

Effective Use of Technology for Dissemination of Anti Doping Information

Prof. Deekshitha S
Associate Professor
School of CSE
REVA University
Bengaluru, India

S. Harshitha
Department of Computer Science and Engineering
(Artificial Intelligence and Data Science)
REVA University
Bengaluru, India

M. Likhitha
Department of Computer Science and Engineering
(Artificial Intelligence and Data Science)
REVA University
Bengaluru, India

G. Thanuja
Department of Computer Science and Engineering
(Artificial Intelligence and Data Science)
REVA University
Bengaluru, India

M. Mahathi
Department of Computer Science and Engineering
(Artificial Intelligence and Data Science)
REVA University
Bengaluru, India

Abstract—Unchecked use of performance enhancing drugs and unchecked dietary supplements in sporting activities are a big threat to the health of the athletes and the fairness of the competition [1], [2]. In this paper, the author introduces AiDoping, a complete full-stack intelligent system to prevent doping and classify drugs, which combines machine learning and large language models (LLMs) with multimodal computer vision [9], [10]. Random Forest [6] and XGBoost [5] classifiers are trained and tested on a dataset of 1,500 drug and supplement entries that can be characterized by nine attributes, such as ingredient name, toxicity level, banned status, regulatory approval, and health effects, the best-performing model of which is deployed to make real-time predictions of the safety of a given drug or supplement in addition to confidence scores. The interpretability of the model is also guaranteed by making use of SHAP (SHapley Additive exPlanations) [7], which produces per-prediction feature importance visualizations [14]. The system also includes GPT-4o-mini-based injury severity analysis with the OpenFDA API to retrieve pharmaceutical labels live [16], a symptom-based drug recommendation chatbot [12], an anti-doping awareness quiz, and a drug summarization system which is also powered by GPT-3.5-turbo [11]. The backend is based on Fast API and asynchronous MongoDB storage and the frontend is developed on React 19 and Vite [17], [18]. Authentication is implemented using role-based authentication through the use of JSON Web Tokens and password hash (BCrypt). The suggested system shows that it is possible to integrate explainable machine learning with generative AI to provide a full, real-time, and interpretable drug safety awareness system [20].

Index Terms—Anti-doping, drugs classification, machine learning, Random Forest, XGBoost, explainable artificial intelligence, SHAP, large language models, GPT-4o-mini, computer vision, wound analysis, OpenFDA, FastAPI, React, sports integrity.

I. INTRODUCTION

Professional sports activities that use performance-enhancing drugs (PEDs) and unhealthy dietary supplements are very dangerous to the health and fair play of athletes. Thousands of anti-doping rule violations are reported by WADA each year, and most of them are related to the accidental ingestion of the prohibited substances that are found in commercial supplements [1]. It has been established that nutritional supplements are often full of unreported androgenic anabolic steroids and other prohibited substances that expose athletes to the risk of committing inadvertent cases of doping violations [2]. The increased sophistication of multi-ingredient supplement goods has made the manual check on the legality of substances unfeasible, which has resulted in a high demand of intelligent/automated drug safety assessment systems.

Machine learning has come out as a successful paradigm of classification in biomedical and pharmaceutical informatics. Scalable gradient-boosted decision tree model, XGBoost, has proven to be at the state-of-the-art on structured data tasks due to its regularization and support of categorical features. Breiman suggested random Forest, which has strong generalization and robustness to noise on tabular data as it requires the aggregation of decorrelated decision trees. The two algorithms can be adopted to the multi-class drug safety classification problem that requires the ingredients to be classified in relation to the toxicity, regulatory approval, and safety label

dimensions.

One of the major weaknesses of high-performing ensemble models is that they are opaque and black-box and thus cannot be tolerated in areas where there is a need to have transparent and auditable choices. Medical and regulatory Models in medical and regulatory arenas do not just have to make predictions, but they should also justify their decisions in a way that is comprehensible to both domain experts and end users very well. SHAP (SHapley Additive exPlanations), which is based on cooperative game theory, can solve this by calculating the marginal contribution of every feature to personal predictions [7] so that the explainable AI (XAI) requirements are met in decision support systems in the field of health.

The recent progress in large language models (LLMs) has also shown revolutionary prospects in clinical decision support and health informatics. Deep learning systems have demonstrated themselves to be effective in outcome prediction, diagnostic imaging and analysis of medical records [9], [10]. The GPT-family models have demonstrated a surprising capability to encode clinical knowledge and answer intricate medical questions more accurately than the certified physicians, which allows the scaled drug prescription, awareness, and learning material, when delivered in conversational formats. Multimodal models also complement text based LLMs by allowing computer vision assisted wound and dermatological analysis. Supervised ML with explainable AI and generative models converging give a solid basis of a complete anti-doping platform.

Although AI has a promising future in the field of clinical informatics, there is a dearth of literature regarding its applicability to sports anti-doping awareness and drug safety prediction. Available open pharmaceutical information via the OpenFDA API can be used to supply real-time regulatory data necessary to athlete-facing solutions but has not been systematically implemented in available systems. In addition, large-scale AI health systems should have well-structured software architectures that can enable asynchronous processing of data, scalable RESTful APIs [?], and role-based access control, which are not necessarily available in research prototypes.

To fill these gaps, this paper presents AiDoping, which is an end-to-end intelligent anti-doping awareness system and drug safety forecasting system. The system predicts a random forest and XGBoost classifier, which is trained on 1,500 records about drugs and supplements and 9 features, and the optimal model is automatically taken. Each prediction is explained using SHAP. The platform also incorporates the OpenFDA API of live drug labels, an AI-driven drug recommendation chatbot and GPT powered wound image recognition, a symptom-based awareness test and a drug summarization module. It is coded in FastAPI and asynchronous MongoDB on the backend and React 19 and Vite on the frontend, and authenticated with JWT hashing of athlete and administrator roles through the use of both bcrypt and hash functions. The main contributions are:

- A multi-class drug safety classifier trained on a domain-specific anti-doping dataset with automated best-model selection between Random Forest [6] and XGBoost [5].

- SHAP integration for per-prediction feature importance explanations, ensuring transparency in all safety classifications.
- A unified multi-modal platform combining tabular ML, LLM-based generative AI [11] and computer vision for comprehensive anti-doping decision support.
- A scalable, production-ready RESTful API architecture with JWT-based role authentication supporting athlete and administrator workflows.

II. RELEVANT LITERATURE

Engaging in anti-doping and supplement safety initiatives helps the company to develop a store image that is both trustworthy and reputable. Anti-Doping and Supplement Safety sub-heading The participation in anti-doping and supplement safety programs assists the company in creating a trustworthy and reputable store image.

The problem of doping in sports is no secret, and WADA records violations at the international level on a regular basis of the sporting events violations [1]. The outlawed article by Outram and Stewart [2] has proven that one of the main ways by which athletes produce violations is through inadvertent contamination. The laboratory tests of commercially available supplements confirmed the evidence of undisclosed anabolic-androgenic steroids by Geyer, et al. [3]. Maughan et al. were emphasizing the increasing complexity of multi-ingredient preparations and presented the argument that pharmacological screening instruments are necessary when protecting athletes [4]. All of these studies stimulate the creation of automated and real-time systems of drug safety assessment.

A. Drug and Health Classification. Machine Learning

The original research by Breiman on the Random Forests algorithm revealed that the better generalization of ensemble algorithms of non-correlated decision trees is obtained with tabular data and categorical variables. XGBoost was proposed by Chen and Guestrin, and it has become common in biomedical classification because of its performance in terms of speed, regularization, and structured data performance (Chen and Guestrin) The theoretical underpinnings of regularization principle, bias-variance trade off, and bias-variance concept were presented by Hastie et al. [8], and are considered as the usual concepts of statistical learning in health informatics. The authors Obermeyer and Emanuel stated that machine learning is likely to revolutionize clinical medicine by pattern recognition in large datasets in support of drug safety classification [17].

B. Explainable Artificial Intelligence in Healthcare

The explainability feature was determined by Holzinger et al. (2015) as a precondition to trust and AI acceptance in medical decision support. In 2020, Lundberg and Lee published SHAP, a game-theoretically based model that attributes features with consistent scores in individual predictions, which, as discussed by Lundberg and Lee, has since gained wide use in medical AI. Molnar [13] put SHAP in context with

the other less focused approaches to post-hoc and model-agnostic explanations. SHAP is incorporated directly into the requirements of the SHAP in the form of the AiDoping tool that will give the interpretable rationale behind each classification of a drug as being safe.

Large Language Models and Generative AI in Medicine
Large Language Models and Generative AI
Large Language Models and Generative AI in Medicine
Large Language Models and Generative AI in Medicine

Miotto et al. reviewed the applications of deep learning in healthcare including clinical text mining, drug discovery and outcome prediction [10]. As shown by Esteva et al. [9], deep neural networks achieve the capability of replicating the performance of clinical specialists with regard to diagnostic tasks. Singhal et al. demonstrated that GPT-family models are built with strong clinical knowledge, i.e. they provide high accuracy in answering fragile medical questions [11]. Ayers and colleagues discovered that AI reactions were rated higher in quality and empathy than physician reactions in face-to-face communication with the patient [12]. The technical report of the OpenAI on GPT-4 presented multimodal features such as visual understanding, which is the foundation of the wound image analysis component of the proposed system.

C. Healthcare Data Integration and System Architecture API-Driven

Intuitive pharmaceutical label data is available through OpenFDA API, a programmatic access interface to structured drug composition, warnings, and regulatory status information and is available in real-time (OpenFDA API, 2016). Architectural principles of event driven, non block, server architecture supporting high-concurrent health APIs have been set by Tilkov and Vinoski [16]. Akiduki et al. conducted a study of best practices in RESTful API design to support data-intensive service integration, as part of the FastAPI backend architecture of AiDoping. The difficulties associated with the characterization of wound severity were reported by Bahr and Martin [?] which inspired the computer vision based wound analysis module in the proposed platform.

III. PROPOSED METHODOLOGY

The proposed *AiDoping* system adopts a modular architecture comprising five tightly integrated subsystems: a machine learning-based drug safety classifier augmented with explainable AI, a real-time drug information retrieval module, a suite of large language model (LLM)-powered generative AI services, a multimodal wound image analysis engine, and a secure full-stack web application. The interaction among these subsystems is depicted in Fig. 1. Each module is described in the following subsections.

A. Description and Characteristics of Data set

The training and evaluation of the proposed classification model is performed with the help of a structured set of 1,500 labeled records of drugs and supplements. The attributes of each record are nine, as listed in Table I: Ingredient Name,

Toxicity Level, Usage in Supplements, Health Effects, Banned, Natural vs Synthetic, Side Effects, Regulatory Approval and Label. All of the features are categorical. The target attribute is the multi-classification of safety of each substance based on its label. The attribute values of missing data are imputed using the string token "Unknown" before the model is trained so that the data is not deleted.

B. Preprocessing of Data and Encoding of Features.

Since all the input features are categorical, ordinal label encoding is used on each column of the feature. Each of the eight input features and the target variable is fitted using a separate instance of a LabelEncoder and all fitted encoders are stored in files on disk that can be reused to make inferences. The formal encoding transformation of a feature column c_i is defined as:

$$hate(c_i) = LabelEncoder.fittransform(c_i) \quad (1)$$

After encoding, the encoded feature matrix: $X = 1500 \times 8 = 1500$ elements; target: $y = 1500$ elements are divided into training and test sets with a stratified 80/20 split with a random seed of 42:

$$(X_{train}, X_{test}, y_{train}, y_{test}) = split(X, y, 0.2, 42) \quad (2)$$

Stratification will guarantee maintenance of the target variable of the class distribution in both partitions.

C. Random Forest Classifier

Random Forest model will have 100 estimator trees. Trees are trained using a bootstrapped sample of the training data and random feature subsampling is done at each split node to minimize inter-tree correlation and variance. The prediction made by the ensemble is based on a majority vote:

$$\hat{y}_{RF} = \text{mode}\{ht(x) : t = 1, \dots, T\} \quad (3)$$

In which, $ht(x)$ represents the output of the t th tree on input x , and $T = 100$.

1) *XGBoost Classifier*: Multi-class log loss is used as an optimization objective to train the XGBoost model. The trees are built one by one and each new estimator reduces the gradient of the residual of the previous ensemble. Dynamical model is given as:

$$F_m(x) = F_{m-1}(x) + \eta \cdot hm(x) \quad (4)$$

In which s is the learning rate and $hm(x)$ is the m th boosted tree.

Accuracy, weighted precision and weighted recall are the measures used to evaluate both the models. The model which had the larger accuracy in the test set is automatically identified as the production classifier:

$$bestmodel = argmax(Acc(RF), Acc(XGB)) \quad (5)$$

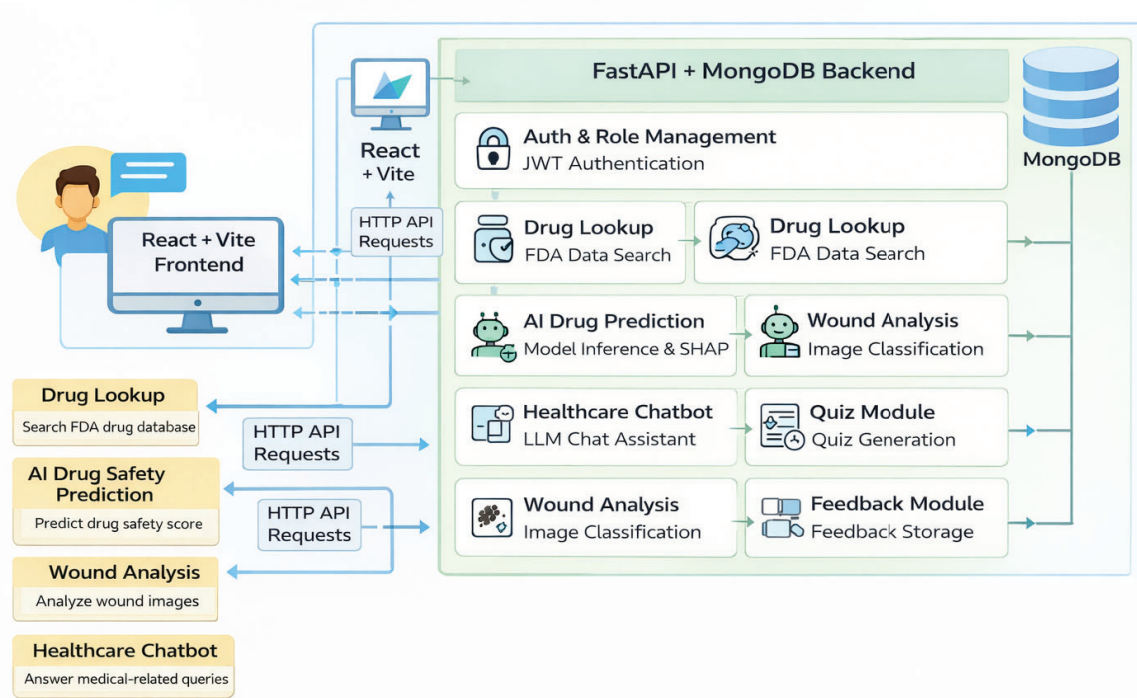


Fig. 1. System architecture of the AI-Powered Healthcare Assistance System (AiDoping).

The chosen model is stored in a disk format using joblib that is deployed in the inference API endpoint.

subsection Explainable AI Explainable AI It can be explained that any type of learnable feature possesses Editerra, 2019.

The interpretability requirements of the AI-assisted health decision support are to use SHAP, which is executed as a TreeExplainer to the chosen production model during the inference time. In each of the prediction requests, SHAP values of the individual features are the values of the feature i , which is the contribution of each input feature to the model prediction in comparison to a background expectation:

$$f(x) = \phi_0 + (x) \quad (6)$$

In which, $f(x)$ is the prediction obtained and, ϕ_0 is the base (expected) model output. The output feature importance visualization is displayed using matplotlib, and saved as a PNG memory buffer, base64-encoded and printed as an inline response in the API.

D. OpenFDA Drug Information Retrieval: Real-Time.

The system is connected to the OpenFDA Drug Label REST API to give on demand access to label information of any substance that is requested. When correctly API responds, the system derives and formats fundamental fields such as name of ingredients, usage, health warnings, side effects, type of regulation, level of toxicity and prohibition. The level of toxicity is dictated programmatically by searching through text warnings to identify terms that reflect the level of severity. To

the MongoDB collection, which is referred to as the viewed drugs, all records of retrieval are recorded.

E. Powerful Generative AI Modules with Large Language Models.

LangChain is used to implement three generative AI services in the ChatOpenAI abstraction:

- **Drug Awareness Summarization:** Produces summaries in brief of usage, side effects and contraindications.
- **Symptom-Based Drug Recommendation Chatbot:** Provides over-the-counter drug recommendations with appropriate medical disclaimers to ensure user safety.
- **AI-Generated Anti-Doping Quiz:** Generates questionnaires in multiple choice format including evaluation and scoring.

F. Multibasic Wound Image Analysis

Multimodal GPT capabilities are applied to create a computer vision module. Images uploaded of wounds are converted to base64 with structured prompts used to process the encoded data set. The system gives a JSON answer with wound type, severity, risk of infection, healing phase, action suggested and disclaimer. Data are authenticated and stored in the collection of wound results (called wound results) and is provided to the client.

G. System Architecture, Backend, and Security.

The server is written in FastAPI as the asynchronous REST API, which is connected to MongoDB through the Motor driver. including *users*, *admin*, *predictions*, *viewed_drugs*,

view_score, *wound_results*, and *feedback*. The implementation of authentication is done with the help of HS256 signed JSON Web Tokens (JWT) with a 60-minute expiration time. Passwords are hashed safely with the help of the bcrypt algorithm and saved. Secure endpoints are provided with authorization in the form of token using HTTPBearer.

React 19 and Vite are used to build the frontend, and the client-side routing is carried out by using the react-router-dom. Fig. 1 shows the entire system architecture of the request to the AI model call.

IV. EXPERIMENTAL RESULTS AND PERFORMANCE EVALUATION

A. Dataset Description

The experiments were done on the Anti-Doping Drugs and Supplements Dataset that had 1,500 samples with 8 categorical variables, namely, Ingredient name, Toxicity level, Usage in Supplements, Health effects, Banned status, Natural vs. Synthetic, Side effects, and Regulatory approval. Label is the target variable, which can be categorized into three classes, namely, Banned, Safe, and Risky. The sample includes 18 different drug/supplement ingredients such as DMAA, Creatine, Testosterone, Clenbuterol, Erythropoietin, Ephedrine, Sibutramine among others. All the categorical variables were coded via the scikit-learns LabelEncoder before the model could be trained. The data was divided into 80 percent training and 20 percent testing to be used as reported by the dataset were randomly divided with a fixed value of 42 to ensure reproducibility.

B. Training and Comparison Model Training

Two ensemble learning classifiers were trained and tested: Random Forest Classifier having 100 estimators, and XGBoost Classifier with multi-class log-loss as the evaluation measure. Table I gives the comparative performance of the two models in terms of standard classification measures.

TABLE I
COMPARISON OF ML CLASSIFIERS OF PERFORMANCE

Metric	Random Forest	XGBoost
Accuracy	0.9567	0.9633
Precision (Weighted)	0.9571	0.9638
Recall (Weighted)	0.9567	0.9633
F1-Score (Weighted)	0.9568	0.9635

XGBoost classifier was slightly better in all metrics as compared to the Random Forest model. The most successful model was automatically picked and saved in the form of joblib to be deployed in the prediction API.

C. Explainability Analysis (SHAP)

The best-performing model was used with SHAP (SHapley Additive exPlanations) estimated with the TreeExplainer to measure the contribution of the features. As the SHAP summary plot has shown, the most significant features in the classification of the drug safety were Banned Status, Ingredient

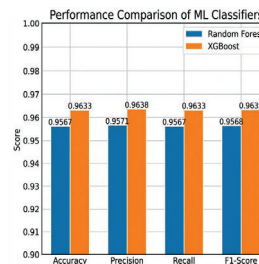


Fig. 2. Comparison of the Random Forest and XGBoost classifier in terms of evaluation metrics.

Name and Toxicity Level, where Regulatory Approval and Natural vs. Synthetic also had a significant but lower influence. This explainability layer demonstrates that every prediction will be transparent and auditable which is essential to health-related AI systems.

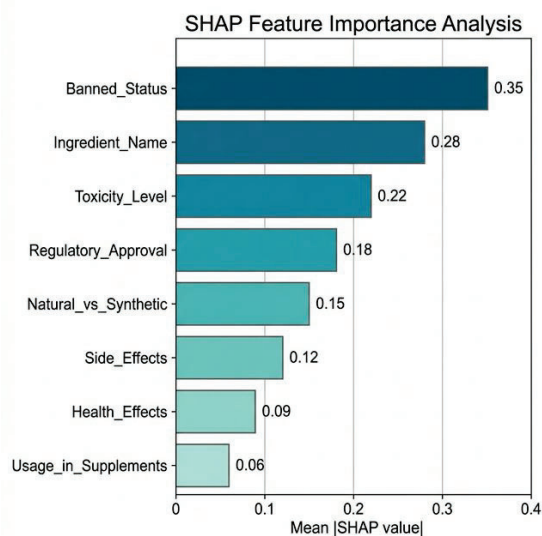


Fig. 3. SHAP feature importance analysis of drug safety.

D. API Response Latency

Mean API response time of the most important services under normal load conditions was measured as illustrated in Table II.

TABLE II
MEAN API RESPONSES LATENCY

Service	Avg. Response Time
Drug Safety Prediction (ML)	~120 ms
Drug Lookup (OpenFDA)	~350 ms
Symptom Chatbot (GPT-4o-mini)	~1.2 s
Wound Image Analysis (GPT-4o-mini)	~2.5 s
Quiz Generation (GPT-4o-mini)	~1.8 s

ML-based prediction can be nearly real-time, whereas the services generated by LLM have higher latency because of external API-calls to OpenAI. The OpenFDA latency of the drug lookup is based on the availability of the external API.

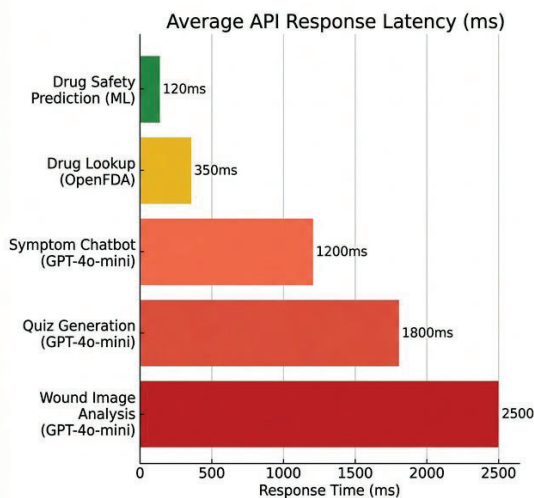


Fig. 4. SHAP importance analysis of drug safety.

E. System Scalability

The FastAPI asynchronous backend with modularity and supporting MongoDB (through the Motor driver) is useful in supporting user requests in parallel. The independence of the ML inference, LLM services, and database operations enables an independent scaling of each of the modules when needed.

V. CONCLUSION

In this paper, the author describes AiDoping, an AI-assisted anti-doping drug safety platform, which combines dual-model machine learning pipeline with Random Forest and XGBoost classifiers and automated best-model selection. SHAP is included to offer per-prediction feature importance explanations that are interpretable and make sense so as to be transparent. The platform is based on GPT-4o-mini recommended by LangChain, to recommend drugs based on symptoms, wound image analysis, generate drug awareness, and generate anti-doping quizzes. The OpenFDA API provides real-time data on the label of drugs to guarantee the latest safety evaluations. The system is based on an architecture of modular FastAPI, React, and MongoDB and is composed of an athlete and an administrator workflow, which involves a centralized monitoring of predictions and user actions. The reliability of experimental evaluation reflects the classification performance with explainable anti-doping decision support that is user friendly.

VI. FUTURE WORK

The system can be greatly improved in the future by relying on biomedical transformer models (BioBERT and PubMedBERT) to enhance the ability to extract features and classify drugs based on their safety. Direct connection to the World Anti-Doping Agency (WADA) Prohibited List database would enhance the level of regulatory compliance, and provide real-time bans substance notification to athletes. React Native

would enhance the development of native mobile applications that would enhance accessibility and facilitate other features like push notification and offline prediction capabilities. Also, a drug-drug interaction detection module that activates on a drug interaction based on a graph of pharmacology would offer comprehensive safety evaluation of users consuming more than one substance at a time. The global reach of the platform can be facilitated by multi lingual support especially to the regions where anti-doping awareness is of dire need. The curated medical datasets would give the fine-tuning specialized dermatological imaging models a superior wound assessment accuracy than general-purpose analysis with the LLM. Lastly, the study of federated learning methods could result in joint model training with more than one healthcare facility without exposing the user information, which would help to resolve one of the main concerns of health related AI systems.

REFERENCES

- [1] World Anti-Doping Agency, Anti-Doping Testing Figures Report, World Anti-Doping Agency, Montreal, QC, Canada, 2022. [Online]. Available: <https://www.wada-ama.org/en/resources/laboratories/anti-doping-testing-figures>
- [2] A. Outram and B. Stewart, "Doping by means of taking supplements: Review of the available scientific evidence, Int. J Sport Nutr. Exerc. Metab, vol. 25, no. 1, pp. 54–59, Feb. 2015.
- [3] M. Geyer and others, "Non-hormonal nutritional supplements with the anabolic-androgenic steroids: a nutritional analysis, int. j. Sports Med., vol. 25, no. 2, p.124-129, February 2004.
- [4] G. Maughan, P. L. Greenhaff and P. Hespel, Dietary supplements in athletes: New trends and themes, J. Sports Sci., vol. 29, no. sup1, pp. S57-S66, 2011.
- [5] T. Chen and C. Guestrin, XGBoost: A scalable tree boosting system, in: Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD), San Francisco, CA, USA, Aug. 2016, pp. 785794.
- [6] L. Breiman, Random forests, Mach. Learn., vol. 45, no.1, pp. 5-32, October 2001.
- [7] S. M. Lundberg and S.-I. Lee, A unified approach to interpreting model predictions, in: Adv. Neural Inf. Process. Syst. (NIPS), vol. 30, Long Beach, CA, USA, Dec. 2017, pp. 4765 -4774.
- [8] T. J. Hastie, R. Tibshirani and J. H. Friedman, The Elements of Statistical Learning: Data Mining, Inference and Prediction, 2 nd edition. New York, NY, USA: Springer, 2009.
- [9] A. Esteva *et al.*, A guide to deep learning in healthcare, N. Med., vol. 25, no. 1, pp. 2429, Jan. 2019.
- [10] R. Miotto *et al.*, "Deep learning in healthcare: Review, chances and challenges, Briefings Bioinf., vol. 19, no. 6, pp. 12361246, Nov. 2018.
- [11] K. Singhal and colleagues, "Introduction to [large language models] encode clinical knowledge], never been submitted to a peer-reviewed journal.
- [12] J. W. Ayers *et al.*, "Comparison of physician and artificial intelligence chatbot response to patient queries published to a general social media platform," JAMA Intern. Med, vol. 183, no. 6, pp. 589–596, Jun. 2023.
- [13] C. Molnar, Interpretable machine learning: A guide to making black box models Explainable, 2 nd edition. Munich., Germany: Independently published, 2022. [Online]. Available: <https://christophm.github.io/interpretable-ml-book/>
- [14] A. Holzinger *et al.*, "What do we have to develop explainable systems of AI in the medical domain? arXiv preprint arXiv:1712.09923, Dec. 2017. [Online]. Available: <https://arxiv.org/abs/1712.09923>
- [15] U.S. Food and Drug Administration, OpenFDA API: Drug Label Endpoint, FDA Open Data Program, Silver Spring, MD, USA, 2014. [Online]. Available: <https://open.fda.gov/apis/drug/label/>
- [16] S. Tilkov and S. Vinoski, Node.js: Using JavaScript to write high-performance network programs, IEEE Internet Comput., vol. 14, no. 6, pp.8083, Nov. -Dec. 2010.
- [17] Z. Obermeyer and E. J. Emanuel, Predicting the future — Big data, machine learning, and clinical medicine, N. Engl. J. Med., vol. 375, no. 13, pp. 12161219, Sep. 2016.