 0.63 | 0.62 | 0.61 | 0.40 |
| | SP | 0.74 | 0.81 | 0.84 | 0.88 | 0.93 |

distance, cumulative distance, feature contributions from each steroid parameter, and the variance explained by each principal component. The 3D plot illustrates that the position of $\mathbf{x}_T$ is distant from both the CoM and the other samples within the athlete's profile, allowing experts to easily recognize $\mathbf{x}_T$ as suspicious. Furthermore, a noticeable spike in both the consecutive and cumulative distances covered by the CoM upon the addition of $\mathbf{x}_T$ indicates that the sample does not belong to the same athlete's profile. The values of the proposed evaluation metric $d_{\text{TS}} = 39.3$, $d_{\text{A}} = 8.8$, and $\sigma_A = 4.5$ further support the conclusion that this is a swapped case. In practice, $\mathbf{x}_T$ usually appears at the end of $\mathbf{X}_i$, i.e., as the most recent $\mathbf{x}_{ij}$. However, this example demonstrates that the model functions independently of the testing sample's position, indicating that sample order does not affect detection performance.

## 8.6 Results

### 8.6.1 Performance Comparison

For the comparative study, Table 8.6 shows that the DAP model could differentiate the swapped $\mathbf{X}_i$ from the clean athlete's $\mathbf{X}_i$ based on the proposed decision rule. The model

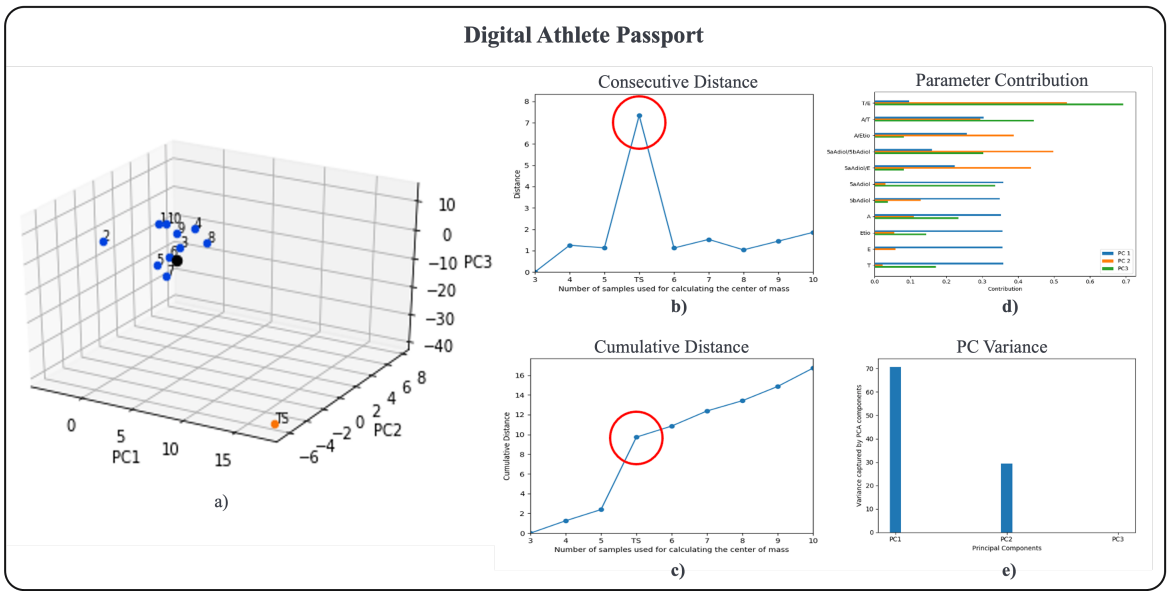


Fig. 8.5 Fully functional overview of the Digital Athlete Passport visualization for a selected athlete's $\mathbf{X}_i$. (a) Projection of $\mathbf{x}_{ij}$ into principal component space (PCS), showing temporal clustering and outlier detection; (b) pairwise distance between consecutive CoM across $\mathbf{x}_{ij}$ segments to detect abrupt shifts; (c) cumulative distance showing the trajectory and deviation magnitude over time; (d) contributions of each steroid parameter to the first three principal components, indicating their relevance in variance separation; (e) proportion of variance explained by each principal component, highlighting the dominance of PC1 in the data representation.

achieves an accuracy of 81% on $\mathbf{X}_i$ of male athletes. However, slightly lower performance was observed on $\mathbf{X}_i$ of female athletes. This is because the female athletes' $\mathbf{x}_{ij}$ have less inter-individual variance than the male athletes. The results show that the ensemble method like XGB shows comparable performance to the proposed model in terms of accuracy and specificity. Since the prevalence of sample swapping cases is very low (<1%) in real-life scenarios, high specificity values are important to minimize false positive cases (cost factor). Overall, the proposed model shows better performance than the Bayesian approach (SoTA) as well as all the baseline models in terms of different evaluation metrics.

## 8.7 Case Study

The DAP model was validated on two real-world sample swapping cases that had been confirmed through subsequent DNA analysis conducted by one of WADA's accredited laboratories. These longitudinal profiles contained more than one sample that did not originate from the same athlete. The DAP algorithm successfully flagged both longitudinal

Table 8.6 Performance comparison of the proposed DAP method against baseline models and the current SoTA approach.

| Metrics | Male | | | | | | Female | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **LR** | **SVM** | **RF** | **XGB** | **BM** | **DAP** | **LR** | **SVM** | **RF** | **XGB** | **BM** | **DAP** |
| AC | 0.75 | 0.73 | 0.78 | 0.80 | 0.76 | 0.81 | 0.68 | 0.71 | 0.75 | 0.76 | 0.71 | 0.77 |
| SN | 0.00 | 0.38 | 0.25 | 0.61 | 0.73 | 0.75 | 0.00 | 0.26 | 0.08 | 0.48 | 0.38 | 0.61 |
| SP | 0.89 | 0.88 | 0.92 | 0.93 | 0.82 | 0.92 | 0.84 | 0.82 | 0.88 | 0.90 | 0.85 | 0.89 |

profiles as sample swapping cases, as shown in Table 8.7. Additionally, the model was applied to longitudinal profiles that had been confirmed by the laboratory as positive steroid doping cases (i.e., cases involving the administration of exogenous steroids). In these cases as well, the model successfully identified the profiles as suspicious. The limited number of confirmed sample swapping cases available for evaluation (only two) reflects the low prevalence of such cases in real-world anti-doping investigations. Consequently, identifying such instances remains both a critical and challenging task in anti-doping analysis. These evaluations demonstrate that the proposed method yields promising results and could serve as a valuable enhancement to current approaches for detecting sample swapping cases.

Table 8.7 Evaluation of the DAP model on DNA-verified cases compared with the existing SoTA method. The table shows the percentage of confirmed profiles flagged as anomalous, highlighting DAP's high sensitivity to confirmed doping cases while slightly improving on the BM method's handling of clean profiles.

| **Cases** | **# Confirmed Profiles** | **% Flagged by BM** | **% Flagged by DAP** |
|---|---|---|---|
| Sample Swapping | 2 | 100% | 100% |
| Steroid Doping | 5 | 100% | 100% |
| Clean Profiles | 23 | 78% | 82% |

## 8.8 Summary

This chapter presented the concept of the Digital Athlete Passport (DAP), an interpretable and cost-effective anomaly detection framework designed to identify sample swapping in longitudinal steroid profiles. The motivation for DAP arises from the growing concern over fraudulent practices in elite sports, particularly the use of sample swapping to evade doping detection. While current verification techniques such as DNA testing provide conclusive evidence, they are financially and logistically demanding. DAP offers a data-driven alternative

that supports decision-makers by flagging suspicious profiles without relying on costly laboratory confirmation for every case.

DAP directly addresses RQ3, which concerns how anomaly detection systems can be designed to provide domain-informed reasoning and interpretability. In this framework, interpretability involves not only identifying anomalies but also explaining the rationale behind them in a manner consistent with biological expectations. DAP achieves this through a visual analytics pipeline that combines dimensionality reduction with distance-based evaluation. Steroid profiles are projected into a low-dimensional principal component space capturing physiological axes of variation, and each athlete's trajectory is modeled over time. New samples are assessed by their cumulative and consecutive distances from historical trajectories, enabling both anomaly detection and contextual reasoning. Interpretability is further improved by decomposing biomarker loadings on the principal components, which shows which steroid parameters contribute most to flagged deviations and how these shifts align with domain knowledge. For example, if a suspicious sample diverges due to a sharp T/E imbalance, DAP highlights this feature-level change and links it to potential doping behavior or physiological disruption. This explicit attribution, combined with trajectory visualization, provides transparency and supports a human-in-the-loop decision process, bridging algorithmic outputs with expert review.

Empirical evaluation on real-world sample swapping cases confirmed DAP's effectiveness, showing improved performance even without labeled training data. The framework outperformed baseline and domain-specific models by providing not just improved detection but also actionable interpretability. Because DAP relies on unsupervised methods and simple statistical rules, it is particularly suited for deployment in resource-constrained or time-sensitive contexts, such as during international sporting events.

In conclusion, DAP contributes an interpretable and resource-efficient solution for anomaly detection in anti-doping workflows. By integrating trajectory modeling, principal component analysis, and biomarker attribution into a unified visual analytics framework, DAP addresses the limitations of black-box models while advancing traceability and transparency. It exemplifies how visual analytics and statistical reasoning can be combined to meet the dual requirements of operational efficiency and scientific rigor in elite sports monitoring.

# Section V: Conclusion and Limitations

# Chapter 9

# Software Framework for Longitudinal Anomaly Detection

## 9.1 Introduction

The analysis of longitudinal clinical data plays an important role in clinical monitoring [6], where it supports decision-making in anti-doping. These datasets help practitioners to identify physiological anomalies that may indicate prohibited interventions [254]. In this thesis, anti-doping is treated as a key application domain, providing a concrete use case for demonstrating the utility of anomaly detection methods. However, it is important to note that anti-doping is not a separate user group; rather, it represents one of the real-world domains where experts require tools to interpret longitudinal profiles.

This chapter introduces CASPIAN, a software framework developed to apply anomaly detection in longitudinal clinical analysis. CASPIAN is designed as a practical utility for the domain experts (users), such as laboratory analysts and anti-doping regulatory authorities, who need transparent and interpretable decision support in evaluating individual profiles. Therefore, the purpose of this chapter is not to present new scientific contributions, but to demonstrate how the methodological advances of this thesis can be integrated into a usable software system that meets the practical needs of its users.

The framework has been designed around three user-driven requirements: (i) to provide robust anomaly detection in longitudinal profiles, (ii) to integrate domain knowledge, such as biochemical metabolism pathways, into anomaly validation, and (iii) to deliver interpretable outputs that can be reviewed and trusted by human experts. CASPIAN has been implemented as a full-stack software system with a graphical user interface that allows users to select models and visualize analytical outputs. Emphasizing configurability and modularity, the

system can be adapted to diverse clinical datasets while remaining accessible to non-technical experts. By applying the methodological contributions of this thesis in a practical software environment, CASPIAN bridges the gap between research and application. The following sections provide a detailed overview of its architecture and user-facing interface, with specific attention to how it can be deployed in the anti-doping domain.

## 9.2 CASPIAN Framework

As shown in Fig. 9.1, the CASPIAN software framework represents a multi-layered software system developed to address the analytical and interpretative challenges in anomaly detection within longitudinal clinical profiles. Structured around three conceptually distinct parallel layers, CASPIAN allows independent yet coordinated processing of input longitudinal profiles through modular, model-specific components. The system integrates these layers within a unified interface, allowing for simultaneous anomaly scoring, pathway-aware validation, and interpretive reporting. This layered approach not only supports end-to-end detection and explanation but also facilitates human-in-the-loop decision-making, ensuring that the outputs are both statistically grounded and clinically interpretable.

### 9.2.1 Longitudinal Anomaly Detection

The first layer focuses on detecting irregularities in longitudinal profiles based on their temporal structure.

- *SACNN*: This model captures global temporal dependencies using convolution and self-attention mechanisms. It is particularly effective for profiles with three or more samples and provides intra-individual modeling by learning profile-specific dynamics.
- *SCNN*: Designed for limited-sample profiles, SCNN can handle profiles with as few as two samples. It uses strategic subsampling and convolutional feature extraction to estimate trajectory-level deviations even under data constraints.

Both models serve as the first-pass detectors, flagging anomalies for further domain-based scrutiny.

### 9.2.2 Domain Knowledge Integration

Recognizing that statistical anomalies may not always correlate to biological implausibility, CASPIAN incorporates domain prior knowledge through two specialized models:

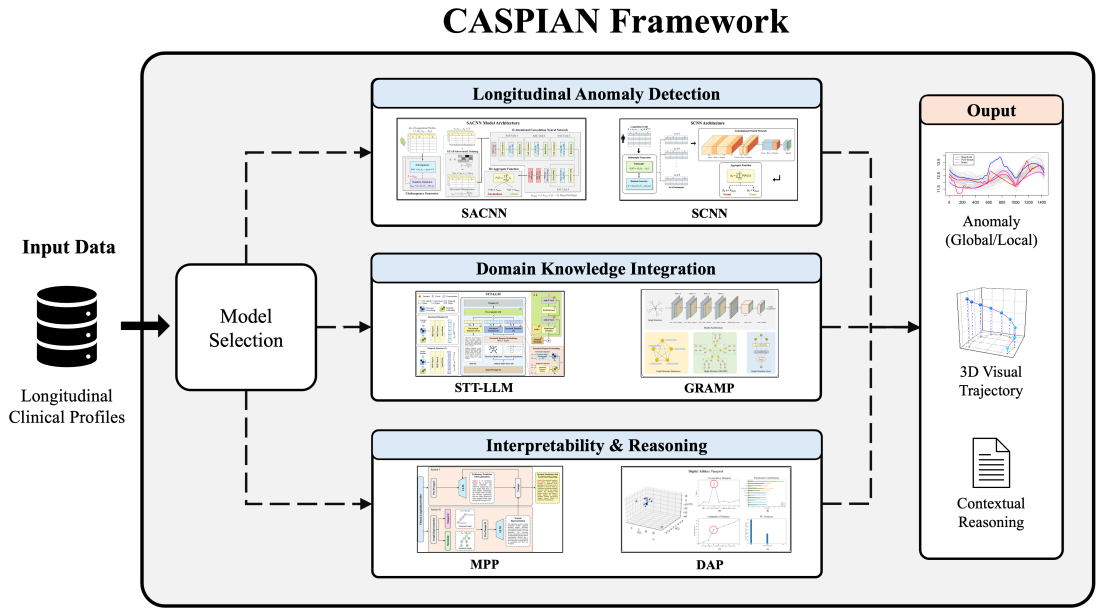


Fig. 9.1 Overall architecture of the CASPIAN framework designed for anomaly detection in longitudinal clinical profiles. It comprises three primary modules: (i) Longitudinal Anomaly Detection using models such as SACNN and SCNN, (ii) Domain Knowledge Integration through STT-LLM and GRAMP to incorporate metabolic and temporal pathway information, and (iii) Interpretability & Reasoning using MPP and DAP to improve transparency via visual trajectory and contextual textual explanations.

- *STT-LLM*: This model tokenizes longitudinal input into structured embeddings and leverages large language models to reason about profile evolution. It can provide zero-shot anomaly detection with temporal-biological awareness.
- *GRAMP*: This model integrates metabolic and biochemical interactions via graph neural networks. By modeling pathway-level constraints and reactions, GRAMP validates whether the detected anomalies violate expected metabolic behavior.

These models help contextualize anomalies, ensuring that flagged deviations are not just statistically significant but also biologically meaningful.

### 9.2.3 Interpretability & Contextual Reasoning

CASPIAN focuses on transparency and user interpretability by incorporating dedicated modules for explanation generation:

- *MPP*: This model generates domain-specific textual reasoning using biologically-grounded prompts. For each anomaly, it offers biochemical justifications and trend patterns.

- *DAP*: This model performs trajectory-based visual analytics. Using PCA projections and centroid deviation metrics, DAP shows how the individual profile diverges from population norms over time.

The interpretability layer closes the loop between model output and expert decision-making, supporting evidence-based anomaly validation.

## 9.3 Capabilities Across the Stack

The CASPIAN framework integrates a set of models, each designed to address specific computational and domain-level requirements for anomaly detection in longitudinal clinical data. The models span three functional layers, allowing the system to operate flexibly across various clinical and regulatory contexts. SACNN and SCNN form the foundation of the anomaly detection layer, offering complementary capabilities in sequence modeling. By leveraging structured attention mechanisms, SACNN is adapted to detect subtle yet meaningful deviations in profiles with higher temporal resolution. On the other hand, SCNN addresses a significant real-world challenge, i.e., detecting anomalies in limited data. Its ability to model trajectories with only two samples makes it particularly valuable in longitudinal monitoring settings where frequent sampling is not feasible. Both models contribute not only to anomaly localization but also to intra-individual trend modeling.

Beyond detection, CASPIAN's strength lies in its layered incorporation of domain knowledge and interpretability mechanisms, both of which are important for clinical validation and regulatory decision-making. The STT-LLM model bridges temporal structure with textual reasoning by reformulating metabolite pathway trajectories into structured prompts for large language models, offering an interpretable and generalizable reasoning layer. GRAMP introduces structured biological knowledge by modeling metabolic pathways as graphs, capturing dynamic temporal-pathway interactions. Together, these modules allow CASPIAN to evaluate the biological plausibility of detected anomalies rather than relying on statistical thresholds alone. On the interpretability level, MPP generates structured domain-specific textual explanations that help contextualize the anomaly, while DAP quantifies and visualizes deviations in trajectory space, providing experts with a visual diagnostic interface. These interpretability components are important for integrating CASPIAN into real-world clinical workflows, where trust, transparency, and expert oversight are important. The functional distribution across models, as shown in Table 9.1, reflects the framework's focus on modularity and domain alignment.

Table 9.1 Comprehensive comparison of different models' capabilities within the CASPIAN framework for anomaly detection, domain knowledge integration, and interpretability.

| **Models** | **Longitudinal Anomaly Detection** | | | **Domain Knowledge Integration** | | **Interpretability & Reasoning** | |
|---|---|---|---|---|---|---|---|
| | Temporal Modeling | Handles Limited Profile | Intra-Individual Modeling | Metabolic / Biochemical Prior | Temporal-Pathway Interaction | Visual Trajectory Analysis | Textual Reasoning |
| SACNN | ✓ | ✓ | ✓ | | | | |
| SCNN | ✓ | ✓ | ✓ | | | | |
| STT-LLM | ✓ | | | ✓ | ✓ | | ✓ |
| GRAMP | | | | ✓ | | | |
| MPP | ✓ | | | ✓ | ✓ | | ✓ |
| DAP | | ✓ | ✓ | | | ✓ | |

## 9.4 User Interface and Workflow

CASPIAN is implemented as an interactive software platform that enables real-time exploration and interpretation of longitudinal clinical data. At the center of the framework is a user-facing dashboard that allows experts to assemble analysis pipelines tailored to their investigative or regulatory tasks (see Fig. 9.2). Rather than functioning as a monolithic black-box tool, CASPIAN provides modular components that can be configured to meet different requirements. For example, an anti-doping laboratory analyst might prioritize immediate anomaly detection and clear visual reporting to flag suspicious patterns in an athlete's biological profile, while an anti-doping regulatory authority may focus on transparent explanations and pathway-informed reasoning to support expert panel reviews and adjudication processes.

Each module within the interface has dedicated components that present key outputs in a clear and interpretable way. Anomaly detectors return classification results with confidence scores; pathway-based models display visual overlays of metabolic structures combined with model inferences; reasoning model generates structured textual explanations; and visualization model presents trajectory deviations and principal component contributions. This layered design supports both high-level summaries for operational decision-making and detailed drill-downs for expert validation. Therefore, the interface accommodates two modes of use: research-oriented exploration, where reproducibility and fine-grained analysis are emphasized, and operational deployment, where clarity and transparency are prioritized. By combining configurability with interpretability, CASPIAN provides a practical decision-support environment that can be applied across biomedical domains.

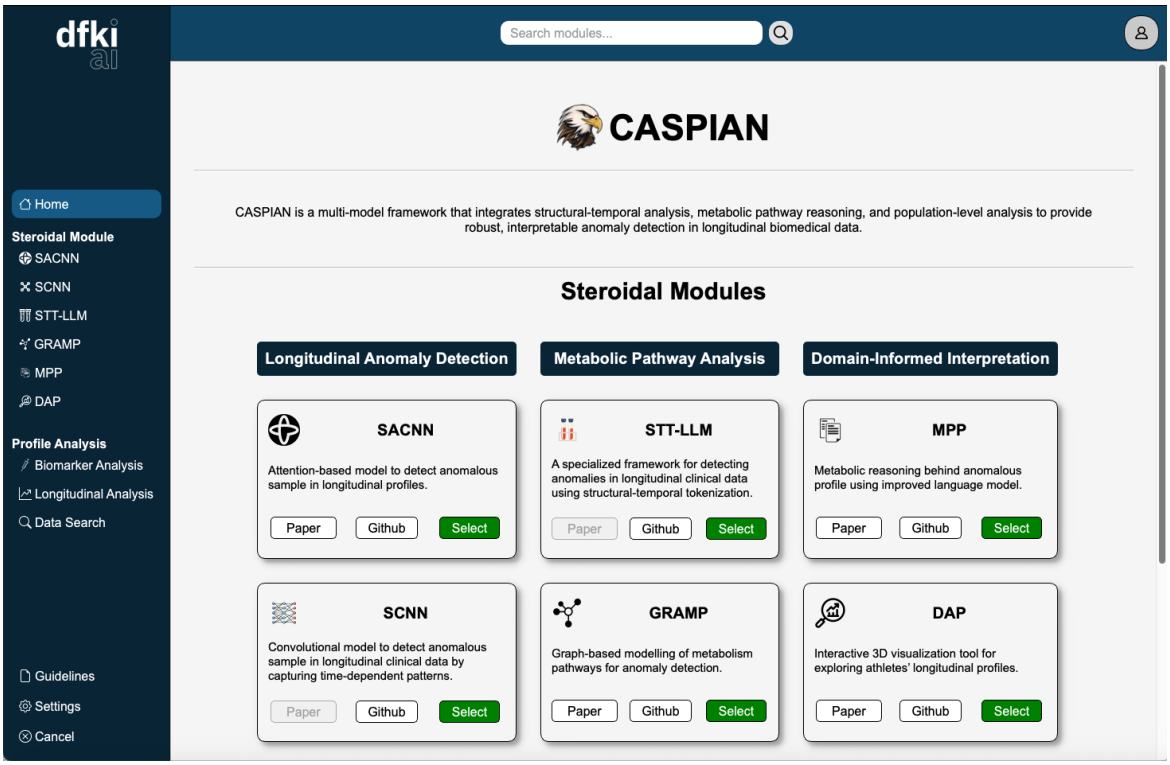


Fig. 9.2 User interface and functional components of the CASPIAN software, which is designed for detecting anomalies in longitudinal clinical data. The dashboard allows users to select specific models tailored to their individual needs and requirements.

## 9.5 Summary

This chapter presented the design and operational implementation of CASPIAN, a multi-model software framework developed to support biologically grounded and interpretable anomaly detection in longitudinal clinical profiles. By structuring the system into three modular and parallel analytical layers, CASPIAN addresses the key practical challenges outlined in this thesis. It offers a flexible analytical environment that supports diverse data contexts, including limited sample sizes, while maintaining fidelity to metabolic and clinical principles. From deep sequence modeling to graph-based biological validation and LLM-enabled explanations, each component of the framework is implemented to contribute meaningfully to the detection pipeline, ensuring that outputs are not only accurate but also interpretable and actionable. CASPIAN's implementation as an interactive software platform transforms the underlying methodological contributions of this work into a usable decision-support system for anti-doping and the regulatory domain. Its dashboard interface, modular configuration, and model-specific visualizations enable both research exploration and expert-driven diagnostics. Whether applied to anti-doping investigations or biomarker trajectory

analysis, CASPIAN provides an end-to-end solution for identifying and explaining anomalies in complex longitudinal clinical data.

# Chapter 10

# Conclusion

This thesis addressed the problem of anomaly detection in longitudinal clinical data by developing models that are robust to temporal complexity, domain variability, and data imperfections. Longitudinal datasets are foundational to clinical monitoring [331] and anti-doping in sports [70]. Therefore, this work was motivated by three important limitations in the current state-of-the-art methods: i) lack of robustness to temporal structure, ii) limited integration of domain-specific knowledge, and iii) insufficient model interpretability. This work aimed to bridge these gaps through a multifaceted modeling approach that highlights personalized pattern learning, biologically grounded representations, and explainable outputs.

## Modeling Temporal Dynamics for Anomaly Detection

A fundamental challenge in longitudinal clinical analysis lies in detecting anomalous patterns when labeled anomaly samples are rare or unavailable. Direct supervised learning methods, which rely heavily on annotated datasets, are often unsuitable in this setting [266]. In anti-doping, obtaining explicit labels for anomalies is not only resource-intensive but can also be practically impossible, as anomalies may be subtle, evolve slowly over time, or only become apparent retrospectively. Therefore, there is a pressing need for models that can learn normal temporal dynamics and identify deviations without requiring labeled anomaly data. To address this problem, this thesis proposed two complementary models: the Self Attention-based Convolutional Neural Network (SACNN) and the Subsampling-based Convolutional Neural Network (SCNN). Both models are designed to operate in semi-supervised regimes, capable of learning from the structure of the longitudinal profiles themselves.

The SACNN model introduced a novel combination of convolutional feature extraction and self-attention mechanisms. The convolutional layers captured local structural patterns among biomarkers at each time point, while the self-attention layers learned global temporal

dependencies across the longitudinal sequence. This dual mechanism enabled SACNN to model both short-term variations and long-range trends within clinical profiles, regardless of the profile's length. In particular, the architecture allowed for the generation of structural-temporal embedding maps, preserving important temporal patterns that could signify subtle anomalous shifts even across long monitoring periods. Extensive experimental evaluations showed SACNN's effectiveness. When applied to real-world longitudinal steroid profiles of athletes, SACNN achieved an AUROC of 0.92 on the male athlete dataset (Steroid-M) and maintained a sensitivity of over 73% at a high specificity threshold of 99%. Compared to baselines such as Isolation Forest, $\beta$-VAE, and ensemble methods, SACNN consistently outperformed across all metrics. A key strength of SACNN lies in its robustness, i.e., it effectively handles longitudinal profiles of varying lengths and irregular sampling without reliance on explicit anomaly labels. However, SACNN's sophisticated architecture, particularly the use of multiple self-attention layers and convolutional blocks, resulted in a relatively high computational cost. While this complexity was beneficial for achieving better performance, it necessitated greater computational resources during both training and inference.

The SCNN model is developed as a computationally lighter alternative, which is attractive for resource-constrained environments due to its efficiency. SCNN retains much of SACNN's temporal modeling capabilities while significantly reducing model size and complexity. It introduces a subsampling-based approach, generating multiple subsequences from the available data and aggregating the learned representations through convolutional layers. In experimental validations, SCNN achieved strong performance, maintaining an AUROC above 0.81 on different longitudinal datasets and achieving a sensitivity exceeding 50% even when applied to limited profiles containing only two samples (Steroid-M$_{lim}$ and Steroid-F$_{lim}$ datasets). Although SCNN's performance was slightly lower than SACNN's, it shows remarkable efficiency and flexibility. The principal advantage of SCNN lies in its lightweight design, enabling faster training and inference, with only a modest trade-off in predictive performance. However, because it used a simpler representation strategy, SCNN was somewhat less sensitive to subtle, long-range temporal anomalies compared to SACNN. Both models were evaluated on real-world longitudinal steroid profiles, demonstrating their ability to detect anomalies in a clinical context. The results highlighted the potential of these models to improve the detection of subtle deviations in longitudinal data, paving the way for more effective monitoring and intervention strategies in clinical practice.

The development of SACNN and SCNN directly addresses RQ1, which examines how anomalies in longitudinal clinical data can be modeled and detected despite data complexity like irregular sampling, heterogeneous profile lengths, and the scarcity of labeled anomaly data. Both models adopt a semi-supervised paradigm that learns individualized baselines from

within-subject histories rather than relying on population thresholds or annotated anomalies. SACNN addresses irregularity and heterogeneity by constructing structured subsequences and applying structural-temporal embeddings with attention-weighted convolution, thereby enabling the detection of subtle anomalies. In contrast, SCNN addresses sparsity and limited-sample scenarios by generating temporally ordered subsamples and learning implicit differential consistency through convolutional encoders, enabling meaningful detection even with only two samples. In both approaches, anomalies emerge not from direct supervision but as deviations from learned intra-individual trajectories. This ability to learn personalized normality and detect deviations without sufficient ground-truth anomaly labels is central to answering RQ1 and is especially critical in domains such as anti-doping, where labeled anomalies are scarce, costly, or unverifiable without invasive follow-up testing.

## Incorporation of Metabolism Pathway Structure into Anomaly Detection

In longitudinal clinical analysis, biological markers do not evolve independently; rather, they are intricately linked through underlying biochemical pathways such as metabolism, endocrine regulation, or immune signaling [301]. Ignoring these structural relationships can result in models that detect statistical anomalies without considering biological plausibility, leading to false positives or clinically irrelevant findings. Therefore, effectively incorporating domain-specific knowledge, such as metabolic pathway structures, into anomaly detection models is important for improving both performance and interpretability. To address this need, this thesis proposed two major contributions: Structural-Temporal Tokenization for Large Language Models (STT-LLM) and GRAph-based modeling for Metabolism Pathway (GRAMP). Both approaches were designed to leverage the inherent structure of biological systems to improve anomaly detection in longitudinal profiles.

STT-LLM introduced a new framework for adapting large language models to longitudinal clinical data by embedding metabolic and temporal structure into token representations. Unlike conventional LLM applications focused on natural language, this method developed a specialized tokenization mechanism that integrates both pathway-informed relationships and temporal progression of different biomarkers into transformer-compatible input sequences. The structural tokenizer captured connections among biomarkers based on known biochemical pathways, while the temporal tokenizer encoded the evolution of these markers over time. Together, these components enabled STT-LLM to model complex longitudinal profiles using pretrained language models without modifying their internal architecture. Experimental results showed that STT-LLM outperformed baseline LLMs in both zero-shot and few-shot anomaly detection tasks across multiple datasets. The model also demonstrated better generalization from limited examples and reduced training costs by employing lightweight

fine-tuning techniques. STT-LLM offers a promising way for integrating domain-specific structure into general-purpose models, combining biological grounding with computational efficiency.

Building on a different architectural paradigm, the GRAMP model embedded the metabolism pathway structure directly into a graph neural network. Each metabolite was modeled as a node, with directed edges representing enzymatic reactions in the steroid metabolism pathway. This graph-based representation enabled GRAMP to leverage known biochemical dependencies during training and inference, allowing the model to learn physiological patterns that extend beyond individual biomarker trajectories. GRAMP used graph attention mechanisms to dynamically weigh relationships among metabolites, learning to highlight interactions that were informative for detecting anomalies. When evaluated on longitudinal steroid datasets, GRAMP achieved AUROC scores of 0.91 for male and 0.85 for female athletes, outperforming baseline models that ignored domain structure. GRAMP not only improved accuracy but also improved biological specificity, i.e., flagged anomalies were more consistent with plausible metabolic disruptions, as verified through domain expert review and case study. The model's reliance on metabolic pathway knowledge represents both its strength and its limitations. While it enables high-fidelity anomaly detection when accurate graphs are available, GRAMP's applicability may be restricted in domains lacking well-defined biochemical maps or in cases where the biological processes are only partially understood.

Together, STT-LLM and GRAMP provide two complementary approaches to embedding biological knowledge into anomaly detection pipelines. These development directly addresses RQ2, which examines why the integration of domain knowledge is important for improving anomaly detection in longitudinal clinical data. By embedding relational structure into the learning process, both models demonstrated increased sensitivity to clinically meaningful deviations and reduced the incidence of biologically implausible false positives. STT-LLM achieves this through token-level encoding of structure and time into language models, while GRAMP uses graph-based learning to enforce biochemical constraints during message passing. These findings support the broader claim that embedding domain expertise into model design can significantly improve the reliability of anomaly detection systems in high-stakes clinical applications.

## Interpretability and Domain-Informed Reasoning for Anti-Doping

In domains such as anti-doping, where decisions based on anomaly detection can have significant ethical or legal consequences [61], interpretability is as important as predictive performance. Black-box models that flag anomalies without providing understandable justifi-

cations are insufficient for clinical or regulatory acceptance. Thus, a requirement for anomaly detection models in longitudinal clinical data is to produce outputs that are transparent and aligned with domain-expert reasoning. To meet this requirement, this thesis introduced two complementary contributions: the Metabolism Pathway-driven Prompting (MPP) framework and the Digital Athlete Passport (DAP) system. Both approaches were designed to support post-hoc interpretability by offering biologically grounded and visually accessible justifications for anomaly decisions, facilitating communication between automated systems and domain experts such as anti-doping officials and clinicians.

The MPP framework serves as a mechanism for generating human-readable explanations using large language models. MPP encodes prior biological knowledge into structured natural language prompts, guiding the pretrained LLM to produce context-sensitive reasoning of why a given longitudinal profile may be considered anomalous. Instead of providing only an anomaly score, MPP enables the model to reference relevant metabolic pathways, temporal patterns, and inter-marker relationships in its explanations. The framework was applied to various pre-trained LLMs, including LLaMA-2 and Mistral-7B, and evaluated through qualitative studies and expert feedback. Results showed that explanations generated by MPP were not only more informative than standard anomaly scores but also aligned with expert expectations regarding biochemical plausibility and temporal coherence. This interpretability layer helped bridge the gap between machine learning predictions and real-world clinical or regulatory reasoning. However, the quality of explanations depended on prompt design and the intrinsic reasoning capabilities of the underlying LLM, highlighting the importance of structured domain-informed prompting strategies.

In parallel, the Digital Athlete Passport (DAP) system was designed to provide a visual representation of profile evolution and to indicate the status of anomalies. DAP reduces high-dimensional longitudinal data into three principal components and overlays temporal data to construct a personalized centroid per individual. New samples are projected into this reduced space, and their distance from the centroid is used to assess anomaly likelihood. Unlike purely numerical outputs, DAP offers an intuitive visualization of how an individual's longitudinal profile is evolving relative to their own history, enabling experts to detect trends such as sudden spikes, gradual shifts, or emerging outliers. In evaluations on longitudinal steroid datasets, DAP achieved over 85% sensitivity while offering a graphical summary interpretable by users without extensive machine learning expertise. The system's independence from complex neural architectures and compatibility with standard clinical tools made it a practical solution for daily monitoring. However, as with all projection-based methods, the risk of information loss in dimensionality reduction remains, and subtle anomalies dispersed across many biomarkers may be underrepresented.

The combined development of MPP and DAP directly addresses RQ3, which examines how anomaly detection models can be made interpretable and suitable for decision support in anti-doping and clinical applications. MPP provides a textual explanation layer grounded in biological reasoning and enabling transparency in model predictions. DAP complements this by offering a visual summary of temporal deviations, facilitating intuitive analysis and expert validation. Together, these systems ensure that anomaly detection does not remain a technical exercise confined to algorithmic output but becomes a collaborative process where domain knowledge and machine learning insights are integrated to inform sensitive, high-impact decisions. Their successful application shows that interpretability in longitudinal anomaly detection is not only desirable but feasible, and it can be operationalized through a combination of prompt-based reasoning and human-centered visualization.

## CASPIAN Software Framework

The models developed throughout this work were integrated into a unified framework, CASPIAN. It adapts multiple complementary anomaly detection strategies, combining structural-temporal modeling, explanation generation, and visual trajectory analysis within a single system. It was designed in direct response to the challenges identified in this work. To this end, CASPIAN accommodates both high-complexity and lightweight components, allowing users to select appropriate models based on their computational resources and interpretability requirements.

CASPIAN offers a flexible pipeline where SACNN and SCNN provide the temporal anomaly detection backbone, GRAMP embeds domain-specific metabolic pathway structure, STT-LLM and MPP deliver interpretability through tokenized reasoning and prompt-based explanations, and DAP provides intuitive visual diagnostics. Through this modular architecture, CASPIAN operates as a practical decision-support system rather than a rigid pipeline, enabling domain experts to adapt workflows to their analytical needs. This flexibility makes it well-suited for longitudinal monitoring tasks in diverse fields, ranging from anti-doping to healthcare. By combining deep learning, domain knowledge, and transparent reasoning in one deployable system, CASPIAN demonstrates the feasibility of building anomaly detection frameworks that are not only technically robust but also practically usable in high-stakes decision-making contexts.

## 10.1 Limitations

Despite the substantial progress achieved through the models presented in this thesis, several limitations of this work require careful consideration. These limitations stem from both the inherent challenges of working with longitudinal clinical data and trade-offs in model design, domain integration, and system implementation. Recognizing these constraints is important for guiding future research and ensuring responsible deployment in real-world settings.

First, there are inherent limitations in the nature and scope of the available datasets. The primary datasets used in this thesis were derived from real-world longitudinal steroid profiles collected by anti-doping agencies and affiliated laboratories. While these datasets provided valuable ground for experimentation, they were largely composed of healthy, elite athlete populations. As a result, the demographic diversity and biological variability present in general clinical populations were underrepresented. This may limit the external validity and generalizability of the findings beyond the specific domain of anti-doping in sports. Additionally, the number of confirmed anomalous events in the data was limited, and in many cases, ground truth labels were inferred through expert agreement rather than direct clinical validation.

Second, limitations arise from architectural complexity and computational requirements. The SACNN model, while achieving SoTA performance in capturing structural and temporal dependencies, incurred substantial computational costs due to its multi-layered self-attention and convolutional design. Such resource demands may hinder deployment in low-power or latency-sensitive environments. Although SCNN was introduced to address these concerns through a lighter-weight architecture, it demonstrated a modest trade-off in performance. This highlights the classic trade-off between model expressiveness and efficiency, which remains a continuing challenge in longitudinal anomaly detection.

Third, the integration of biological pathway knowledge into modeling introduces specific challenges related to flexibility and generalizability. The STT-LLM framework, while offering a scalable and token-efficient method for incorporating structural and temporal information into large language models, depends on carefully engineered tokenization schemes that reflect accurate and consistent biological relationships. In domains where pathway information is noisy or sparsely annotated, the effectiveness of STT-LLM may be diminished, and its generalization across biological systems could be constrained. On the other hand, GRAMP embeds predefined metabolic pathways into graph neural networks, improving the detection of biologically plausible anomalies. However, its reliance on fixed and curated pathway structures introduces rigidity. In clinical scenarios where biochemical pathways are incomplete or dynamically evolving, GRAMP's applicability may be limited. Both

approaches highlight the trade-off between incorporating domain-specific structure and maintaining adaptability across diverse clinical contexts.

Finally, interpretability mechanisms also exhibit important limitations. The MPP framework improved the quality of explanations, but the reliability and consistency of generated outputs varied significantly across different LLMs and model configurations. This variability necessitates further research into explanation auditing and consistency checking. The DAP system, while visually intuitive and accessible to non-technical users, relies on dimensionality reduction methods such as PCA, which may obscure critical anomaly signals when these are dispersed across multiple low-dimensional spaces. Consequently, DAP may under-represent subtle but clinically important deviations, particularly in complex or high-dimensional biomarker landscapes.

## 10.2 Future Works

Building on the limitations identified in this thesis, several promising directions for future research can be outlined to improve the generalizability and real-world applicability of anomaly detection in longitudinal clinical analysis. A step-by-step trajectory of future work can be expected, moving from data and methodological improvements toward broader integration impact.

The first step lies in broadening the scope and diversity of datasets. The current study has focused primarily on longitudinal steroid profiles from real-world athletes, which limits the generalizability of findings to other populations. Future research should extend these approaches to diverse clinical cohorts, including patients with chronic illnesses, individuals across different age groups, and populations exposed to varying sociocultural and environmental conditions. Collaborative efforts to construct large-scale and expertly annotated longitudinal datasets would provide a foundation for standardized benchmarking and enable more reliable evaluation across heterogeneous settings. The next step is the optimization of model architectures for practical deployment. While the models developed in this thesis demonstrated strong performance, their computational requirements vary, and some may be unsuitable for resource-constrained environments. Future research could explore model compression techniques such as pruning and quantization, as well as adaptive architectures that dynamically adjust their complexity depending on data density and computational resources. Such developments would facilitate the deployment of anomaly detection systems in embedded clinical devices and anti-doping field laboratories.

Beyond efficiency, an important step is the deeper integration of biological domain knowledge. Current models incorporate pathway structure to improve plausibility, but

future research should explore patient-specific or dynamic pathway modeling approaches. These would reflect individual baselines more accurately, rather than relying on static, population-level networks. Advances in biomedical natural language processing could also be leveraged to automate the extraction and continuous updating of pathway information from scientific literature and databases, enabling models to evolve alongside expanding biomedical knowledge. Such personalization and adaptability would significantly increase the clinical relevance of anomaly detection outputs.

Improving interpretability remains another central direction. For visualization-based approaches, future work could explore more expressive methods such as manifold learning or graph-based trajectory representations, which may reduce information loss and better capture complex temporal structures. For language model-based explanations, future research should focus on improving explanation quality by fine-tuning models on domain-specific longitudinal datasets, integrating techniques such as contrastive or counterfactual explanation generation, and incorporating mechanisms for uncertainty quantification. These developments would strengthen the transparency and trustworthiness of anomaly detection models in clinical and regulatory applications.

Finally, future research should open the door to broader horizons by extending anomaly detection beyond its current boundaries. One avenue involves exploring causal anomaly detection frameworks, which would distinguish pathological changes from adaptations and enable simulation of counterfactual patient trajectories. Another route lies in multi-modal integration, combining laboratory biomarkers with data from wearables, imaging, and electronic health records to support real-time anomaly detection in continuous monitoring contexts. Extending these approaches beyond human clinical and anti-doping applications to domains such as veterinary health or environmental biosurveillance could also demonstrate the versatility and transferability of pathway-informed longitudinal anomaly detection.

# References


[1] Abdin, M., Aneja, J., Behl, H., Bubeck, S., Eldan, R., Gunasekar, S., Harrison, M., Hewett, R. J., Javaheripi, M., Kauffmann, P., and Zhang, Y. (2024). Phi-4 technical report.

[2] Agnieszka, D. and Magdalena, L. (2018). Detection of outliers in the financial time series using arima models. In *2018 Applications of Electromagnetics in Modern Techniques and Medicine (PTZE)*, pages 49–52. IEEE.

[3] Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. pages 2623–2631.

[4] Al-Tashi, Q., Saad, M. B., Muneer, A., Qureshi, R., Mirjalili, S., Sheshadri, A., and Wu, J. (2023). Machine learning models for the identification of prognostic and predictive cancer biomarkers: a systematic review. *International Journal of Molecular Sciences*, 24(9):7781.

[5] Alber, M., Buganza Tepole, A., Cannon, W. R., De, S., Dura-Bernal, S., Garikipati, K., Karniadakis, G., Lytton, W. W., Perdikaris, P., Petzold, L., and Kuhl, E. (2019). Integrating machine learning and multiscale modeling—perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences. *npj Digital Medicine*, 2(1):115.

[6] Albert, P. S. (1999). Longitudinal data analysis (repeated measures) in clinical trials. *Statistics in Medicine*, 18(13):1707–1732.

[7] Alladio, E., Caruso, R., Gerace, E., Amante, E., Salomone, A., and Vincenti, M. (2016). Application of multivariate statistics to the steroidal module of the athlete biological passport: a proof of concept study. *Analytica chimica acta*, 922:19–29.

[8] Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Goffinet, , Hesslow, D., Launay, J., Malartic, Q., Mazzotta, D., Noune, B., Pannier, B., and Penedo, G. (2023). The falcon series of open language models.

[9] Alqudah, A. M. and Moussavi, Z. (2025). A review of deep learning for biomedical signals: Current applications, advancements, future prospects, interpretation, and challenges. *Computers, Materials & Continua*, 83(3).

[10] Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., and McDermott, M. (2019). Publicly available clinical bert embeddings.

[11] Amer, M., Goldstein, M., and Abdennadher, S. (2013). Enhancing one-class support vector machines for unsupervised anomaly detection. In *Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description*, pages 8–15. ACM.

[12] Anandharaj, A. and Sivakumar, P. B. (2019). Anomaly detection in time series data using hierarchical temporal memory model. In *2019 3rd International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 1287–1292. IEEE.

[13] Antonelli, D., Bruno, G., and Chiusano, S. (2013). Anomaly detection in medical treatment to discover unusual patient management. *IIE Transactions on Healthcare Systems Engineering*, 3(2):69–77.

[14] Aoki, M. (2013). *State Space Modeling of Time Series*. Springer Science & Business Media.

[15] Apostol, E. S., Truică, C. O., Pop, F., and Esposito, C. (2021). Change point enhanced anomaly detection for iot time series data. *Water*, 13(12):1633.

[16] Bandara, K., Hyndman, R. J., and Bergmeir, C. (2025). Mstl: A seasonal-trend decomposition algorithm for time series with multiple seasonal patterns. *International Journal of Operational Research*.

[17] Baron, D. A. and Foley, T. (2009). Doping in sports. *Psychiatrike*, 20(4):336–341.

[18] Baron, D. A., Martin, D. M., and Abol Magd, M. S. (2007). Doping in sports and its spread to at-risk populations: An international review. *World Psychiatry: Official Journal of the World Psychiatric Association (WPA)*, 6(2):118–123.

[19] Barua, A., Muthirayan, D., Khargonekar, P. P., and Al Faruque, M. A. (2020). Hierarchical temporal memory-based one-pass learning for real-time anomaly detection and simultaneous data prediction in smart grids. *IEEE Transactions on Dependable and Secure Computing*, 19(3):1770–1782.

[20] Becchi, M., Aguilera, R., Farizon, Y., Flament, M. M., Casabianca, H., and James, P. (1994). Gas chromatography/combustion/isotope-ratio mass spectrometry analysis of urinary steroids to detect misuse of testosterone in sport. *Rapid communications in mass spectrometry RCM*, 8(4):304–308.

[21] Belyaeva, A., Cosentino, J., Hormozdiari, F., Eswaran, K., Shetty, S., Corrado, G., and Furlotte, N. A. (2023). Multimodal llms for health grounded in individual-specific data. In *Workshop on Machine Learning for Multimodal Healthcare Data*, pages 86–102, Cham. Springer Nature Switzerland.

[22] Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Podstawski, M., Gianinazzi, L., and Hoefler, T. (2024). Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 17682–17690.

[23] Blázquez-García, A., Conde, A., Mori, U., and Lozano, J. A. (2021). A review on outlier/anomaly detection in time series data. *ACM Computing Surveys (CSUR)*, 54(3):1–33.

[24] Bock, C., Aubet, F. X., Gasthaus, J., Kan, A., Chen, M., and Callot, L. (2022). Online time series anomaly detection with state space gaussian processes. *arXiv preprint arXiv:2201.06763*.

[25] Bolton, R. J. and Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science*, 17(3):235–255.

[26] 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.

[27] 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*.

[28] Bordes, A., Weston, J., Collobert, R., and Bengio, Y. (2011). Learning structured embeddings of knowledge bases. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pages 301–306. AAAI Press.

[29] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

[30] Callao, M. P. and Rius, A. (2003). Time series: A complementary technique to control charts for monitoring analytical systems. *Chemometrics and Intelligent Laboratory Systems*, 66(1):79–87.

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

[32] Callegari, C., Giordano, S., Pagano, M., and Pepe, T. (2012). WAVE-CUSUM: Improving cusum performance in network anomaly detection by means of wavelet analysis. *Computers & Security*, 31(5):727–735.

[33] Carriquiry, A., Hofmann, H., Tai, X., and VanderPlas, S. (2019). Machine learning in forensic applications. *Significance*, 16:29–35.

[34] CBS News / Associated Press (2011). Armstrong on doping claims: Never a failed test. https://www.cbsnews.com/news/armstrong-on-doping-claims-never-a-failed-test/. Accessed 2025-09-17.

[35] Chakrabarty, S., Talwadker, R., and Mukherjee, T. (2021). ScarceGAN: Discriminative classification framework for rare class identification for longitudinal data with weak prior. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM)*, pages 140–150. ACM.

[36] Chang, L.-W., Li, C.-T., Yang, C.-P., and Lin, S.-D. (2025). Learning on missing tabular data: Attention with self-supervision, not imputation, is all you need. *ACM Transactions on Intelligent Systems and Technology*, 16(3):1–24.

[37] Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., and Xie, X. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

[38] Chaovalitwongse, W. A., Fan, Y. J., and Sachdeo, R. C. (2007). On the time series $k$-nearest neighbor classification of abnormal brain activity. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 37(6):1005–1016.

[39] Chen, C.-F. R., Panda, R., Ramakrishnan, K., Feris, R., Cohn, J., Oliva, A., and Fan, Q. (2021). Deep analysis of cnn-based spatio-temporal representations for action recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6161–6171.

[40] Chen, J., Tan, X., Rahardja, S., Yang, J., and Rahardja, S. (2024). Joint selective state space model and detrending for robust time series anomaly detection. *IEEE Signal Processing Letters*. In press.

[41] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM.

[42] Chen, X., Deng, L., Zhao, Y., and Zheng, K. (2023). Adversarial autoencoder for unsupervised time series anomaly detection and interpretation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 267–275. ACM.

[43] Cheng, D., Xiang, S., Shang, C., Zhang, Y., Yang, F., and Zhang, L. (2020). Spatio-temporal attention-based neural network for credit card fraud detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01).

[44] Cheng, K. W., Chen, Y. T., and Fang, W. H. (2015). Gaussian process regression-based video anomaly detection and localization with hierarchical feature representation. *IEEE Transactions on Image Processing*, 24(12):5288–5301.

[45] Chirkova, N. and Troshin, S. (2023). Codebpe: Investigating subtokenization options for large language model pretraining on source code. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*.

[46] Cho, W., Kim, Y., and Park, J. (2019). Hierarchical anomaly detection using a multioutput gaussian process. *IEEE Transactions on Automation Science and Engineering*, 17(1):261–272.

[47] Choi, Y., Lim, H., Choi, H., and Kim, I. J. (2020). Gan-based anomaly detection and localization of multivariate time series data for power plant. In *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 71–74. IEEE.

[48] Christodoulou, V. and Bi, Y. (2015). A combination of cusum-ewma for anomaly detection in time series data. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–8. IEEE.

[49] Churová, V., Vyškovský, R., Maršálová, K., Kudláček, D., and Schwarz, D. (2021). Anomaly detection algorithm for real-world data and evidence in clinical research: Implementation, evaluation, and validation study. *JMIR Medical Informatics*, 9(5):e27172.

[50] Collins, L. M. (2006). Analysis of longitudinal data: The integration of theoretical model, temporal design, and statistical model. *Annual Review of Psychology*, 57(1):505–528.

[51] Constantinou, A. C., Guo, Z., and Kitson, N. K. (2023). The impact of prior knowledge on causal structure learning. *Knowledge and Information Systems*, 65(8):3385–3434.

[52] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

[53] Dagan, G., Synnaeve, G., and Roziere, B. (2024). Getting the most out of your tokenizer for pre-training and domain adaptation. In *Proceedings of the Forty-first International Conference on Machine Learning (ICML 2024)*.

[54] Das, B. C., Amini, M. H., and Wu, Y. (2025). Security and privacy challenges of large language models: A survey. *ACM Computing Surveys*, 57(6):1–39.

[55] de Hon, O., Kuipers, H., and van Bottenburg, M. (2015). Prevalence of doping use in elite sports: a review of numbers and methods. *Sports Medicine*, 45(1):57–69.

[56] De Wilde, L., Van Renterghem, P., Van Eenoo, P., and Polet, M. (2020). Development and validation of a fast gas chromatography combustion isotope ratio mass spectrometry method for the detection of epiandrosterone sulfate in urine. *Drug Testing and Analysis*, 12(8):1006–1018.

[57] DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., and team, D. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

[58] Deng, D. (2020). Research on anomaly detection method based on dbscan clustering algorithm. In *2020 5th International Conference on Information Science, Computer Technology and Transportation (ISCTT)*, pages 439–442. IEEE.

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

[60] Deutsche Welle (2025). Lance armstrong confesses to doping. https://www.dw.com/en/lance-armstrong-confesses-to-doping/a-16531718. Accessed 2025-09-17.

[61] Di Pietro, B., Verdi, F., Di Gianfrancesco, A., and Isidori, E. (2025). Legal and educational approaches to anti-doping: integrating compliance and ethical awareness. In *INTED2025 Proceedings*, pages 1877–1886. IATED.

[62] Diggle, P. J. (2002). *Analysis of Longitudinal Data*. Oxford University Press, Oxford, 2nd edition.

[63] Dilaveris, P. E. and Kennedy, H. L. (2017). Silent atrial fibrillation: Epidemiology, diagnosis, and clinical impact. *Clinical Cardiology*, 40(6):413–418.

[64] Dimeo, P. (2014). Why lance armstrong? historical context and key turning points in the 'cleaning up' of professional cycling. *The International Journal of the History of Sport*, 31(8):951–968.

[65] Dimitrova, D., Kaishev, V., and Tan, S. (2020). Computing the kolmogorov–smirnov distribution when the underlying cdf is purely discrete, mixed or continuous. *Journal of Statistical Software*, 95(10):1–42.

[66] Ding, N., Gao, H., Bu, H., and Ma, H. (2018a). Radm: Real-time anomaly detection in multivariate time series based on bayesian network. In *2018 IEEE International Conference on Smart Internet of Things (SmartIoT)*, pages 129–134. IEEE.

[67] Ding, N., Gao, H., Bu, H., Ma, H., and Si, H. (2018b). Multivariate-time-series-driven real-time anomaly detection based on bayesian network. *Sensors*, 18(10):3367.

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

[69] Donike, M., Ueki, M., Kuroda, Y., Geyer, H., Nolteernsting, E., Rauth, S., and Fujisaki, M. (1995). Detection of dihydrotestosterone (dht) doping: alterations in the steroid profile and reference ranges for dht and its 5 alpha-metabolites. *The Journal of Sports Medicine and Physical Fitness*, 35(4):235–250.

[70] Dragčević, D., Jakšić, V. P., and Jakšić, O. (2024). Athlete biological passport: longitudinal biomarkers and statistics in the fight against doping. *Archives of Industrial Hygiene and Toxicology*, 75(1):24.

[71] Dubin, J. A. and Müller, H.-G. (2005). Dynamical correlation for multivariate longitudinal data. *Journal of the American Statistical Association*, 100(471):872–881.

[72] Dudek, G. (2023). Std: A seasonal-trend-dispersion decomposition of time series. *IEEE Transactions on Knowledge and Data Engineering*, 35(10):10339–10350.

[73] Edwards, L. J. (2000). Modern statistical techniques for the analysis of longitudinal data in biomedical research. *Pediatric Pulmonology*, 30(4):330–344.

[74] Elbayad, M., Besacier, L., and Verbeek, J. (2018). Pervasive attention: 2D convolutional neural networks for sequence-to-sequence prediction. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 97–107, Brussels, Belgium.

[75] Elghazel, H., Deslandres, V., Kallel, K., and Dussauchoy, A. (2007). Clinical pathway analysis using graph-based approach and markov models. In *2007 2nd International Conference on Digital Information Management*, volume 1, pages 279–284. IEEE.

[76] Elliott, S. (2008). Erythropoiesis-stimulating agents and other methods to enhance oxygen transport. *British Journal of Pharmacology*, 154(3):529–541.

[77] Ellore, A., Mishra, S., and Hota, C. (2020). *Sequential Anomaly Detection Using Feedback and Prioritized Experience Replay*, pages 245–260.

[78] Erfani, S. M., Rajasegarar, S., Karunasekera, S., and Leckie, C. (2016). High-dimensional and large-scale anomaly detection using a linear one-class svm with deep learning. *Pattern Recognition*, 58:121–134.

[79] Evans, R. S. (2016). Electronic health records: Then, now, and in the future. *Yearbook of Medical Informatics*, Suppl 1:S48–S61.

[80] Feher, D., Minixhofer, B., and Vulić, I. (2024). Retrofitting (large) language models with dynamic tokenization.

[81] Feng, C. and Tian, P. (2021). Time series anomaly detection for cyber-physical systems via neural system identification and bayesian filtering. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2858–2867. ACM.

[82] 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*.

[83] Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2012). *Applied Longitudinal Analysis*. John Wiley & Sons, Hoboken, NJ, 2nd edition.

[84] Foorthuis, R. (2021). On the nature and types of anomalies: A review of deviations in data. *International Journal of Data Science and Analytics*, 12(4):297–331.

[85] Frattallone-Llado, G., Kim, J., Cheng, C., Salazar, D., Edakalavan, S., and Weiss, J. C. (2024). Using multimodal data to improve precision of inpatient event timelines. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 322–334.

[86] Friston, K. J., Harrison, L., and Penny, W. (2003). Dynamic causal modelling. *NeuroImage*, 19(4):1273–1302.

[87] Fuse, T. and Kamiya, K. (2017). Statistical anomaly detection in human dynamics monitoring using a hierarchical dirichlet process hidden markov model. *IEEE Transactions on Intelligent Transportation Systems*, 18(11):3083–3092.

[88] Garg, A., Zhang, W., Samaran, J., Savitha, R., and Foo, C. S. (2021). An evaluation of anomaly detection and diagnosis in multivariate time series. *IEEE Transactions on Neural Networks and Learning Systems*, 33(6):2508–2517.

[89] Gashi, M., Gursch, H., Hinterbichler, H., Pichler, S., Lindstaedt, S., and Thalmann, S. (2022). MEDEP: Maintenance event detection for multivariate time series based on the PELT approach. *Sensors*, 22(8):2837.

[90] Gatterer, K., Gumpenberger, M., Overbye, M., Streicher, B., Schobersberger, W., and Blank, C. (2020). An evaluation of prevention initiatives by 53 national anti-doping organizations: Achievements and limitations. *Journal of Sport and Health Science*, 9(3):228–239.

[91] Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1):3–42.

[92] Ghanvatkar, S. and Rajan, V. (2023). Graph-based patient representation for multimodal clinical data: Addressing data heterogeneity. medRxiv preprint.

[93] Gibbons, R. D., Hedeker, D., and DuToit, S. (2010). Advances in analysis of longitudinal data. *Annual Review of Clinical Psychology*, 6(1):79–107.

[94] Gledhill, N. (1982). Blood doping and related issues: A brief review. *Medicine and Science in Sports and Exercise*, 14(3):183–189.

[95] Glick, I. D. (2004). Undiagnosed bipolar disorder: New syndromes and new treatments. *Primary Care Companion to the Journal of Clinical Psychiatry*, 6(1):27–33.

[96] Goetz, L. H. and Schork, N. J. (2018). Personalized medicine: Motivation, challenges, and progress. *Fertility and Sterility*, 109(6):952–963.

[97] Goldman, O., Caciularu, A., Eyal, M., Cao, K., Szpektor, I., and Tsarfaty, R. (2024). Unpacking tokenization: Evaluating text compression and its correlation with model performance. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2274–2286.

[98] Graham, M. R., Davies, B., Grace, F. M., Kicman, A., and Baker, J. S. (2008). Anabolic steroid use: Patterns of use and detection of doping. *Sports Medicine*, 38:505–525.

[99] Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., and the LLaMA Team (2024). The llama 3 herd of models.

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

[101] Gu, M., Fei, J., and Sun, S. (2020). Online anomaly detection with sparse gaussian processes. *Neurocomputing*, 403:383–399.

[102] Gu, R., Wang, G., Song, T., Huang, R., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., and Zhang, S. (2021). CA-net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. *IEEE Transactions on Medical Imaging*, 40(2):699–711.

[103] Guk, K., Han, G., Lim, J., Jeong, K., Kang, T., Lim, E.-K., and Jung, J. (2019). Evolution of wearable devices with real-time disease monitoring for personalized healthcare. *Nanomaterials*, 9(6):813.

[104] Gunawardana, A., Meek, C., and Xu, P. (2011). A model for temporal dependencies in event streams. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 24.

[105] Güler, A. K., Alsan, H. F., and Arsan, T. (2025). Anomaly detection and performance analysis with exponential smoothing model powered by genetic algorithms and meta optimization. *IEEE Access*.

[106] Haab, J., Deutschmann, N., and Martínez, M. R. (2022). Is attention interpretation? a quantitative assessment on sets. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pages 303–321, Cham. Springer Nature Switzerland.

[107] 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)*.

[108] Han, Z., Gao, C., Liu, J., Zhang, J., and Zhang, S. Q. (2024). Parameter-efficient fine-tuning for large models: A comprehensive survey.

[109] Harerimana, G., Kim, J. W., and Jang, B. (2022). A multi-headed transformer approach for predicting the patient's clinical time-series variables from charted vital signs. *IEEE Access*, 10:105993–106004.

[110] Harris, S., Dowling, M., and Houlihan, B. (2021). An analysis of governance failure and power dynamics in international sport: the russian doping scandal. *International Journal of Sport Policy and Politics*, 13(3):359–378.

[111] Hasan, U., Hossain, E., and Gani, M. O. F. (2023). A survey on causal discovery methods for temporal and non-temporal data.

[112] Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K. L., Narang, A., Kumar, N., Han, Z., Guizani, M., Niyato, D., and Hussain, A. (2024). Interpreting black-box models: A review on explainable artificial intelligence. *Cognitive Computation*, 16(1):45–74.

[113] Hauskrecht, M., Batal, I., Valko, M., Visweswaran, S., Cooper, G. F., and Clermont, G. (2013). Outlier detection for patient monitoring and alerting. *Journal of Biomedical Informatics*, 46(1):47–55.

[114] He, K., Huang, S., and Qian, X. (2019). Early detection and risk assessment for chronic disease with irregular longitudinal data analysis. *Journal of Biomedical Informatics*, 96:103231.

[115] He, Y., Yan, T., Zhan, Y., Feng, Z., and Xia, Y. (2024). SGFM: Conditional flow matching for time series anomaly detection with state space models. *IEEE Internet of Things Journal*.

[116] Hedeker, D. and Gibbons, R. D. (2006). *Longitudinal Data Analysis*. John Wiley & Sons, Hoboken, NJ.

[117] Hegselmann, S., Buendia, A., Lang, H., Agrawal, M., Jiang, X., and Sontag, D. (2023). Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, pages 5549–5581. PMLR.

[118] Hejazi, M. and Singh, Y. P. (2013). One-class support vector machines approach to anomaly detection. *Applied Artificial Intelligence*, 27(5):351–366.

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

[120] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2017). beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*.

[121] Hoblos, J. (2025). Anomaly detection on univariate time series data using exponentially weighted moving average (anewma). In *Proceedings of the 10th International Conference on Internet of Things, Big Data and Security (IoTBDS 2025)*, pages 402–409. Science and Technology Publications, Lda.

[122] Homayouni, H., Ghosh, S., Ray, I., Gondalia, S., Duggan, J., and Kahn, M. G. (2020). An autocorrelation-based lstm-autoencoder for anomaly detection on time-series data. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 5068–5077. IEEE.

[123] Hou, J., Acharya, L., Zhu, D., and Cheng, J. (2016). An overview of bioinformatics methods for modeling biological pathways in yeast. *Briefings in Functional Genomics*, 15(2):95–108.

[124] Hou, N., Li, M., He, L., Xie, B., Wang, L., Zhang, R., Yu, Y., Sun, X., Pan, Z., and Wang, K. (2020). Predicting 30-days mortality for mimic-iii patients with sepsis-3: A machine learning approach using xgboost. *Journal of Translational Medicine*, 18:1–14.

[125] Huang, X., Zhang, F., Wang, R., Lin, X., Liu, H., and Fan, H. (2023). KalmanAE: Deep embedding optimized kalman filter for time series anomaly detection. *IEEE Transactions on Instrumentation and Measurement*, 72:1–11.

[126] Huang, Y., Mao, X., Guo, S., Chen, Y., Shen, J., Li, T., and Wan, H. (2025). Std-plm: Understanding both spatial and temporal properties of spatial-temporal data with plm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11817–11825.

[127] Huang, Z., Lu, X., and Duan, H. (2012). Anomaly detection in clinical processes. In *AMIA Annual Symposium Proceedings*, volume 2012, page 370. American Medical Informatics Association.

[128] Ibrahim, J. G. and Molenberghs, G. (2009). Missing data methods in longitudinal studies: A review. *Test*, 18(1):1–43.

[129] Iida, Y., Fukuda, K., Nishida, K., and Matsuo, Y. (2021). Tabbie: Pretrained representations of tabular data. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3446–3456.

[130] Jamshidi, E. J., Yusup, Y., Kayode, J. S., and Kamaruddin, M. A. (2022). Detecting outliers in a univariate time series dataset using unsupervised combined statistical methods: A case study on surface water temperature. *Ecological Informatics*, 69:101672.

[131] Jelkmann, W. and Lundby, C. (2011). Blood doping and its detection. *Blood: The Journal of the American Society of Hematology*, 118(9):2395–2404.

[132] Jeong, D. P., Garg, S., Lipton, Z. C., and Oberst, M. (2024). Medical adaptation of large language and vision-language models: Are we making progress?

[133] Jia, J., Gao, J., Xue, B., Wang, J., Cai, Q., Chen, Q., and Gai, K. (2025). From principles to applications: A comprehensive survey of discrete tokenizers in generation, comprehension, recommendation, and information retrieval.

[134] Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., and El Sayed, W. (2023). Mistral 7b.

[135] Jiang, J., Zhang, C., Ke, L., Hayes, N., Zhu, Y., Qiu, H., Zhang, B., Zhou, T., and Wei, G.-W. (2025). A review of machine learning methods for imbalanced data challenges in chemistry. *Chemical Science*, 16(18):7637–7658.

[136] Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J. Y., Shi, X., Chen, P.-Y., Liang, Y., Li, Y.-F., Pan, S., et al. (2023). Time-llm: Time series forecasting by reprogramming large language models.

[137] 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).

[138] Khan, M. A., Jang, J. H., Iqbal, N., Jamil, H., Naqvi, S. S. A., Khan, S., and Kim, D. H. (2025). Enhancing patient rehabilitation predictions with a hybrid anomaly detection model: Density-based clustering and interquartile range methods. *CAAI Transactions on Intelligence Technology*. In press.

[139] Kim, K., Park, J. H., Lee, M., and Song, J. W. (2022). Unsupervised change point detection and trend prediction for financial time-series using a new cusum-based approach. *IEEE Access*, 10:34690–34705.

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

[141] Kleppner, D. and Kolenkow, R. (1973). *An Introduction to Mechanics (2nd ed.)*. McGraw-Hill Publications.

[142] Klerman, E. B., Wang, W., Phillips, A. J., and Bianchi, M. T. (2017). Statistics for sleep and biological rhythms research: Longitudinal analysis of biological rhythms data. *Journal of Biological Rhythms*, 32(1):18–25.

[143] Knoth, S. and Schmid, W. (2004). Control charts for time series: A review. In Wilrich, P. T. and Schmid, W., editors, *Frontiers in Statistical Quality Control 7*, pages 210–236. Springer.

[144] Kobierecka, A. and Kobierecki, M. (2019). The negative implications of russia's doping scandal on the country's international image. *Eastern Review*, 8:161–182.

[145] Kölle, K., Biester, T., Christiansen, S., Fougner, A. L., and Stavdahl, Ø. (2019). Pattern recognition reveals characteristic postprandial glucose changes: Non-individualized meal detection in diabetes mellitus type 1. *IEEE Journal of Biomedical and Health Informatics*, 24(2):594–602.

[146] Kowsar, I., Rabbani, S. B., and Samad, M. D. (2024). Attention-based imputation of missing values in electronic health records tabular data. In *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, pages 177–182. IEEE.

[147] Kozitsin, V., Katser, I., and Lakontsev, D. (2021). Online forecasting and anomaly detection based on the arima model. *Applied Sciences*, 11(7):3194.

[148] Kreindler, D. M. and Lumsden, C. J. (2016). The effects of the irregular sample and missing data in time series analysis. In Guastello, S. J. and Gregson, R. D., editors, *Nonlinear Dynamical Systems Analysis for the Behavioral Sciences Using Real Data*, pages 149–172. CRC Press, Boca Raton, FL.

[149] Kumari, S. and Jayaram, B. (2016). Measuring concentration of distances—an effective and efficient empirical index. *IEEE Transactions on Knowledge and Data Engineering*, 29(2):373–386.

[150] Laird, N. M. (1988). Missing data in longitudinal studies. *Statistics in Medicine*, 7(1–2):305–315.

[151] Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974.

[152] Langfu, C., Zhang, Q., Yan, S., Yang, L., Wang, Y., Wang, J., and Bai, C. (2023). A method for satellite time series anomaly detection based on fast-dtw and improved-knn. *Chinese Journal of Aeronautics*, 36(2):149–159.

[153] Lazaridou, A., Kuncoro, A., Gribovskaya, E., Agrawal, D., Liska, A., Terzi, T., and Blunsom, P. (2021). Mind the gap: Assessing temporal generalization in neural language models. *Advances in Neural Information Processing Systems*, 34:29348–29363.

[154] Le Novere, N. (2015). Quantitative and logic modelling of molecular and gene networks. *Nature Reviews Genetics*, 16(3):146–158.

[155] Lee, C. K., Cheon, Y. J., and Hwang, W. Y. (2021). Studies on the gan-based anomaly detection methods for the time series data. *IEEE Access*, 9:73201–73215.

[156] Lee, E. C., Fragala, M. S., Kavouras, S. A., Queen, R. M., Pryor, J. L., and Casa, D. J. (2017). Biomarkers in sports and exercise: tracking health, performance, and recovery in athletes. *The Journal of Strength & Conditioning Research*, 31(10):2920–2937.

[157] Leeuwenberg, A. and Moens, M.-F. (2020). Towards extracting absolute event timelines from english clinical reports. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2710–2719.

[158] Lei, Z., Zhu, L., Fang, Y., Li, X., and Liu, B. (2020). Anomaly detection of bridge health monitoring data based on knn algorithm. *Journal of Intelligent & Fuzzy Systems*, 39(4):5243–5252.

[159] Leon-Lopez, K. M., Mouret, F., Arguello, H., and Tourneret, J.-Y. (2021). Anomaly detection and classification in multispectral time series based on hidden markov models. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11.

[160] Lesouple, J., Baudoin, C., Spigai, M., and Tourneret, J.-Y. (2021). Generalized isolation forest for anomaly detection. *Pattern Recognition Letters*, 149:109–119.

[161] Lever, J., Krzywinski, M., and Altman, N. (2017). Principal component analysis. *Nature*, 14:641–642.

[162] Li, D., Chen, D., Jin, B., Shi, L., Goh, J., and Ng, S.-K. (2019). MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks. In *International Conference on Artificial Neural Networks*, pages 703–716, Cham. Springer International Publishing.

[163] Li, J., Pedrycz, W., and Jamal, I. (2017). Multivariate time series anomaly detection: A framework of hidden markov models. *Applied Soft Computing*, 60:229–240.

[164] Li, P., Zhong, P., Mao, K., Wang, D., Yang, X., Liu, Y., Yin, J., and See, S. (2021a). Act: an attentive convolutional transformer for efficient text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15).

[165] Li, Y., Peng, X., Zhang, J., Li, Z., and Wen, M. (2021b). Dct-gan: Dilated convolutional transformer-based gan for time series anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(4):3632–3644.

[166] Liang, D., Gonen, H., Mao, Y., Hou, R., Goyal, N., Ghazvininejad, M., Zettlemoyer, L., and Khabsa, M. (2023). Xlm-v: Overcoming the vocabulary bottleneck in multilingual masked language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13142–13152.

[167] Lin, S., Clark, R., Birke, R., Schönborn, S., Trigoni, N., and Roberts, S. (2020). Anomaly detection for time series using vae-lstm hybrid model. In *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4322–4326. IEEE.

[168] Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). Isolation-based anomaly detection. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE.

[169] Liu, J., Yang, D., Zhang, K., Gao, H., and Li, J. (2023). Anomaly and change point detection for time series with concept drift. *World Wide Web*, 26(5):3229–3252.

[170] Liu, L., Yu, S., Wang, R., Ma, Z., and Shen, Y. (2024a). How can large language models understand spatial-temporal data?

[171] Liu, Q., Chen, B., Guo, J., Ziyadi, M., Lin, Z., Chen, W., and Lou, J. G. (2021a). Tapex: Table pre-training via learning a neural sql executor.

[172] Liu, X., Hu, J., Li, Y., Diao, S., Liang, Y., Hooi, B., and Zimmermann, R. (2024b). Unitime: A language-empowered unified model for cross-domain time series forecasting. In *Proceedings of the ACM Web Conference 2024 (WWW '24)*, pages 4095–4106.

[173] Liu, X., Yang, B., Liu, D., Zhang, H., Luo, W., Zhang, M., Zhang, H., and Su, J. (2021b). Bridging subword gaps in pretrain-finetune paradigm for natural language generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6001–6011.

[174] Liu, Z. and Hauskrecht, M. (2015). Clinical time series prediction: Toward a hierarchical dynamical system framework. *Artificial Intelligence in Medicine*, 65(1):5–18.

[175] Losi, E., Venturini, M., Manservigi, L., Ceschini, G. F., and Bechini, G. (2019). Anomaly detection in gas turbine time series by means of bayesian hierarchical models. *Journal of Engineering for Gas Turbines and Power*, 141(11):111019.

[176] Loy, C. C., Xiang, T., and Gong, S. (2011). Detecting and discriminating behavioural anomalies. *Pattern Recognition*, 44(1):117–132.

[177] Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., and Zhang, G. (2018). Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 31(12):2346–2363.

[178] Lu, Y., Yan, J., Ou, G., and Fu, L. (2023). A review of recent progress in drug doping and gene doping control analysis. *Molecules*, 28(14):5483.

[179] Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30. Curran Associates, Inc.

[180] Lundby, C., Robach, P., and Saltin, B. (2012). The evolving science of detection of 'blood doping'. *British Journal of Pharmacology*, 165(5):1306–1315.

[181] Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., and Liu, T.-Y. (2022). BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.

[182] Luo, Y., Liu, Z., Wang, L., Wu, B., Zheng, J., and Ma, Q. (2024). Knowledge-empowered dynamic graph network for irregularly sampled medical time series. *Advances in Neural Information Processing Systems*, 37:67172–67199.

[183] Ma, J. and Perkins, S. (2003). Time-series novelty detection using one-class support vector machines. In *Proceedings of the International Joint Conference on Neural Networks, 2003*, volume 3, pages 1741–1745. IEEE.

[184] Maennig, W. (2017). Major sports events: Economic impact. Technical Report 58, Hamburg Contemporary Economic Discussions.

[185] Malhotra, P., Vig, L., Shroff, G., and Agarwal, P. (2015). Long short term memory networks for anomaly detection in time series. In *Proceedings*, volume 89, page 94.

[186] Mandrikova, O., Fetisova, N., and Polozov, Y. (2021). Hybrid model for time series of complex structure with arima components. *Mathematics*, 9(10):1122.

[187] Mareck, U., Geyer, H., Fußhöller, G., Schwenke, A., Haenelt, N., Piper, T., and Schänzer, W. (2010). Reporting and managing elevated testosterone/epitestosterone ratios—novel aspects after five years' experience. *Drug Testing and Analysis*, 2(11-12):637–642.

[188] Mareck, U., Geyer, H., Opfermann, G., Thevis, M., and Schänzer, W. (2008). Factors influencing the steroid profile in doping control analysis. *Journal of Mass Spectrometer*, 43:877–891.

[189] 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., and Aquino Neto, F. R. (2005). Dna typing: an accessory evidence in doping control. *Journal of Forensic Science*, 50(3):587–592.

[190] Matarazzo, A. and Torlone, R. (2025). A survey on large language models with some insights on their capabilities and limitations.

[191] McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

[192] McGuigan, M. (2017). *Monitoring Training and Performance in Athletes*. Human Kinetics, Champaign, IL.

[193] McLaren, R. H. (2016). Mclaren report part i and ii: Independent investigation into state-sponsored doping in russia. Technical Report Part I (July 18, 2016); Part II (December 9, 2016), World Anti-Doping Agency (WADA), Montreal/Toronto. Commissioned by WADA; Part I published July 18, 2016; Part II published December 9, 2016.

[194] Minixhofer, B., Ponti, E., and Vulić, I. (2024). Zero-shot tokenizer transfer. In *Proceedings of the Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS 2024)*.

[195] Mohamed, S. K., Nounu, A., and Nováček, V. (2021). Biological applications of knowledge graph embedding models. *Briefings in Bioinformatics*, 22(2):1679–1693.

[196] Mohs, R. C., Schmeidler, J., and Aryan, M. (2000). Longitudinal studies of cognitive, functional and behavioural change in patients with alzheimer's disease. *Statistics in Medicine*, 19(11):1401–1409.

[197] Mok, W. Q., Wang, W., and Liaw, S. Y. (2015). Vital signs monitoring to detect patient deterioration: An integrative literature review. *International Journal of Nursing Practice*, 21:91–98.

[198] Monakhov, V., Thambawita, V., Halvorsen, P., and Riegler, M. A. (2023). Gridhtm: Grid-based hierarchical temporal memory for anomaly detection in videos. *Sensors*, 23(4):2087.

[199] Monostory, K. and Dvorak, Z. (2011). Steroid regulation of drug-metabolizing cytochromes p450. *Current Drug Metabolism*, 12(2):154–172.

[200] Montagna, S. and Hopker, J. (2018). A bayesian approach for the use of athlete performance data within anti-doping. *Frontiers in Physiology*, 9:884.

[201] Monteiro, M. S., Carvalho, M., Bastos, M. L., and Guedes de Pinho, P. (2013). Metabolomics analysis for biomarker discovery: advances and challenges. *Current Medicinal Chemistry*, 20(2):257–271.

[202] Moschini, G., Houssou, R., Bovay, J., and Robert-Nicoud, S. (2021). Anomaly and fraud detection in credit card transactions using the arima model. *Engineering Proceedings*, 5(1):56.

[203] Moston, S. and Engelberg, T. (2016). *Detecting Doping in Sport*. Routledge.

[204] Musen, M. A., Middleton, B., and Greenes, R. A. (2021). Clinical decision-support systems. In *Biomedical Informatics: Computer Applications in Health Care and Biomedicine*, pages 795–840. Springer International Publishing, Cham, 5 edition.

[205] Mørkeberg, J. (2013). Blood manipulation: Current challenges from an anti-doping perspective. *Hematology 2013, the American Society of Hematology Education Program Book*, 2013(1):627–631.

[206] Naumann, N., Walpurgis, K., Rubio, A., Thomas, A., Paßreiter, A., and Thevis, M. (2023). Detection of doping control sample substitutions via single nucleotide polymorphism-based id typing. *Drug Testing and Analysis*, 15(11-12):1521–1533.

[207] Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., and Mian, A. (2023). A comprehensive overview of large language models.

[208] Negro-Calduch, E., Azzopardi-Muscat, N., Krishnamurthy, R. S., and Novillo-Ortiz, D. (2021). Technological progress in electronic health record system optimization: Systematic review of systematic literature reviews. *International Journal of Medical Informatics*, 152:104507.

[209] Nguyen, L. H. and Goulet, J. A. (2019). Real-time anomaly detection with bayesian dynamic linear models. *Structural Control and Health Monitoring*, 26(9):e2404.

[210] Nikolentzos, G., Tixier, A., and Vazirgiannis, M. (2020). Message passing attention networks for document understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8544–8551.

[211] Niu, Z., Yu, K., and Wu, X. (2020). Lstm-based vae-gan for time-series anomaly detection. *Sensors*, 20(13):3738.

[212] Noroozizadeh, S., Weiss, J. C., and Chen, G. H. (2023). Temporal supervised contrastive learning for modeling patient risk progression. In *Machine Learning for Health (ML4H)*, pages 403–427.

[213] NPR (2016). Report: Russia used 'mouse hole' to swap urine samples of olympic athletes. Online; https://www.npr.org/sections/thetorch/2016/07/19/486595080/report-russia-used-mouse-hole-to-swap-urine-samples-of-olympic-athletes. Accessed 2025-09-17.

[214] Oh, M. and Zhang, L. (2022). Generalizing predictions to unseen sequencing profiles via deep generative models. *Scientific Reports*, 12(1):7151.

[215] O'Hara, L., Livigni, A., Theo, T., Boyer, B., Angus, T., Wright, D., and Freeman, T. C. (2016). Modelling the structure and dynamics of biological pathways. *PLoS Biology*, 14(8).

[216] Parvin, P., Chessa, S., Kaptein, M., and Paternò, F. (2019). Personalized real-time anomaly detection and health feedback for older adults. *Journal of Ambient Intelligence and Smart Environments*, 11(5):453–469.

[217] Patharkar, A., Cai, F., Al-Hindawi, F., and Wu, T. (2024). Predictive modeling of biomedical temporal data in healthcare applications: Review and future directions. *Frontiers in Physiology*, 15:1386760.

[218] 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.

[219] Peng, J., Lee, K., and Ingersoll, G. (2002). An introduction to logistic regression analysis and reporting. *Journal of Educational Research*, 96:3–14.

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

[221] Petkovski, A. and Shehu, V. (2023). Anomaly detection on univariate sensing time series data for smart aquaculture using k-means, isolation forest, and local outlier factor. In *2023 12th Mediterranean Conference on Embedded Computing (MECO)*, pages 1–5. IEEE.

[222] Pickering, T. G., Eguchi, K., and Kario, K. (2007). Masked hypertension: A review. *Hypertension Research*, 30(6):479–488.

[223] Piper, T., Geyer, H., Haenelt, N., Huelsemann, F., Schaenzer, W., and Thevis, M. (2021). Current insights into the steroidal module of the athlete biological passport. *International Journal of Sports Medicine*, 42:1–16.

[224] Plumb, J. O., Otto, J. M., and Grocott, M. P. (2016). 'blood doping' from armstrong to prehabilitation: Manipulation of blood to improve performance in athletes and physiological reserve in patients. *Extreme Physiology & Medicine*, 5(1):5.

[225] Pottgiesser, T., Sottas, P.-E., Echteler, T., Robinson, N., Umhau, M., and Schumacher, Y. O. (2011). Detection of autologous blood doping with adaptively evaluated biomarkers of doping: A longitudinal blinded study. *Transfusion*, 51(8):1707–1715.

[226] Pourahmadi, M. (2013). *High-Dimensional Covariance Estimation: With High-Dimensional Data*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, NJ.

[227] Provotar, O. I., Linder, Y. M., and Veres, M. M. (2019). Unsupervised anomaly detection in time series using lstm-based autoencoders. In *2019 IEEE International Conference on Advanced Trends in Information Theory (ATIT)*, pages 513–517. IEEE.

[228] Puder, A., Zink, M., Seidel, L., and Sax, E. (2024). Hybrid anomaly detection in time series by combining kalman filters and machine learning models. *Sensors*, 24(9):2895.

[229] Qin, Y. and Lou, Y. (2019). Hydrological time series anomaly pattern detection based on isolation forest. In *2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, pages 1706–1710. IEEE.

[230] Rahman, M., Bejder, J., Bonne, T. C., Andersen, A. B., Huertas, J. R., Aikin, R., Nordsborg, N. B., and Maass, W. (2022a). Detection of erythropoietin in blood to uncover doping in sports using machine learning. In *Proceedings of the IEEE International Conference on Digital Health (ICDH)*, pages 193–201.

[231] Rahman, M. R., Hussain, M., Piper, T., Geyer, H., Equey, T., Baume, N., Aikin, R., and Maass, W. (2023). Modelling metabolism pathways using graph representation learning for fraud detection in sports. In *2023 IEEE International Conference on Digital Health (ICDH)*, pages 158–168.

[232] Rahman, M. R., Khaliq, L. A., Piper, T., Geyer, H., Equey, T., Baume, N., Aikin, R., and Maass, W. (2024a). 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)*.

[233] Rahman, M. R., Liu, R., and Maass, W. (2024b). Incorporating metabolic information into llms for anomaly detection in clinical time-series. Time Series in the Age of Large Models: Workshop at NeurIPS 2024.

[234] Rahman, M. R., Piper, T., Geyer, H., Equey, T., Baume, N., Aikin, R., and Maass, W. (2022b). Data analytics for uncovering fraudulent behaviour in elite sports. In *Proceedings of International Conference on Information System (ICIS)*.

[235] Raiaan, M. A. K., Mukta, M. S. H., Fatema, K., Fahad, N. M., Sakib, S., Mim, M. M. J., and Azam, S. (2024). A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, 12:26839–26874.

[236] Ravì, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., and Yang, G.-Z. (2016). Deep learning for health informatics. *IEEE Journal of Biomedical and Health Informatics*, 21(1):4–21.

[237] Reardon, C. L. and Creado, S. (2014). Drug abuse in athletes. *Substance Abuse and Rehabilitation*, 5:95–105.

[238] Refaeilzadeh, P., Tang, L., and Liu, H. (2009). Cross-validation. In *Encyclopedia of Database Systems*. Springer, USA.

[239] 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.

[240] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM.

[241] Rindfleisch, A., Malter, A. J., Ganesan, S., and Moorman, C. (2008). Cross-sectional versus longitudinal survey research: Concepts, findings, and guidelines. *Journal of Marketing Research*, 45(3):261–279.

[242] 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.

[243] Rodriguez, M. A., Kotagiri, R., and Buyya, R. (2018). Detecting performance anomalies in scientific workflows using hierarchical temporal memory. *Future Generation Computer Systems*, 88:624–635.

[244] Ruan, Y., Lan, X., Ma, J., Dong, Y., He, K., and Feng, M. (2024). Language modeling on tabular data: A survey of foundations, techniques and evolution.

[245] Rudenko, V. (2014). Main modern problems of doping in sport. *Pedagogics, Psychology, Medical-Biological Problems of Physical Training and Sports*, 6.

[246] Rutkove, S. B. (2015). Clinical measures of disease progression in amyotrophic lateral sclerosis. *Neurotherapeutics: The Journal of the American Society for Experimental NeuroTherapeutics*, 12(2):384–393.

[247] Salamin, O., Kuuranne, T., Saugy, M., and Leuenberger, N. (2018). Erythropoietin as a performance-enhancing drug: Its mechanistic basis, detection, and potential adverse effects. *Molecular and Cellular Endocrinology*, 464:75–87.

[248] Sato, S., Sakuma, J., Yoshinaga, N., Toyoda, M., and Kitsuregawa, M. (2020). Vocabulary adaptation for domain adaptation in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4269–4279.

[249] Saugy, M., Lundby, C., and Robinson, N. (2014). Monitoring of biological markers indicative of doping: The athlete biological passport. *British Journal of Sports Medicine*, 48(10):827–832.

[250] Schiffer, L., Barnard, L., Baranowski, E. S., Gilligan, L. C., Taylor, A. E., Arlt, W., Shackleton, C. H. L., and 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.

[251] Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., and Langs, G. (2017). Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *Information Processing in Medical Imaging*, pages 146–157, Cham. Springer International Publishing.

[252] Schneider, P. and Xhafa, F. (2022). *Anomaly Detection and Complex Event Processing over IoT Data Streams: With Application to EHealth and Patient Data Monitoring*. Academic Press, Cambridge, MA.

[253] Schulsinger, F., Mednick, S. A., and Knop, J., editors (2012). *Longitudinal Research: Methods and Uses in Behavioral Science*, volume 1. Springer Science & Business Media, New York.

[254] Schüssler-Fiorenza Rose, S. M., Contrepois, K., Moneghetti, K. J., Zhou, W., Mishra, T., Mataraso, S., and Snyder, M. P. (2019). A longitudinal big data approach for precision health. *Nature Medicine*, 25(5):792–804.

[255] Sennrich, R., Haddow, B., and Birch, A. (2015). Neural machine translation of rare words with subword units.

[256] Shapley, L. S. (1953). A value for $n$-person games. In Kuhn, H. W. and Tucker, A. W., editors, *Contributions to the Theory of Games II*, volume 28 of *Annals of Mathematics Studies*, pages 307–317. Princeton University Press, Princeton, NJ.

[257] Shen, L., Li, Z., and Kwok, J. (2020). Timeseries anomaly detection using temporal hierarchical one-class network. *Advances in Neural Information Processing Systems*, 33:13016–13026.

[258] Shen, T., Zhou, T., Long, G., Jiang, J., Pan, S., and Zhang, C. (2018). Disan: Directional self-attention network for rnn/cnn-free language understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

[259] Shen, Y., Yu, J., Zhou, J., and Hu, G. (2025). Twenty-five years of evolution and hurdles in electronic health records and interoperability in medical research: Comprehensive review. *Journal of Medical Internet Research*, 27:e59024.

[260] Shukla, S. N. and Marlin, B. M. (2018). Modeling irregularly sampled clinical time series.

[261] Shumway, R. H. and Stoffer, D. S. (2017). Arima models. In *Time Series Analysis and Its Applications: With R Examples*, pages 75–163. Springer International Publishing, Cham, 4 edition.

[262] Sidey-Gibbons, J. and Sidey-Gibbons, C. (2019). Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*, 19.

[263] Sim, I., Gorman, P., Greenes, R. A., Haynes, R. B., Kaplan, B., Lehmann, H., and Tang, P. C. (2001). Clinical decision support systems for the practice of evidence-based medicine. *Journal of the American Medical Informatics Association*, 8(6):527–534.

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

[265] Singer, J. D. and Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford University Press, Oxford.

[266] Singh, A., Thakur, N., and Sharma, A. (2016). A review of supervised machine learning algorithms. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 1310–1315. IEEE.

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

[268] Sohal, H. S. (2025). Data anomalies and outlier detection in chronic kidney disease: A statistical perspective. In *2025 International Conference on Ambient Intelligence in Health Care (ICAIHC)*, pages 1–6. IEEE.

[269] Song, W., Shi, C., Xiao, Z., Duan, Z., Xu, Y., Zhang, M., and Tang, J. (2019). In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. ACM.

[270] Sottas, P.-E., Baume, N., Saudan, C., et al. (2007). Bayesian detection of abnormal values in longitudinal biomarkers with an application to t/e ratio. *Biostatistics*, 8:285–296.

[271] 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).

[272] Sottas, P.-E., Robinson, N., Rabin, O., and Saugy, M. (2011). The athlete biological passport. *Clinical Chemistry*, 57(7):969–976.

[273] Sottas, P.-E. and Vernec, A. (2012). Current implementation and future of the athlete biological passport. *Bioanalysis*, 4(13):1645–1652.

[274] Stehle, F. K., Vandelli, W., Zahn, F., Avolio, G., and Fröning, H. (2024). Deephydra: A hybrid deep learning and dbscan-based approach to time-series anomaly detection in dynamically-configured systems. In *Proceedings of the 38th ACM International Conference on Supercomputing*, pages 272–285. ACM.

[275] Su, Y., Zhao, Y., Niu, C., Liu, R., Sun, W., and Pei, D. (2019). Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2828–2837. ACM.

[276] Sun, G., Yin, C., Xia, T., Lu, Y., and Mao, J. (2024). An improved arima based anomaly detection method for time series data. In *2024 IEEE 8th Conference on Energy Internet and Energy System Integration (EI2)*, pages 5132–5138. IEEE.

[277] Sun, M., Zhou, K., He, X., Wang, Y., and Wang, X. (2022). Gppt: Graph pre-training and prompt tuning to generalize graph neural networks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1717–1727.

[278] Sun, Y., Yu, W., Chen, Y., and Kadam, A. (2019). Time series anomaly detection based on gan. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 375–382. IEEE.

[279] Tanner, R. and Gore, C. (2012). *Physiological Tests for Elite Athletes*. Human Kinetics, Champaign, IL.

[280] Thang, T. M. and Kim, J. (2011). The anomaly detection by using dbscan clustering with multiple parameters. In *2011 International Conference on Information Science and Applications*, pages 1–5. IEEE.

[281] Thevis, M., Geyer, H., Mareck, U., Sigmund, G., Henke, J., Henke, L., and Schänzer, W. (2007). Detection of manipulation in doping control urine sample collection: A multidisciplinary approach to determine identical urine samples. *Analytical and Bioanalytical Chemistry*, 388:1539–1543.

[282] 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.

[283] Thevis, M., Kohler, M., and Schänzer, W. (2008). New drugs and methods of doping and manipulation. *Drug Discovery Today*, 13(1-2):59–66.

[284] Thill, M., Konen, W., Wang, H., and Bäck, T. (2021). Temporal convolutional autoencoder for unsupervised anomaly detection in time series. *Applied Soft Computing*, 112:107751.

[285] Tipirneni, S. and Reddy, C. K. (2022). Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(6):1–17.

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

[287] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., and Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models.

[288] Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. pages 4489–4497.

[289] Tsikerdekis, M., Waldron, S., and Emanuelson, A. (2021). Network anomaly detection using exponential random graph models and autoregressive moving average. *IEEE Access*, 9:134530–134542.

[290] Van de Kerkhof, D. H., De Boer, D., Thijssen, J. H. H., and Maes, R. A. A. (2000). Evaluation of testosterone/epitestosterone ratio influential factors as determined in doping analysis. *Journal of Analytical Toxicology*, 24(2):102–115.

[291] Van Renterghem, P., van Eenoo, P., Geyer, H., Schänzer, W., and Delbeke, F. T. (2010). 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.

[292] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

[293] Veeravalli, B., Deepu, C. J., and Ngo, D. (2017). Real-time, personalized anomaly detection in streaming data for wearable healthcare devices. In Bessis, N. and Dobre, C., editors, *Handbook of Large-Scale Distributed Computing in Smart Healthcare*, pages 403–426. Springer.

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

[295] Vellido, A., Martín-Guerrero, J., Rossi, F., and Lisboa, P. (2011). Seeing is believing: The importance of visualisation in real-world machine learning applications. In *Proceedings of 19th European Symposium on Artificial Neural Networks*.

[296] Verbeke, G. (2000). Linear mixed models for longitudinal data. In *Linear Mixed Models in Practice: A SAS-Oriented Approach*, pages 63–153. Springer New York, New York, NY.

[297] Vinutha, H. P., Poornima, B., and Sagar, B. M. (2018). Detection of outliers using interquartile range technique from intrusion dataset. In *Information and Decision Sciences: Proceedings of the 6th International Conference on FICTA*, pages 511–518. Springer Singapore.

[298] von Rueden, L., Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., Kirsch, B., Walczak, M., Pfrommer, J., Pick, A., Ramamurthy, R., Rieck, B., and Schuecker, J. (2021). Informed machine learning – a taxonomy and survey of integrating prior knowledge into learning systems. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):614–633.

[299] Walker, W. H., Walton, J. C., DeVries, A. C., and Nelson, R. J. (2020). Circadian rhythm disruption and mental health. *Translational Psychiatry*, 10:28.

[300] Wang, J., Jiang, H., Liu, Y., Ma, C., Zhang, X., Pan, Y., and Zhang, S. (2024). A comprehensive review of multimodal large language models: Performance and challenges across different tasks.

[301] Wang, J., Maxwell, C. A., and Yu, F. (2019a). Biological processes and biomarkers related to frailty in older adults: a state-of-the-science literature review. *Biological Research for Nursing*, 21(1):80–106.

[302] Wang, L., Li, J., Bhatti, U. A., and Liu, Y. (2019b). Anomaly detection in wireless sensor networks based on knn. In *Artificial Intelligence and Security: 5th International Conference, ICAIS 2019, New York, NY, USA, July 26–28, 2019, Proceedings, Part III*, volume 11645 of *Lecture Notes in Computer Science*, pages 632–643. Springer International Publishing.

[303] Wang, S., Li, C., and Lim, A. (2021). A model for non-stationary time series and its applications in filtering and anomaly detection. *IEEE Transactions on Instrumentation and Measurement*, 70:1–11.

[304] Wang, W.-Y., Chan, T.-F., Peng, W.-C., Yang, H.-K., Wang, C.-C., and Fan, Y.-C. (2022). How is the stroke? inferring shot influence in badminton matches via long short-term dependencies. *ACM Transactions on Intelligent Systems and Technology*, 14(1):1–22.

[305] Wen, Q., Gao, J., Song, X., Sun, L., Xu, H., and Zhu, S. (2019). Robuststl: A robust seasonal-trend decomposition algorithm for long time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5409–5416.

[306] Wen, Q., Zhang, Z., Li, Y., and Sun, L. (2020). Fast robuststl: Efficient and robust seasonal-trend decomposition for time series with complex patterns. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2203–2213. ACM.

[307] Weng, Y. and Liu, L. (2018). A sequence anomaly detection approach based on isolation forest algorithm for time-series. In *International Workshop on High Performance Computing for Advanced Modeling and Simulation in Nuclear Energy and Environmental Science*, pages 198–207, Singapore. Springer Singapore.

[308] West, S. G. and Hepworth, J. T. (1991). Statistical issues in the study of temporal data: Daily experiences. *Journal of Personality*, 59(3):609–662.

[309] Wibisono, S., Anwar, M. T., Supriyanto, A., and Amin, I. H. A. (2021). Multivariate weather anomaly detection using dbscan clustering algorithm. In *Journal of Physics: Conference Series*, volume 1869, page 012077. IOP Publishing.

[310] 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.

[311] Wong, L., Liu, D., Berti-Equille, L., Alnegheimish, S., and Veeramachaneni, K. (2022). AER: Auto-encoder with regression for time series anomaly detection. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 1152–1161. IEEE.

[312] World Anti-Doping Agency (2005). Anti-doping administration management system (adams). https://www.wada-ama.org/en/what-we-do/adams. Accessed 2025-09-17.

[313] World Anti-Doping Agency (2018). 2018 anti-doping testing figures. Technical report. Accessed: 2025-09-17.

[314] World Anti-Doping Agency (2021a). Decision limits for the confirmatory quantification of exogenous threshold substances by chromatography-based analytical methods, wada technical document – td2021dl. Technical report. Accessed: 2025-09-17.

[315] World Anti-Doping Agency (2021b). International standard for the protection of privacy and personal information (ispppi). Technical report. Accessed: 2025-09-17.

[316] World Anti-Doping Agency (2021c). Measurement and reporting of endogenous anabolic androgenic steroid (eaas) markers of the urinary steroid profile, wada technical document – td2021eaas. Technical report. Accessed: 2025-09-17.

[317] World Anti-Doping Agency (2025). Athlete biological passport. https://www.wada-ama.org/en/athlete-biological-passport. Accessed 2025-09-17.

[318] World Anti-Doping Agency (WADA) (2023). International standard for testing and investigations (isti) 2023. https://www.wada-ama.org/sites/default/files/2022-12/isti_2023_w_annex_k_final_clean.pdf. Accessed 2025-09-17.

[319] World Anti-Doping Agency (2025). Who we are. Accessed: 2025-09-17.

[320] Woźniak, A., Krawiec, K., and Książek, R. (2024). Application of basic machine-learning classifiers for automatic anomaly detection in shewhart control charts. *Decision Making in Manufacturing and Services*, 18:83–98.

[321] Wu, D., Jiang, Z., Xie, X., Wei, X., Yu, W., and Li, R. (2019). Lstm learning with bayesian and gaussian processing for anomaly detection in industrial iot. *IEEE Transactions on Industrial Informatics*, 16(8):5244–5253.

[322] Wu, J., Zeng, W., and Yan, F. (2018). Hierarchical temporal memory method for time-series-based anomaly detection. *Neurocomputing*, 273:535–546.

[323] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., and Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation.

[324] Xu, H., Pang, G., Wang, Y., and Wang, Y. (2023a). Deep isolation forest for anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12591–12604.

[325] Xu, J. and Cai, Y. (2019). Incorporating context-relevant knowledge into convolutional neural networks for short text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):10067–10068.

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

[327] Xu, L., Xu, K., Qin, Y., Li, Y., Huang, X., Lin, Z., Ji, X., et al. (2022). Tgan-ad: Transformer-based gan for anomaly detection of time series data. *Applied Sciences*, 12(16):8085.

[328] Xu, Y., Xu, S., Ramprassad, M., Tumanov, A., and Zhang, C. (2023b). Transehr: Self-supervised transformer for clinical time series data. In *Machine Learning for Health (ML4H)*, pages 623–635. PMLR.

[329] Yaacob, A. H., Tan, I. K., Chien, S. F., and Tan, H. K. (2010). Arima based network anomaly detection. In *2010 Second International Conference on Communication Software and Networks*, pages 205–209. IEEE.

[330] Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., and the Qwen Team (2025). Qwen2.5 technical report.

[331] Yao, Y., Li, L., Astor, B., Yang, W., and Greene, T. (2023). Predicting the risk of a clinical event using longitudinal data: The generalized landmark analysis. *BMC Medical Research Methodology*, 23(1):5.

[332] Yaro, A. S., Maly, F., and Prazak, P. (2023). Outlier detection in time-series receive signal strength observation using z-score method with $s_n$ scale estimator for indoor localization. *Applied Sciences*, 13(6):3900.

[333] Yaro, A. S., Maly, F., Prazak, P., and Malý, K. (2024). Outlier detection performance of a modified z-score method in time-series rss observation with hybrid scale estimators. *IEEE Access*, 12:12785–12796.

[334] Yin, C., Zhang, S., Wang, J., and Xiong, N. (2020). Anomaly detection based on convolutional recurrent autoencoder for iot time series. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(1):112–122.

[335] Zamanzadeh Darban, Z., Webb, G. I., Pan, S., Aggarwal, C., and Salehi, M. (2024). Deep learning for time series anomaly detection: A survey. *ACM Computing Surveys*, 57(1):1–42.

[336] Zhang, C., Li, S., Zhang, H., and Chen, Y. (2019a). Velc: A new variational autoencoder based model for time series anomaly detection. *arXiv preprint arXiv:1907.01702*.

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

[338] Zhang, D., Feng, T., Xue, L., Wang, Y., Dong, Y., and Tang, J. (2025). Parameter-efficient fine-tuning for foundation models.

[339] Zhang, J., Guo, S., Qu, Z., Zeng, D., Zhan, Y., Liu, Q., and Akerkar, R. (2021). Adaptive federated learning on non-iid data with resource constraint. *IEEE Transactions on Computers*, 71(7):1655–1667.

[340] Zhang, Y., Ruan, W., Wang, F., and Huang, X. (2020a). Generalizing universal adversarial attacks beyond additive perturbations. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 1412–1417, Los Alamitos, CA, USA. IEEE Computer Society.

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

[342] Zhao, Y. and Hryniewicki, M. K. (2018). Xgbod: Improving supervised outlier detection with unsupervised representation learning. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

[343] Zhao, Y., Hu, X., Cheng, C., Wan, C., Wang, W., Yang, J., Bai, H., Li, Z., Xiao, C., Wang, Y., Qiao, Z., Sun, J., and Akoglu, L. (2021). Suod: Accelerating large-scale unsupervised heterogeneous outlier detection.

[344] Zhao, Y., Nasrullah, Z., Hryniewicki, M. K., and Li, Z. (2019). LSCP: locally selective combination in parallel outlier ensembles. In *Proceedings of the 2019 SIAM International Conference on Data Mining, SDM 2019*, pages 585–593.

[345] Zhou, Z. G. and Tang, P. (2016a). Continuous anomaly detection in satellite image time series based on z-scores of season-trend model residuals. In *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3410–3413. IEEE.

[346] Zhou, Z. G. and Tang, P. (2016b). Improving time series anomaly detection based on exponentially weighted moving average (ewma) of season-trend model residuals. In *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3414–3417.

[347] Zhu, G., Zhao, H., Liu, H., and Sun, H. (2019). A novel lstm-gan algorithm for time series anomaly detection. In *2019 Prognostics and System Health Management Conference (PHM-Qingdao)*, pages 1–6. IEEE.

[348] Zitnik, M., Nguyen, F., Wang, B., Leskovec, J., Goldenberg, A., and Hoffman, M. M. (2019). Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Information Fusion*, 50:71–91.

[349] Zouhar, V., Meister, C., Gastaldi, J., Du, L., Sachan, M., and Cotterell, R. (2023). Tokenization and the noiseless channel. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5184–5207.