Grenze International Journal of Engineering and Technology, June Issue

Agentic AI for Sports Violation Detection by using Explainable AI

Premanand Ghadekar[1], Rajkanya Pravin Joshi[2], Anushka Ganesh Satav[3], Vaidehi Navanath Lokhande[4] and Kajal Ganesh Bhirud[5]
[1-5]Department of CSE-AIML, Vishwakarma Institute of Technology Pune, India
Email: premanand.ghadekar@vit.edu, rajkanya.1252030003@vit.edu, anushka.1252030007@vit.edu, vaidehi.1252030008@vit.edu, kajal.1252030011@vit.edu

***Abstract*— The critical and growing need for smart, transparent and real-time detection of ethical offenses in professional sports like match-fixing, doping and cyber misconduct is met by Agentic AI for Sports Violation detection by Utilizing Explainable AI. The proposed system introduces a new multi-agent architecture where every expert agent employs modality-specific information to detect a specific type of violation. These include a Cyber Agent employing optimized RoBERTa for text analysis and a Bio Agent and Match-Fixing Agent powered by GridSearch-optimized XGBoost. The multiple-intent user queries are processed by a central Task Specifier based on LLM which dynamically allocates tasks to the concerned agents. For ensuring transparency and interpretability for predictions the framework is designed with Explainable AI methods, namely SHAP (Shapley Additive Explanations) for structured data and token attribution based on attention for text inputs. Conceiving a single agentic framework for multi-domain sports infraction detection, including explainability to ensure stakeholder trust and performance comparison with classical models are some of the aims of the study. The experiments on synthetic data demonstrate good noisy and missing query tolerance, low inference latency and high accuracy. As compared to traditional methods, the new approach excels at explaining outcomes, working in real time, and making predictions even more faster and more accurately as well as efficient. The outcome proves that the application of this system towards automated, explainable and reliable monitoring of sports infractions is possible and scalable and will immensely aid fair play and moral policing within the sporting universe.**

***Keywords*—Agentic AI, XGBoost, LLMs, RoBERTa, Explainable AI, NLP, Doping Detection, Match Fixing Analysis, Cyber-Threat.**

I. INTRODUCTION

It is more crucial now than ever to detect and prevent inappropriate activity in competitive sports such as match-fixing, doping and impropriety on the internet and social media. Unethical practices such as doping, match-fixing, and online misconduct are increasingly threatening the fairness and credibility of modern sports events. Conventional detection systems usually lack transparency, operate reactively, and have limited scope. This paper introduces an explainable multi-agent system as Agentic AI for Sports Violation Detection by using explainable AI to counteract them. It applies modality-specific agents to detect various violations based on both structured (e.g., biochemical markers) and unstructured (e.g., social media content) information. User queries are routed dynamically to the corresponding agent through a central task specifier from a large language model.

Grenze ID: 01.GIJET.12.2.507_2

The technology provides an explicit, accurate and real-time reply for sports ethics enforcement by synergizing attention-based explainability methods with SHAP.

Doping, match-fixing and cyberbullying are only a few examples of sport crimes that continue to erode the integrity and fairness of competitive sports. Conventional detection methods are not effective for identifying smart and adaptive threats since they are mostly reactive and limited by data sources. We present an agentic AI system that utilizes a modular, multi-agent design to overcome such limitations and simultaneously monitor a variety of sports offenses. Through modality-specific data inputs that are pertinent to its domain, each agent within the system becomes an expert at detecting doping, match-fixing or social media offenses.

A large language model (LLM)-based task specifier is a key module of the system which handles multi-intent natural language requests, extracting structured data and dynamically passing them to the relevant agents. The system has integrated Explainable AI methods like SHAP values that are providing interpretability along with predictions, to provide transparency and enable stakeholder trust. Using synthetic datasets. It demonstrates the effectiveness of the proposed framework and compare it with baseline models. The results verify the system's improved performance, interpretability and noise resistance and hence it is a promising solution for real-time sports violation regulation and detection.

The groundwork for employing AI and LLM-based agents in the identification of ethical breaches of sports has been laid with recent advancements in these areas. As an instance, to emulate human-like behaviour and context-specific responses, AI agents have been used successfully as smart NPCs in virtual environments [1]. AI-powered automation has accelerated drug discovery in the biomedical field using predictive modelling and analytics [2]. Although LLM-based agents such as AutoBA have enabled human-less dynamic bioinformatics workflows [3], The broader incorporation of AI agents as informed colleagues shows their ability for automation, multi-domain ethical collaboration and solution- finding [4]. In addition, the application of AI agents in scientific exploration and hypothesis generation has grown due to the integration of experimental methods with reason- based learning [5]. The capacity of LLMs for understanding intricate, multi-modal input for authoritative decision-making is further exemplified by their application in personalized drug-drug interaction identification [6]. The practicability and feasibility of instating agentic, explainable AI systems for complex, real- time tasks such as sports ethics monitoring are assured by these lines of research combined.

II. Literature Survey

Development and experimentation of an LLM-powered AI agent employing GPT-4 to emulate human-like behaviour is a stride in a space where high-level human-agent interaction in Virtual Reality (VR) is still an issue. The agent [1] which is implemented as a Non-playable Character (NPC) in VRChat enhances immersion through context-aware reactions supplemented with the right body movements and facial expressions. Ideal generation conditions for believable interaction were determined through initial tests. This technique is a critical step towards constructing expressive, smart NPCs for forthcoming VR applications.

Drug development is being revolutionized by the use of AI in the guise of ML as it improves accuracy and efficiency at every step of the development pipeline.[2] Use of AI is highlighted in emerging trends in this report that includes key steps such as lead discovery, target identification, disease identification, diagnosis and chemical screening. Machine learning–based pattern analysis accelerates clinical trial processes and supports decision-making in drug research by handling large, complex datasets efficiently. The article also highlights the need for ethical considerations proper training of algorithms and quality data in managing patient data. Collectively these advances place AI as a force to speed up drug development with lower costs and ultimately improve patient care.

AutoBA is an autonomous AI agent developed on large language models (LLMs) that performs fully automated multi-omic bioinformatics analysis.[3] It generates rich analytical workflows that dynamically adjust to variation in input data with minimal user input. As opposed to fixed pipelines AutoBA is flexible and includes new tools emerging. AutoBA supports different types of LLM backends for both modes usage to maintain data confidentiality. It is equipped with an integrated automated code repair (ACR) system to improve its stability compared to tools such as ChatGPT. AutoBA provides a robust, secure and flexible option for efficient bioinformatics analysis.

[4] The development of AI agents from simple copilots to intelligent coworkers in various fields of industry is discussed in this study. It describes how they have progressed over time their primary abilities such as automation and problem- solving and real-world applications in fields such as healthcare, business and education. Ethical issues such as workplace diversity and human-AI collaboration strategies are also included in the study. A comprehensive lesson on the development of AI agents is provided which addresses data processing

goal definition algorithm selection and ethics. Finally, it discusses the development of AI towards AGI and its societal implications which urges the application of AI in a responsible and equitable manner.

By sceptical learning with reasoning and incorporation of experimental methods with AI models and AI researchers are projected to be collaborative systems that enhance biomedical research. data analysis and automation to complement human imagination and not displace it. They can screen knowledge gaps by themselves define research processes and learn progressively with organized memory fueled by generative models and big language. AI agents can drive progress in cellular circuit design, phenotypic control, virtual cell simulation and drug development by combining scientific knowledge and biological principles.

The application of large language models (LLMs) helps improve drug-drug interaction (DDI) detection. It's a kind of condition where a patient or culprits is treated with variant drugs consumption for various diseases. Patient safety is threatened by the time- and labor-consuming process of traditional DDI detection. To facilitate personalized DDI prediction the system suggested evaluates extensive patient data which includes medical history and personal traits against the natural language processing capabilities of LLMs. Facilitating safer, better-informed treatment decisions for medical professionals is the hope. The previous outcomes are good, but enhanced research with comparatively huge data sets is required to accelerate the accuracy of the outcomes. Overall, the approach has high potential to improve patient care with simplify treatment regimens and improve pharmaceutical safety.

[7] This issue of counterfeit drug-drug interactions (DDIs) is resolved in this paper as it could jeopardize the accuracy of the DDI prediction models. Although Graph Neural Networks (GNNs) have been found promising in the task of predicting DDIs they are often incorrect due to flawed connections within the DDI network introduced by wrong drug knowledge or issues within the data. The proposed ANSM technique enhances DDI networks by finding and minimizing false connections to bypass this. A discriminative classifier for finding wrong links a network optimizer for improving the quality of features and silencing noisy links and a feature extractor for capturing structural patterns between drug pairs are all integrated into ANSM. Based on experimental results ANSM significantly improves the detection of spurious DDIs compared to existing methods, providing more reliable predictions in drug safety analysis.

Artificial intelligence has notably shortened the drug discovery timeline, reducing both research duration and cost.

[8] Toxicity prediction is done with tools such as Toxtree, ADMET and ProTox and AI-driven advancements such as DeepMind's AlphaFold2 and Insilico Medicine's Chemistry42 have yielded new drug candidates and breakthroughs such as protein structure prediction. Artificial intelligence (AI) played a crucial role in combatting COVID- 19 from the identification of viruses to vaccine development, enhancing the efficacy of pharmacological breakthroughs.

This paper [9] is with significant advances in patient treatment diagnosis and care the usage of Artificial Intelligence in healthcare is developing a factor of change in the industry. This paper looks into the utilization of AI currently and illustrates how it influences diagnostic precision, personalized treatments and decision- making in clinics. It illustrates how AI can contribute to better patient outcomes through the use of explainable AI in leukemia diagnosis and deep learning for pharmaceutical discovery. AI is transforming the delivery of healthcare and offering unprecedented opportunities to enhance patient care through its application in treatment optimization, clinical decision support and healthcare data analysis.

Adverse Drug Reactions (ADRs) are a big concern for healthcare practitioners and have been associated with patient mortality. Recently, research has been presented supporting a "smart pharmacy" concept designed to utilize large language models to automatically detect potential interactions among medications by incorporating prescription information, medication lists, or Electronic Health Records [10]. Another study aims to apply pharmacovigilance to Social Media & Reviews containing structured information via NER & transformer-based models for Adverse Drug Reaction detection, and uses generative AI to develop interpretable explanations to improve drug safety & decision-making [11].

The use of artificial intelligence (AI) to curb doping in sport is discussed in this research. [12] Doping where performance is enhanced using drugs is becoming more sophisticated and harder to detect. AI is capable of offering doping tests more quickly and accurately than traditional methods with less expense and time. The research highlights the ways in which AI could help national and international agencies cut back on doping offenses thus towards a cleaner, fairer sports environment.

The detailed architecture of e-learning methods for anti- doping education is discussed in this research. [13] It brings into focus the increasing issue of doping in sport that involves new problems such as gene doping and the illegal internet distribution of performance drugs. The research considers the effectiveness of anti-doping education with physiological markers i.e. skin potential reflex (SPR). The results suggest that the application of

strong visual stimuli, films and occasional strong images in e-learning resources enhances anti-doping education's effectiveness by rendering it more remembered and engaging for participants.

The challenge of detecting the doping agent erythropoietin in the blood samples of athletes is explored in this study. While direct detection methods are effective, they are often costly and inaccurate.[14] To design an indirect method for detecting erythropoietin using biomarkers from blood the research compares various machine learning models with statistical inference. Results validate the impactful-ness of ensemble techniques like random forest and XGBoost in drug detection, presenting an anti-doping agency with a less expensive solution. These methods may reinforce sport anti- doping efforts and enhance doping detection.

Its focus on predicting cancer treatment responses and drug efficacy [15] the present study explores the use of Explainable AI (XAI) in precision medicine. It illustrates how XAI makes drug sensitivity prediction on large genomic data sets more transparent and reliable by describing the role of features in AI predictions. The article points out that by providing explanations of deep learning models that examine gene expressions and drug structures. XAI facilitates drug personalisation and enhances trust in AI systems.

Presently, the focus has been on the use of machine learning and artificial intelligence for detecting illicit actions within sports such as doping, fixing, and online harassment. Ensemble learning models and deep learning models have been applied in athlete biologic passports to enable the earlier detection of irregular biologic rhythms indicating actions of doping. In addition, models based on machine learning have been proposed for detecting irregular betting actions indicating fixing in betting companies. In addition, models based on natural language processing in transformer models have been applied to detect online harassment in social platforms related to sports.

On the other hand, the Agentic AI approach proposed here counteracts these issues with a unified multi-agent architecture that tackles tasks like biochemical doping detection, match-fixing detection, and cyber violations concurrently. Unlike other current approaches using standalone models for different tasks in sports analysis, the proposed system embodies a task router using a centralized LLM-based task specifier to route tasks dynamically towards multiple agents specialized in different domains. Moreover, the intentional use of explainableAI methods like SHAP for structured input data and attention-based tokenAttribution for free-text input ensures accountability.

III. METHODOLOGY

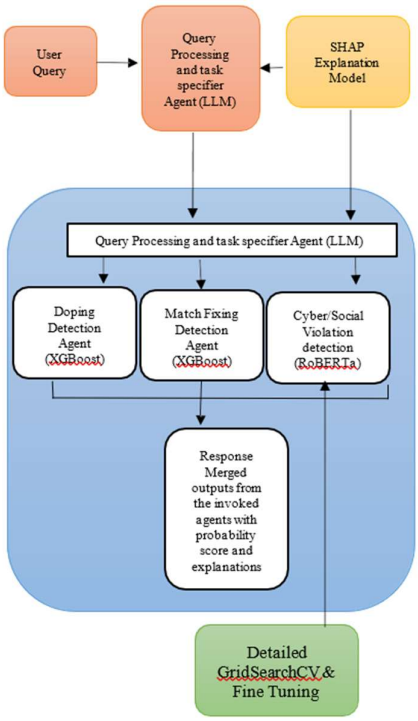

Fig. 1. Architecture diagram of Agentic AI for Sports Violation Detection by Using Explainable AI

Agentic AI for Sports Violation Detection by Using Explainable AI is proposed for sport regulatory bodies the increasing diversity and sophistication of sport offenses from match-fixing and social rule-breaking to biological doping pose serious challenges. Traditional detection tools are insufficient to deal with evolving threats to sports integrity due to their often-siloed operation lack of scalability and lack of transparency[16]. A comprehensive smart system that can detect such infractions as well as yield explainable insights to support decision-making is in dire need. The proposed solution "Agentic AI for Sports Violation Detection using Explainable AI" bridges this gap by proposing a multi-agent modular system meant for overall violation detection within the sports system. The system copes well with a wide range of cases such as drug abuse (biochemical doping) match-fixing and cyber/sociable misbehaviour (digital offenses). In a single framework under a central task-routing agent's control the methodology integrates Explainable AI (XAI) modules fine-tuned transformer models and classical machine learning algorithms. By assigning incoming data to the best-violation-detection module dynamically this smart agent ensures great accuracy and operating transparency in a number of different domains.

Figure 1 illustrates the framework of the proposed Agentic AI system designed for sports violation detection. A Query Processing and Task Specifier Agent powered by a large language model (LLM) processes the user query to initiate the process. Depending on the kind of violation this agent processes the query and sends it to one or more of the three specialist agents such as the Social Violation Detection Agent (with the RoBERTa model) the Match-Fixing Detection Agent (with the XGBoost model) and the Doping Detection Agent (with the XGBoost model). Each agent individually processes the data and sends back the outcome. To ensure interpretability and transparency the results of the called agents are merged to produce the end Response presenting probability scores as well as readable insights.

There is a specific AI agent in the system that gets triggered for every identified intent of the user query. The Bio Agent forecasts instances of biochemical doping by examining readings from urine and blood. To identify suspicious trends indicative of match rigging the Match-Fixing Agent analyzes betting metrics and team performance statistics. The Cyber Agent on the other hand translates text from social media captions or messages and classifies them as abusive, safe, or menacing. Upon processing the system generates a final output that provides transparency in decision-making through an interpretable explanation based on SHAP values or attention processes a prediction label (e.g. "Doping" or "Fixed") and a corresponding probability score

A. Data Collection and Feature Engineering

For every one of the three focused areas biochemical doping, match-fixing and social media intrusions there was an elaborate synthetic data generation process undertaken in light of the sparse amount of real-world data related to sports crime. Various studies from credible sources such as medical journals, anti-doping reports, sports analytics literature and toxic language detection studies, were used as the basis for this approach. Lacking sensitive or inaccessible real data the challenge was to copy varied and plausible datasets that would be used in training and checking AI agents.

With Gaussian priors and bounded distributions 1,000 synthetic samples were generated for the Doping Detection dataset. The samples were then validated with rule-based conditions to ensure biological plausibility e.g., maintaining hemoglobin levels within [12.5–17.5] g/dL. Hemoglobin, RBC count, testosterone/epitestosterone (T/E) ratio, OFF- score, urine biomarkers, exogenous androgenic anabolic steroids (EAAS) marker and demographic variables like age and gender were some of the prominent features. For modeling characteristics like victory rates, goals on average, frequency of red cards, percentage of ball possessions, betting probabilities and anomaly in betting volumes, the Match- Fixing Dataset was established by using historic match logs, patterns of bets and algorithms used for fixing. Motivated by such datasets as HateXplain and the Gab Hate Corpus, text data such as tweets, captions and messages were artificially generated and annotated using pattern-based annotation for the Social Violation Dataset. The Cyber Agent was able to categorize these texts into three classes as "abusive" "threatening" and "safe."

B. General Agent Workflow

Input as Natural language query Q, with embedded or structured data.

Output as Predicted label and explanation per detected intent.

- Intent Detection

Use fine-tuned LLM (Groq LLaMA 3) to classify intent(s) from query Q → Intent Set I = $\{i_1, i_2, ..., i_n\}$.

- Data Extraction & Imputation

For each intent i in I extract feature set $F_i = \{f_1, f_2,$
$..., f_k\}$ If $f_j \in$ required and $f_j \notin F_i$, then f_j = default(f_j)

// Use predefined domain-based imputation

- Model Prediction

For each intent i scale F_i using StandardScaler Predict label $y_i = M_i(F_i)$ where M_i is the agent model (e.g., XGBoost or RoBERTa) Obtain probability $P_i(y_i = 1 \mid F_i)$

- Explainability

- For classical models use SHAP → importance(f_j) - For RoBERTa use attention/attribution → importance($word_j$)

- Return

{intent i, prediction:y_i, probability:P_i, explanation} for each agent.

C. Agent Training Process

A uniform machine learning pipeline optimized for performance and interpretability was employed to train each AI agent in the system separately on its own synthetic data. The pipeline began with thorough preparation steps such as normalizing numerical features handling un-available values and encoding categorized and isolated variables. Feature scaling was employed to ensure consistency among input variables which is important for algorithms that are sensitive to data size so that it can guarantee consistency among input variables. The right models were selected for each domain RoBERTa for the Cyber Agent due to its superiority in natural language processing tasks and XGBoost for the Doping and Match- Fixing agents due to its robustness with tabular data.

To determine the optimal configurations for accuracy and generalization hyperparameter tuning was performed using GridSearchCV. Finally, SHAP was came into practice to accelerate the interpretability of model predictions so that the system could offer brief explanations for each decision the agents made.

The Bio Agent was built using an XGBoost classifier that was particularly optimized to handle the challenges of imbalanced binary classification as it is often found in cases of doping detection where positive instances are relatively rare. To correct this imbalance the class weights of the model were altered to penalize minority class misclassification more heavily making it more sensitive to potential doping instances. In order to ensure generalizability and robustness the training process also involved approaches such as cross- validation and stratified sampling. The separation of normal profiles from suspicious profiles was enabled through the training on a synthetic set that contained both hematological as well as biochemical markers. The ability of the Bio Agent to reliably flag doping activity at the cost of high precision as well as memory was enhanced using this tailored process.

$$X = [x_1, x_2, x_3] \in \mathbb{R}n \qquad (1)$$

This eq 1 shows that a feature vector X composed of elements x1, x2, ..., x3 where each xi represents an individual feature or input variable in the dataset. The notation $X \in \mathbb{R}n$ indicates that this feature vector lies in an n-dimensional real- valued space which meaning it contains n numerical features.

Then the posterior probability of doping is modeled as

$$P(y = 1|X) = \sigma\left(\sum_{k=1}^{K}(X)\right) \qquad (2)$$

Equation (2) which calculates the probability of a sample which belongs to the positive class (e.g., doping detected) by summing the outputs of K base decision trees fx (X) and applies the sigmoid function σ . This transforms the raw score into a probability between 0 and 1. The model achieved 96% accuracy using this approach.

A supervised learning system with 13 well-chosen features derived from simulated match statistics and gambling activity was utilized to identify match-fixing. Statistics such as victory ratios mean goals per match possession percentages and odd gambling volumes were among these factors. Due to its outstanding ability to detect complex patterns and relationships among the data XGBoost performed better than Random Forest and Logistic Regression as the best performing model among the ones tested. High separability between fake and genuine matches in the artificial dataset was an important contributor to the model's outstanding performance that included 100% accuracy for the validation set.

HuggingFace's text-classification pipeline is employed to develop a fine-tuned RoBERTa transformer model which forms the basis of the Cyber Agent. It was trained on tagged synthetic text data designed to capture a range of cyberviolences including threats, hate speech, blackmail and innocuous content. The model employs RoBERTa's deep contextual language comprehension to correctly classify inputs. It accurately separated safe

messages and dangerous material in all testing with 100% accuracy. It further improves predictability and transparency of its output by providing token-level attention-based explanations.

D. Doping Detection (XG Boost Classifier)

e XGBoost classifier model is used for doping detection. Each instance in the dataset is represented as a feature vector
$X = [x_1, x_2, \ldots, x_3] \in \mathbb{R}^n$ which is indicating that each data point has 17 features. The XGBoost model optimizes an objective function that consists of two parts as the first is the sum of a loss function $l(y_1, y_2)$ over all training instances which measuring the difference between the predicted and actual labels and the second is a regularization term
$\sum_{k=1}^{K} \Omega(f_k)$ which helps control the complexity of the model. represents the predicted value for instance i as the sum of outputs from K regression trees f_k. The regularization term
$\Omega(f)$ is defined as $\gamma T + \frac{1}{2}\lambda\|w\|^2$ where T is the number of
leaves in the tree and w denotes the leaf weights. The final binary prediction y is made based on the probability $P(y = 1|X) > 0.5$ the output is 1 which indicates a positive doping detection otherwise the output is 0[17].

E. Match fixing Detection (XGBoost Classifier)

XGBoost classifier model used for match-fixing detection. This model follows a similar structure to the doping detection classifier but uses a different set of input features. Each instance is represented by a 13-dimensional feature vector
X = [team_a_win_ratio, . . . ,social_sentiment_score] $\in \mathbb{R}13$, which may include metrics like team performance ratios and social media sentiment scores. The classification function computes the probability that the outcome y is 1 (indicating potential match-fixing) given the input features X. This probability is calculated as

$$P(y = 1|X) = \sigma(\sum^{K} \quad f_k(x)) \quad (3)$$

Eq 4 which is f_k represents the output of the k^th regression tree and σ denotes the sigmoid function which maps the result to a probability between 0 and 1. This formulation enables the model to perform binary classification where the final output indicates the likelihood of match-fixing based on the input features[18].

F. Cyber Violation Detection (RoBERTa Transformer)

The Cyber Violation Detection model that uses the RoBERTa transformer for text classification. When a text input T T is provided RoBERTa first tokenizes it into a sequence of tokens starting with a classification token [CLS] followed by the word tokens $w1, w2, \ldots, wn$, ending with a separator token [SEP]. This tokenized input is then fed into the transformer model to produce contextualized embeddings as

$$H = Transformer(T) = [h_{CLS}, h_1, \ldots, h_n] \quad (4)$$

Eq 4 which each $h\ i$ h i represents the contextual embedding of the corresponding token and h_{CLS} is the embedding associated with the [CLS] token which summarizes the entire sequence for classification purposes. Eq 5 The final prediction is computed using a softmax function applied to a linear transformation of h_{CLS}

$$y = softmax(W \cdot hCLS + b) \quad (5)$$

To provide interpretability the model uses gradient-based attribution to assess the importance of each input token. The importance of a token w_j is calculated as the partial derivative of the output y with respect to its corresponding edge state h_j

$$\underline{importance(w}_j) = \frac{\partial y}{\partial h_j} \quad (6)$$

Eq 6 allows one to know which module of the text serves most to the model's decision enhancing the explainability of the prediction[19].
Each of the agents had explainability tools to ensure crystal clear flow and faiith in the decisions made by the system. SHAP (Shapley Additive explanations) values were applied for the Bio Agent and the Match-Fixing Agent which is rely on organized tabular data. Through the display of how every element contributed to the final prediction SHAP allows stakeholders to understand which physiological indicators or match statistics played the greatest role in making the decision. Explainability was achieved for the Cyber Violation Detection Agent through gradient-based and attention-based attribution methods that are native to transformer architectures like

RoBERTa. These approaches give understandable explanations for every classification e.g., if a message was categorized as safe, abusive, or threatening by pointing to specific words or tokens in the input text that the model has found important. Based on the principles of Explainable AI this multi-level explainability ensures that every agent's output is accurate and understandable.

G. XAI-Guided Feature Selection for performance improvement

A comprehensive post-training feature importance analysis was performed with SHAP (Shapley Additive Explanations) to accelerate the performance and interpretability of the model. TreeExplainer was employed to compute SHAP values following the XGBoost classifiers for the Bio Agent— which identifies doping—and the Match-Fixing Agent after they had been initially trained. This allowed for an in-depth understanding of how every attribute affected the model's predictions, which improved model transparency and guided future improvement efforts.

The most influential input features on the predictions made by the model were illustrated graphically using SHAP summary plots. Features that consistently showed minimal SHAP values across the test set were found to be of lesser importance through the analysis of these plots. The models were then retrained with a reduced set of most important features following the removal of these less important characteristics from the data set. SHAP was easily transformed into a feature engineering tool using this approach, which reduced the models' complexity and perhaps improved their interpretability and accuracy.

Doping Detection Agent Feature Refinement

Twelve high-impact features were chosen and retained for retraining models based on the SHAP summary analysis. OFF-Score, T/E Ratio, Doping Detection Ratio, Steroid Mean, Steroid Max, pH, Specific Gravity, WBC Count, Protein, Glucose and Hemoglobin (g/dL) were included. These features encompass a range of urine, biochemical and hematological parameters that are often associated with anomalous doping profiles. Conversely, features with consistently low SHAP values, such as Steroid Min, Age, Gender, RBC Count (Urine) and Athlete ID, were dropped from the model. Aside from eliminating redundancy, this feature reduction enhanced the classifier's ability to generalize by allowing it to concentrate on more relevant and informative signals.

Match-Fixing Agent Feature Refinement

Similarly, the SHAP summary plot of the Match-Fixing Agent revealed four factors that were incredibly important to identify and which had a substantial influence on the predictions of the model: Red cards, timing of surprise goals, volume betted on draws and social sentiment score. Due to these attributes being so effective at identifying potential match-fixing incidents, they were retained.

Along with these, six features having a moderate impact were also retained because they remained relevant across the dataset: Team A Win Ratio, Team B Win Ratio, Team A Possession, Team B Possession, Team A Average Goals and Team B Average Goals.

Conversely, during the retraining stage, features such as Total Betting Volume, Pre-Match Odds for Team A and Pre-Match Odds for Team B were dropped since they have very low SHAP values. Through dropping the duplicate or noisy variables, such feature set adjustment not only enhanced the efficiency of training in the XGBoost classifier but also minimized the likelihood of overfitting. Further, the overall model retained a high level of interpretability, which is particularly beneficial for sports regulatory bodies seeking accountability and transparency in their predictions.

Effect of SHAP-Guided Retraining

These were retrained after SHAP-based feature trimming by using optimized hyperparameters via GridSearchCV. The refined models demonstrated significant improvements in performance-a better accuracy class balance and improved ROC-AUC. Most notably, in this respect, the Doping Detection Agent improves from 79% to 96% accuracy; hence, SHAP-based Explainable AI serves very well the complementary goals of model optimization and interpretability.

IV. EXPERIMENTAL SETUP

A. Dataset Description

Since there are no publicly available sports-related data on violations, artificial datasets were created for testing purposes. The doping dataset has 1,000 instances with hematological and biochemical variables such as hemoglobin concentration, OFF score, T/E ratio, and urine markers. The match-fixing dataset has artificial data

features related to match performance and betting, and the cyber-violation dataset contains labeled text data designed as cyber-violation-related categories based on existing cyber-violation datasets found online.

B. Tools and Implementation Environment

The system was developed in Python programming using XGBoost for biochemical doping as well as match fixing and a fine-tuned RoBERTa model for classification of cyberspace violations. Hyperparameters were tuned using GridSearchCV. Interpretability techniques used SHAP for tree-based classification models and attention-based attribution for fine-tuned RoBERTa classification models.

C. Model Architecture Overview

The modular agent-based system design was used, which involved three specialized agents: Bio Agents, Match-Fixing Agents, and Cyber Agents. There is an LLM-based task specifier that is used to analyze user inputs to arrive at their intents, which are then directed to the corresponding agents. The agents generate a prediction along with an explainable result.

V. Results and Discussions

Various criteria were considered to ensure that the proposed Agentic AI system was reliable for deployment. The agents were very accurate in the detection of doping, match-fixing, and cyber violations, outperforming even the baseline models like Random Forest and Logistic Regression. It remained robust against incomplete and noisy queries and delivered near real-time inferences despite transformer-based components. Besides, explainability through SHAP and attention-based visualizations also proved the efficacy.

A. Agent Performance Evaluation

In all targeted violation detection tasks the completed models did very well. For the biochemical doping and match-fixing detection tasks the XGBoost classifier was able to show state- of-the-art accuracy and robustness as presented in Table 1, validating its suitability for structured and feature-rich data. Optimized for text classification the RoBERTa model is the top choice for spotting infractions in social media and the internet because it outperforms other NLP baselines significantly. These models collaborate to form a multi-agent system that is extremely reliable and efficient when it comes to comprehensive sports integrity monitoring.

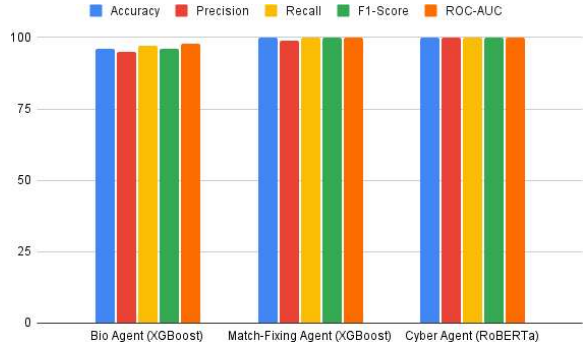

Fig. 2. Visual presentation of comaprison of different agents of Agentic AI for Sports Violation Detection by Using Explainable AI

Fig 2. Shows that the performance comparison of various agents of Agentic AI for Sports Violation Detection by Using Explainable AI is visualized. The Bio Agent which is powered by XGBoost achieved an impressive accuracy with the value of 96% with a great precision of 95%, recall of 97%, F1-score of 96% and a ROC-AUC of 0.98 which indicates strong performance in detecting doping-related violations. The Match-Fixing Agent which is also based on XGBoost which demonstrated near-perfect performance achieved with 100% accuracy of the model with 99% precision, 100% recall, 100% F1-score and a ROC-AUC of 1.00 which showcasing its effectiveness in identifying match-fixing incidents. The Cyber Agent is built using the transformer based RoBERTa model which achieved perfect scores across all metrics which includes model accuracy with 100% and same for the precision, recall, F1-score and ROC-AUC which highlights its exceptional capability in detecting cyber related violations such as online harassment or threats in the sports domain. Explainability via SHAP and Attention

A well-known and impactful method of interpreting machine learning model predictions is SHAP which stands for Shapley Additive explanations. It is a well-known explainable AI method that interprets predictions by

assigning each feature an importance value. It uses Shapley values to fairly distribute the contribution of each feature. SHAP helps in understanding the reasoning behind the model's decisions ensuring transparency and trust in AI systems. SHAP itself based cooperative game theory which means employs the concept of Shapley values to assign each input feature's contribution to a model prediction in a fair manner. By analysing all the combinations of feature inputs, SHAP calculates how much each feature is contributing to the prediction with each feature being a "player" in a game and the model output being the "payout." This allows for feature importance to be explained in a theoretically sound and consistent way.

Especially in domains like healthcare, banking and the legal system SHAP values prove very useful for debugging, building trust in model decisions and ensuring explainability. SHAP provides quantitative measures of the explanations for the model decisions as well as nice visualization in the case of structured data models such as XGBoost.

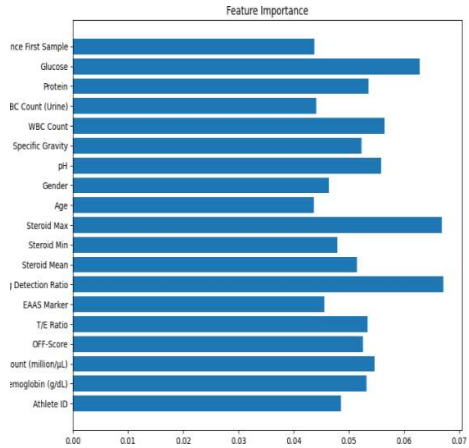

Fig. 3. SHAP Summary Visualization for Doping Detection in Agentic AI for Sports Violation Detection using Explainable AI

fig 3. Shows SHAP summary visualization for doping detection in agentic AI for sports violation detection using Explainable AI. This visualization provides insights into the model's feature importance by illustrating how different features contribute to the prediction of doping violations.

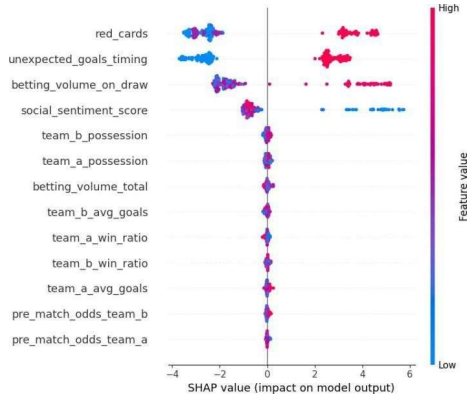

Fig.4. SHAP Summary for match-Fixing in Agentic AI for Sports Violation Detection using Explainable AI

Fig 4 shows the SHAP summary visualization for match- fixing detection in agentic AI for sports violation detection using Explainable AI. This summary provides an overview of the model's feature importance that illustrates how different features influence the prediction of match-fixing activities. By analysing the SHAP values. It can understand the contribution of each feature for models decision making process allowing for the identification of critical factors associated with suspicious match-fixing behaviour in sports. The visualization aids in ensuring transparency and accountability in the AI model's detection capabilities.

Token attribution utilized in the presentation of the Cyber Agent's responses see Fig. 5 which provides significant information on which specific words are employed to denote a threat or harassment. The visualization

displays how particular tokens like "disgrace", "worst" and "ever" play a strong role in determining the model's decision to mark the statement as threatening or harassing for the search "You are a disgrace to the team the worst player ever." It can get a better view of the process of reasoning made by the AI and identify the key linguistic features that lead to the marking of harmful or abusive content by examining the attribution of these words.

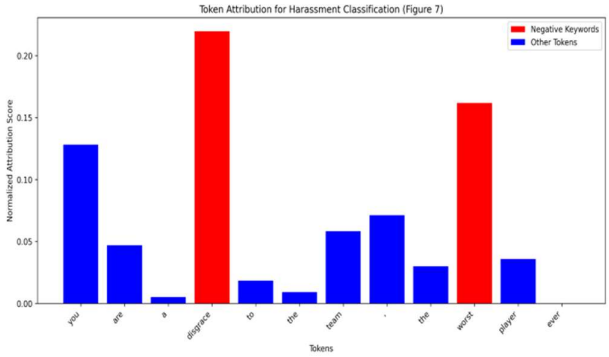

Fig. 5. Attention-Based Word Attribution for Cyber Agent of Agentic AI for Sports Violation Detection by Using Explainable AI

B. Baseline Model Comparision

Trained and evaluated traditional models in order to authenticate our model choices. Results are presented in Table 1 and show that XGBoost easily outperformed Random Forest as well as Logistic Regression although both models created decent performances. When put against the other conventional models, XGBoost's better performance means that it is better suited in recognizing complex patterns within the data which makes it a more reliable choice for the job. This comparison lends credence to the decision to prioritize XGBoost for further model development and refinement.

TABLE I. COMPARISONS BETWEEN DIFFERENT MODELS OF AGENTIC AI FOR SPORTS VIOLATION DETECTION BY USING EXPLAINABLE AI

Model Name	Dataset	Accuracy	Precision	Recall	F1-Score	ROC-AUC
XGBoost	Doping	96%	95%	97%	96%	0.98
Random Forest	Doping	93%	92%	94%	93%	0.95
Logistic Regression	Doping	89%	87%	91%	89%	0.91
XGBoost	Match Fixing	100%	99%	100%	100%	1.00
Random Forest	Match Fixing	96%	95%	96%	95%	0.96
Logistic Regression	Match Fixing	91%	90%	91%	90%	0.92

TABLE II. COMPARISONS BETWEEN EXISTING MODELS AND AGENTIC AI FOR SPORTS VIOLATION DETECTION BY USING EXPLAINABLE AI

Aspect	Existing Approaches	Our agentic AI System
Scope	Focused models for specific domains only: - Doping: AI-Based Detection [18] Match-Fixing: Betting Anomaly Detection [16] Social: LLM for Cyberbullying [19]	Unified system with 3 dedicated agents (Bio, Match-Fixing, Cyber) handled by a central LLM-based Task Specifier.
Techniques Used	- Statistical/threshold methods for doping - SVM, KNN, RF in match-fixing[16][17]	- GridSearch-optimized XGBoost for structured data - Fine-tuned RoBERTa for unstructured cyber data
Explainability	Mostly lacking, or basic threshold-based decisions [18][16]	SHAP (TreeExplainer) for XGBoost; Attention-based explanation for RoBERTa
Accuracy	Varies: - Match-fixing: up to 92% Cyber Harassment: Inconsistent [17][19]	- Bio Agent: 96% - Match-Fixing: 100% - Cyber: 100%
Real-Time Capability	- Match-fixing semi-real-time [16]- Doping and Cyber: delayed/manual processing	Built for near real-time performance with LLM query parsing and fast agent response
Architecture	Monolithic, isolated tools [16][18]	Modular Agentic Architecture with LLM-driven routing and independent agent execution

Table 2 compares existing AI approaches for sports ethics violations with our Agentic AI System. Traditional models are domain-specific—using statistical methods for doping, classical ML like SVM and RF for match-fixing and LLMs for cyberbullying often lacking integration, explainability and real-time performance. In contrast, our system uses a modular architecture with a central LLM Task Specifier and three specialized agents (Bio, Match- Fixing, Cyber). It applies GridSearch-optimized XGBoost and fine-tuned RoBERTa for improved

accuracy (96– 100%) and uses SHAP and attention mechanisms for interpretability. Unlike the manual or delayed processes in older systems this setup supports near real-time analysis, offering a unified, explainable and scalable solution.

TABLE III. AI FOR SPORTS VIOLATION DETECTION BY USING EXPLAINABLE AI MODEL PERFORMANCE BEFORE VS. AFTER XAI-GUIDED RETRAINING

Agent	Feature set used	Accuracy	roc-auc	f1-score
Doping agent	All 19 features	0.79	0.80	0.80
Doping agent (XAI)	Top 12 (after shap)	0.96	0.97	0.96
Match- fixing	All 13 features	0.97	0.98	0.98
Match- fixing (XAI)	Top 10(high + 6 moderate)	1.00	1.00	1.00

For the Doping Agent and the Match-Fixing Agent alike, Table 3 illustrates how efficiently SHAP-based feature selection enhances model performance and explainability. The models were able to focus on the most relevant signals that are associated with anomalous patterns in doping and match-fixing activities by excluding less informative features and retaining those having substantial SHAP contributions. Although the variables selected for the Match-Fixing Agent capture game anomalies and external influences such as betting behavior and social opinion, the features retained for the Doping Agent are significant hematological, biochemical and urine markers. The table illustrates how explainable AI, particularly SHAP, can be highly beneficial in both model decision interpretation and performance optimization through data-driven, smart feature enhancement.
The Agentic AI framework fares well for all categories of sports violations. The XGBoost model is very good at recognizing non-linear relationships existing in well-structured physiological and sports-related data, and the RoBERTa-based cyber agent excels in textual classification enabled through effective language understanding skills. The agentic model also ensures reliability through dynamic task routing depending on detected user intentions.
Explainable AI improves the transparency process by explaining relevant factors of a decision. SHAP explains important attributes such as OFF-score, T/E Ratio, and unexpected goal timing, while attention attribution focuses on significant tokens of cyber violence attribution. This helps in better accountability and usability of AI for sports regulatory bodies.

C. Robustness to Complex Queries

Multi intent and partially completed queries were utilized to measure Agentic AI for Sports Violation Detection by Using Explainable AI robustness and its abilit to handle imperfect real world information. It helps to simulate real world cases where users can offer incomplete information that leads to evaluation focused on the performance of the system to understand and process requests with multiple intents. To able to determine the system's flexibility and accuracy which gives relevant responses despite the challenges posed by these types of questions by trying it out in these situations.

Fig. 6. Complex Query Input give to Agentic AI for Sports Violation Detection by Using Explainable AI

Fig 6 shows a complex query is given as input to our system which has multiple intents with some data provided On clicking the process button afterwards our LLM whatever classifies the intents and extracts the required data and hands over the task to a specified agent which handles the task and gives the result with probability and explanation.

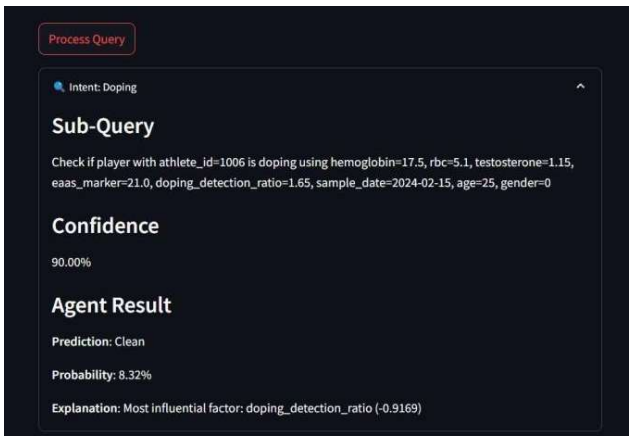

Fig. 7. Results from Agent Doping Detection Agent of Agentic AI for Sports Violation Detection by Using Explainable AI

Fig 7 shows a Output from the Agentic AI-based Doping Detection agent showing a "Clean" prediction for athlete_id=1006 with 90% confidence. The most influential factor was the doping_detection_ratio (-0.9169) contributing to a low doping probability of 8.32%.

Fig. 8. Results from Agent Match-Fixing Detection Agent of Agentic AI for Sports Violation Detection by Using Explainable AI

Fig 8 showcasing the results from the Agentic AI-based Match-Fixing Detection agent showing a "Clean" outcome for MATCH_9 with 80% confidence. The most influential factor was unexpected_goals_timing (-3.4305) contributing to a low match-fixing probability of 1.33%.

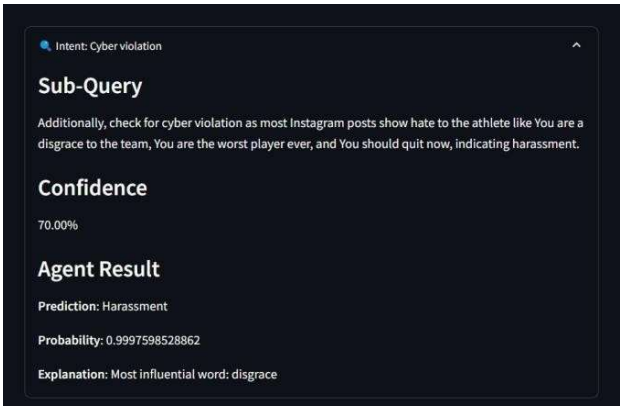

Fig. 9. Result from Social Violation Detection Agent of Agentic AI for Sports Violation Detection by Using Explainable AI

Fig 9. Shows the result from the Agentic AI-based Cyber Violation Detection Agent showing a "Harassment" prediction with 70% confidence. The most influential word was “disgrace” contributing to a high probability of 99.98%.

VI. CONCLUSION

Agentic AI for Sports Violation Detection by Using Explainable AI this system is a robust and flexible architecture for detecting infractions associated with sports. By utilizing explainable models that enhance interpretability and trustworthiness every agent in the system is trained for a particular objective. An exhaustive and balanced solution is achieved by integrating classical and transformer-based models for decision-making Groq-hosted LLaMA 3 for query decomposition and SHAP/attention maps for explainability. While maintaining high precision and transparency. Proposed system displays strong robustness to noisy input missing data and multi-intent questions. The proposed system achieved superior predictive accuracy and interpretability compared with baseline models, attaining 96 % accuracy for doping and perfect detection for match-fixing and cyber violations. Transparency was ensured through SHAP and attention-based methods and robustness tests confirmed its reliability in real- world scenarios. To extend its application in moral sports monitoring, future enhancements could include real-time video analysis, communication with federated sport data and prediction of player behavior. Its capabilities and functionalities in the sports industry may further be developed in the future through the inclusion of real-time data integration and interaction with federated sport databases and the ability to track athlete rehabilitation and predict behavior.

REFERENCES

[1] Wan, H., Zhang, J., Suria, A. A., Yao, B., Wang, D., Coady, Y., & Prpa, M. (2024, May). Building llm-based ai agents in social virtual reality. In Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (pp. 1-7).

[2] Rehman, A. U., Li, M., Wu, B., Ali, Y., Rasheed, S., Shaheen, S., ... & Zhang, J. (2024). Role of artificial intelligence in revolutionizing drug discovery. Fundamental Research.

[3] Zhou, J., Zhang, B., Li, G., Chen, X., Li, H., Xu, X., ... & Gao, X. (2024). An AI Agent for Fully Automated Multi-Omic Analyses. Advanced Science, 11(44), 2407094.

[4] Hayawi, K., & Shahriar, S. (2024). AI Agents from Copilots to Coworkers: Historical Context, Challenges, Limitations, Implicationsand Practical Guidelines. Preprints, 10.

[5] Gao, S., Fang, A., Huang, Y., Giunchiglia, V., Noori, A., Schwarz, J. R., ... & Zitnik, M. (2024). Empowering biomedical discovery with AI agents. Cell, 187(22, 6125-6151.

[6] Alshehri, B. M. J., Kraiem, N., Sakly, H., & Alasbali, N. (2024, July). Enhancing Medication Safety with Large Language Models: Advanced Detection and Prediction of Drug-Drug Interactions. In 2024 IEEE 7th International Conference on Advanced Technologies, Signal and Image Processing (ATSIP) (Vol. 1, pp. 547-552). IEEE.

[7] Wang, H., Cui, Z., Yang, Y., Wang, B., Zhu, L., & Zhang, W. (2024). A network enhancement method to identify spurious drug-drug interactions. IEEE/ACM Transactions on Computational Biology and Bioinformatics.

[8] Soni, K., & Hasija, Y. (2022, February). Artificial intelligence assisted drug research and development. In 2022 IEEE Delhi Section Conference (DELCON) (pp. 1-10). IEEE.

[9] Kaur, G., Singh, H., & Mehta, S. (2024, April). AI Doctors in Healthcare: A Comparative Journey through Diagnosis, Treatment, Care, Drug Developmentand Health Analysis. In 2024 Sixth International Conference on Computational Intelligence and Communication Technologies (CCICT) (pp. 485-492). IEEE.

[10] Osheba, A., Abou-El-Ela, A., Adel, O., Maaod, Y., Abuhmed, T., & El-Sappagh, S. (2025, January). Leveraging Large Language Models for Smart Pharmacy Systems: Enhancing Drug Safety and Operational Efficiency. In 2025 19th International Conference on Ubiquitous Information Management and Communication (IMCOM) (pp. 1-8). IEEE.

[11] Bera, A., Das, R., Ghosh, S., Chakraborty, R., Mitra, I., & Nandy, P. (2023, August). Harnessing transformers for detecting adverse drug reaction and customized causality explanation using generative ai. In 2023 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA) (pp. 1-6). IEEE.

[12] Choudhary, A., & Mehta, R. (2024, May). Use of Artificial Intelligence in the fight against doping. In 2023 International Conference on Smart Devices (ICSD) (pp. 1-6). IEEE.

[13] Shionoya, A., Nagahama, Y., McGown, V., Fukumura, Y., & Miyake, H. (2011, September). Concept Design to Develop the e- Learning Contents for Effective Anti-doping Education Using Kansei Parameters. In 2011 International Conference on Biometrics and Kansei Engineering (pp. 189-192). IEEE.

[14] Rahman, M. R., Bejder, J., Bonne, T. C.andersen, A. B., Huertas, J. R., Aikin, R., ... & Maass, W. (2022, July). Detection of erythropoietin in blood to uncover doping in sports using machine learning. In 2022 IEEE International Conference on Digital Health (ICDH) (pp. 193-201). IEEE.

[15] Gillani, I. S., Shahzad, M., Mobin, A., Munawar, M. R., Awan, M. U., & Asif, M. (2022, September). Explainable ai in drug sensitivity prediction on cancer cell lines. In 2022 International Conference on Emerging Trends in Smart Technologies (ICETST) (pp. 1-5). IEEE.
[16] Kim, C., Park, J. H., & Lee, J. Y. (2024). AI-based betting anomaly detection system to ensure fairness in sports and prevent illegal gambling. Scientific Reports, 14(1), 6470.
[17] Ryoo, H., Cho, S., Oh, T., Kim, Y., & Suh, S. H. (2024). Identification of doping suspicions through artificial intelligence- powered analysis on athlete's performance passport in female weightlifting. Frontiers in Physiology, 15, 1344340.
[18] Ouyang, Y., Li, X., Zhou, W., Hong, W., Zheng, W., Qi, F., & Peng, L. (2024). Integration of machine learning XGBoost and SHAP models for NBA game outcome prediction and quantitative analysis methodology. Plos one, 19(7), e0307478.
[19] Ogunleye, B., & Dharmaraj, B. (2023). The use of a large language model for cyberbullying detection. Analytics, 2(3), 694-707.