Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 22 No. 4 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v22i04.59703

PAPER

BioBERT-XGBoost for Adverse Drug Reaction Prediction: An Interpretable Hybrid Model for Risk-Aware Pharmacovigilance

Alexandra Ramirez[1], Raul Pingo[1], Sandra Wong-Durand[1] (✉), Pedro Castañeda[2], Alejandra Oñate-Andino[3]

[1]Universidad Peruana de Ciencias Aplicadas, Lima, Peru

[2]Universidad Nacional Toribio Rodriguez de Mendoza (UNTRM), Amazonas, Peru

[3]Escuela Superior Politécnica de Chimborazo (ESPOCH), Riobamba, Ecuador

pcsiswon@upc.edu.pe

ABSTRACT

Adverse drug reactions (ADRs) are a critical challenge for patient safety, with over 21,000 alerts reported in Peru in 2024. Current artificial intelligence (AI) models in pharmacovigilance present limitations in external validation, clinical scalability, and algorithmic transparency. This work proposes BioBERT-XGBoost, an interpretable hybrid model that combines biomedical natural language processing with supervised machine learning to predict ADRs. The architecture integrates BioBERT for semantic extraction of pharmacological entities with XGBoost as a calibrated classifier, trained on public datasets (DrugBank, openFDA–FAERS) and anonymized clinical records. The pipeline includes standardized preprocessing through normalized vocabularies, feature engineering with semantic embeddings, class imbalance handling, and probability calibration. Evaluation uses discrimination metrics (AUROC, AUPRC), calibration (Brier score), and explainability (SHAP). The system is deployed on Microsoft Azure through a mobile application that generates risk-stratified clinical alerts, representing a step toward trustworthy clinical decision-support systems for proactive ADR detection.

KEYWORDS

adverse drug reactions (ADRs), pharmacovigilance, biomedical natural language processing (NLP), hybrid machine learning, clinical prediction models, semantic embeddings, risk stratification, patient safety systems

1 INTRODUCTION

Adverse drug reactions (ADRs) represent a critical problem in patient safety, as they increase healthcare costs, prolong hospital stays, and can be fatal in vulnerable populations. In Peru, DIGEMID reported more than 21,000 alerts in 2024, mainly in Lima, La Libertad, and Lambayeque, reflecting the magnitude and frequency of these events [1]. At the international level, it has been shown that therapeutic complexity

Ramirez, A., Pingo, R., Wong-Durand, S., Castañeda, P., Oñate-Andino, A. (2026). BioBERT-XGBoost for Adverse Drug Reaction Prediction: An Interpretable Hybrid Model for Risk-Aware Pharmacovigilance. *International Journal of Online and Biomedical Engineering (iJOE)*, 22(4), pp. 107–122. https://doi.org/10.3991/ijoe.v22i04.59703

Article submitted 2025-11-25. Revision uploaded 2025-12-29. Final acceptance 2025-12-30.

significantly increases the risk of hospitalization due to ADRs, as evidenced by a prospective study in people living with HIV [2].

Artificial intelligence (AI), particularly machine learning (ML), has enabled novel approaches to anticipate ADRs from electronic health records (EHRs). Recent reviews report mean AUC values of approximately 0.81 [3], confirming ML's potential for preventive medicine through early alerts. This approach is particularly crucial for patients with known allergies or polypharmacy, where proactive prevention can avert hospitalizations.

Several studies have applied ML in pharmacovigilance with promising results. A systematic review reported sensitivities and specificities close to 78%, although with limited external validation [4]. Graph neural networks have also been developed to detect drug–drug interaction–related side effects [5], and systems such as PreAlgPro have improved the prediction of allergenic proteins compared to traditional methods [6]. However, there is still a gap between these tools and their stable integration into clinical practice. Current models still exhibit three main weaknesses: lack of external validation [4], high technological demands [5], and limited transparency of algorithms, which generates distrust. There is a need to move toward simpler, auditable, and scalable solutions that can be adapted to different clinical contexts.

This work proposes a modular intelligent model that integrates: (i) a natural language processing (NLP) pipeline (BioBERT) to extract pharmacological entities and detect allergens and (ii) a predictive module based on XGBoost, trained with DrugBank and openFDA. Its innovation lies in targeting digital preventive medicine, generating verifiable, real-time alerts, and prioritizing decision traceability.

The article is organized as follows: Section 2 presents related work; Section 3 describes the methodology and model architecture; Section 4 reports initial results; and Section 5 presents conclusions and clinical implications, along with future pilot tests in hospital environments.

2 RELATED WORKS

Recent evidence highlights the clinical burden of adverse drug reactions across different populations. Population-level studies report substantial mortality related to adverse drug events [7], while observational work has identified high-risk groups such as patients using new antiseizure medications [8] and underserved populations where reporting gaps persist [9]. Similarly, research on cardiovascular drugs [10] and prospective studies in hospitalized older adults have revealed potential "prescribing cascades," emphasizing how regimen complexity can propagate avoidable harm [11].

A persistent limitation of pharmacovigilance systems is underreporting and the resulting uncertainty about actual risks. Expert reviews have outlined the causes and potential solutions to this problem [12], while recent studies based on the FDA Adverse Event Reporting System (FAERS) show how signal detection can reveal disproportionate associations—such as those observed with omalizumab—often without clinical confirmation or causal designs to support such correlations [13].

Data-driven and machine learning–based approaches have emerged in response to these gaps. A systematic review in *Research in Social & Administrative Pharmacy* synthesizes the use of AI to predict ADRs in hospitalized patients and emphasizes the importance of external validation and real-world evaluation [4].

Methodologically, recent models leverage graph neural networks with self-supervised pre-training to learn the structure of drug–drug interaction networks

for side-effect prediction [5], hybrid deep learning (InceptionV3–LSTM) for multilabel ADRs to COVID-19 medications [14], and multi-relation in silico frameworks to estimate high-level drug–drug and drug–disease adverse effects [15].

Taken together, this evidence underscores both the clinical magnitude of ADRs and the limitations of current pharmacovigilance systems, justifying the exploration of more robust, externally validated models that are applicable in real clinical environments. Recent systematic reviews have demonstrated the effectiveness of artificial intelligence techniques, including machine learning and deep learning, in medical imaging for disease classification and segmentation, highlighting the broader potential of intelligent systems across multiple clinical decision support domains [16].

3 SYSTEM DESIGN

3.1 Architecture

Logical architecture. The logical architecture of the system is organized into modular layers that ensure scalability, security, and traceability in the management of clinical information. The main focus lies in integrating a BioBERT-XGBoost machine learning model, which predicts ADR risk based on clinical and pharmacological data.

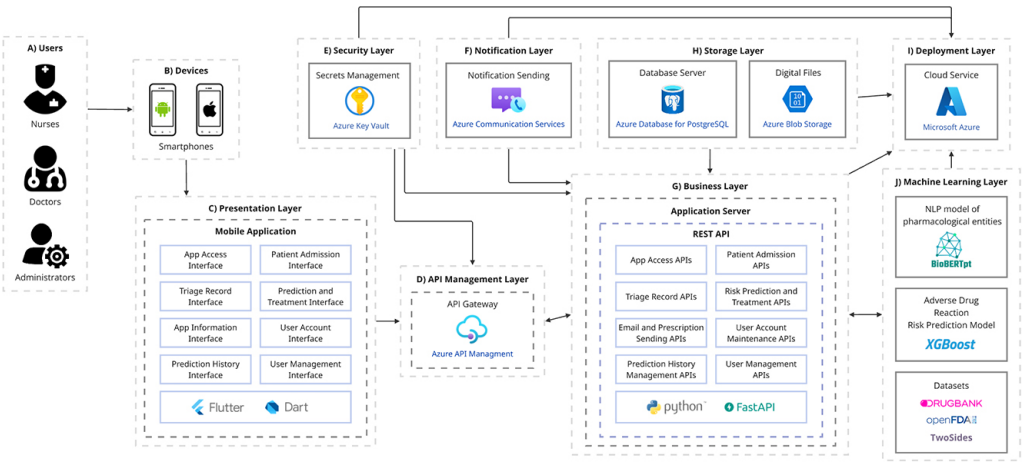

Fig. 1. Logical architecture of the mobile application with integration of the ADR prediction model

In Figure 1, the operation of the mobile application is organized in layers, where the ADR prediction model interacts with the rest of the system. In A) Users, physicians, nurses, and administrators access the system through C) the presentation layer, where the mobile app allows the registration of patient data, triage, and risk queries. These requests are routed to G) the business layer, where a FastAPI server manages REST APIs that connect the interface with the predictive model. Information is stored in H) Storage Layer, using Azure PostgreSQL and Azure Blob storage. Security and access control are managed in E) Security Layer via Azure Key Vault, while D) API Management Layer uses Azure API Management for API governance. The core of the system lies in J) ML Layer, which integrates BioBERT for semantic extraction of pharmacological entities and XGBoost for risk classification, trained with datasets such as DrugBank and openFDA. The intelligent decision

support architecture follows established patterns for AI-driven support systems that prioritize user-centered design, interpretability, and actionable clinical recommendations [17]. Results are returned as clinical alerts within the application, strengthening digital preventive medicine. The entire ecosystem runs over I) Deployment Layer, implemented on Microsoft Azure, ensuring availability and reliable operation. The mobile application interface leverages best practices in interactive mobile technologies to ensure accessibility, usability, and seamless integration into clinical workflows for healthcare professionals [18].

3.2 Methodology

Dataset collection. For model training and validation, we will combine open-source data with locally obtained clinical records. Internationally, we will use the openFDA-FAERS repository, which contains ADR reports coded in MedDRA and enables signal analysis based on spontaneous reports of potential side effects [19]. We will also use DrugBank (Open Data), which provides detailed information on drug properties, ATC classification, and known interactions and contraindications [20]. Both databases are openly accessible and have been used in previous AI-based ADR prediction studies. Complementarily, we will include local records from electronic health records (EHR) and hospital triage data. These will contain demographic variables (age, sex), clinical background (comorbidities, polypharmacy), and medical prescriptions. Before processing, all information will be anonymized and evaluated by the institution's Research Ethics Committee. The protocol will be reviewed in accordance with the Peruvian General Health Law (Law No. 26842, Art. 24–26) [21], ensuring formal approval and compliance with current ethical regulations.

Data preprocessing. Data preprocessing is a critical stage to ensure the quality, consistency, and standardization of clinical data, facilitating its integration into a reproducible, traceable framework for predictive modeling.

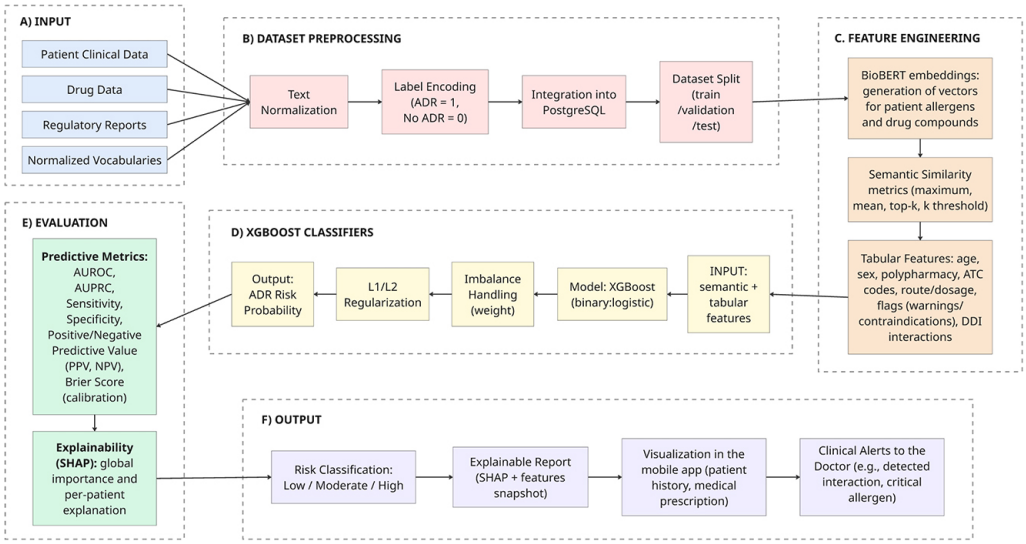

Fig. 2. Pre-processing and feature engineering flow for the ADR prediction model

Figure 2 shows the preprocessing and feature engineering flow designed for the ADR prediction model. The process begins at A) Input, integrating patient clinical

data, pharmacological information, pharmacovigilance reports (openFDA–FAERS, DrugBank), and standardized vocabularies such as RxNorm, ATC, MedDRA, ICD-10, and SNOMED CT. In B) Dataset Preprocessing, data is normalized, binary labels are encoded (ADR = 1; No ADR = 0), and records are consolidated in PostgreSQL, combining clinical histories, prescriptions, and adverse events. The dataset is then split into training, validation, and test sets using stratification strategies that avoid information leakage. In C) Feature Engineering, BioBERT embeddings are generated to represent allergens and drugs as dense semantic vectors. These are combined with similarity metrics (maximum, average, top-k, and threshold) that estimate the distance between the patient's allergy profile and prescribed medications, along with tabular variables such as age, sex, polypharmacy, ATC codes, route and dosage, regulatory warnings, and potential drug–drug interactions (DDIs). In D) XGBoost Classifiers, all features are fed into the model, which applies L1/L2 regularization, class balancing, and binary calibration to estimate the probability of ADR risk. In E) Evaluation, performance is measured via AUROC, AUPRC, sensitivity, specificity, predictive values, and the SHAP explainability module, which interprets predictions both globally and per patient. Finally, in F) Output, the system classifies risk (low, moderate, or high), generates an interpretive report highlighting relevant variables, and shows clinical alerts in the mobile app, supporting medical decision-making and early prevention of adverse events.

Model development. Model development applies tree-ensemble machine learning algorithms to predict ADR risk, optimizing both predictive performance and clinical interpretability.

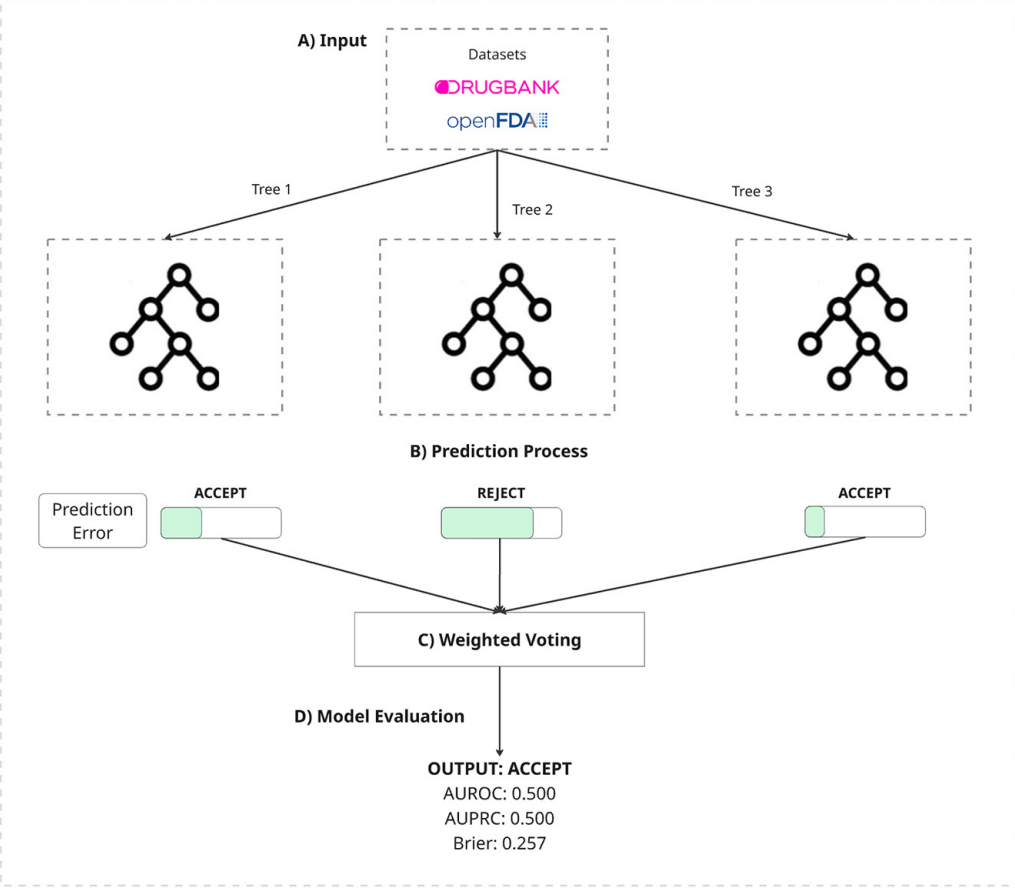

Fig. 3. Development of the XGBoost model for ADR risk prediction

Figure 3 shows the architecture of the XGBoost-based system, organized as an ensemble of decision trees with weighted voting. In A) Input, data from DrugBank and openFDA feed three classifiers (Trees 1–3). Each tree is trained sequentially, correcting the errors of the previous one, which is the central principle of boosting. This approach captures nonlinear relationships between clinical and pharmacological variables, while L1/L2 regularization reduces overfitting and improves generalization. In B) Prediction Process, each classifier generates a binary output (ACCEPT/REJECT) and a confidence score. In C) Weighted Voting, predictions are combined according to their performance on the validation set to produce a final, calibrated decision about ADR risk. Finally, in D) Model Evaluation, the diagram illustrates the iterative development workflow, including logging mechanisms for prediction errors and model diagnostics. It is important to note that the values shown in this figure (AUROC = 0.500, AUPRC = 0.500, Brier = 0.257) represent early-stage baseline results during initial model prototyping with unoptimized hyperparameters and minimal feature engineering. These preliminary metrics served to validate the technical pipeline infrastructure and identify critical areas for improvement. The baseline performance essentially represents random classification (AUROC ≈ 0.5), confirming the need for systematic optimization. Following comprehensive optimization—including preprocessing refinement, feature engineering with BioBERT semantic embeddings (768-dimensional vectors), class balancing through SMOTE, hyperparameter tuning via grid search, and probability calibration using isotonic regression—the final model achieved substantially improved performance metrics, which are comprehensively reported in Section 4 (Results): AUROC = 0.9726, AUPRC = 0.9596, and Brier Score = 0.0604.

Training and testing. The model was trained by integrating structured clinical variables—age, sex, comorbidities, and polypharmacy—with semantic features derived from free-text sources. These semantic features were generated using biomedical embeddings from BioBERT, a language model pre-trained on clinical and biomedical literature that captures contextual relationships between drugs and allergens, thereby enriching information beyond tabular data [22]. For robust evaluation, the dataset comprising 2,655 total records was partitioned using stratified sampling to prevent information leakage. The split was configured as follows: 2,124 records (80.0%) for training and 531 records (20.0%) for testing. Stratification maintained consistent class distribution across splits, with 44.54% serious ADR events in the training set (946 serious vs. 1,178 non-serious) and 44.63% in the test set (237 serious vs. 294 non-serious). This stratified approach ensures that both splits accurately represent the natural prevalence of serious adverse events in the population, preventing temporal leakage and information contamination between training and evaluation phases.

The training set represents data collected from multiple healthcare facilities in Peru, including hospital emergency departments, outpatient pharmacies, and primary care centers. Data sources were aggregated from internationally validated databases (DrugBank for drug properties and interactions, openFDA-FAERS for spontaneous adverse event reports) alongside locally anonymized electronic health records from Peruvian healthcare institutions. This multi-source, geographically diverse approach enhances model generalizability across different clinical settings, patient populations, and prescribing patterns common in Latin American healthcare systems. Training was performed with the XGBoost algorithm configured for binary classification (binary:logistic), incorporating L1/L2 regularization, early stopping, and tuning of the scale_pos_weight parameter to address the strong class imbalance typical of rare ADRs [23]–[24].

Class imbalance was managed through a rigorous two-stage strategy designed to prevent data leakage while ensuring robust minority class representation. First, Synthetic Minority Oversampling Technique (SMOTE) was applied exclusively to the training set after train-test partitioning. SMOTE, configured with k_neighbors = 5, generates synthetic samples by interpolating between existing minority-class instances in feature space, creating new examples that share characteristics with real serious ADR cases without exact duplication. The algorithm balanced the training set from its original imbalanced distribution (946 serious vs. 1,178 non-serious cases) to a perfectly balanced 1:1 ratio (1,178 serious vs. 1,178 non-serious), yielding 2,356 total training samples. Second, XGBoost's scale_pos_weight hyperparameter was set to 1.0, as the perfect 1:1 balance achieved by SMOTE eliminated the need for additional algorithmic class weighting during gradient boosting.

Critically, SMOTE was applied only after the train-test split and exclusively to training data. The test set remained completely unaugmented, preserving its natural class distribution (237 serious, 294 non-serious) to ensure that performance evaluation reflects true generalization capability on real-world, imbalanced data. This methodological approach prevents the common but severe error of applying oversampling before splitting, which artificially inflates performance metrics by allowing synthetic samples to appear in both training and test sets. Our strategy guarantees that all reported test set metrics (AUROC, AUPRC, sensitivity, specificity) represent authentic predictive performance on unseen, naturally distributed clinical data.

Finally, a probability calibration step was applied to ensure clinically interpretable outputs. Two widely validated techniques were evaluated—Platt scaling and isotonic regression—selecting the one that achieved the lowest Brier score on the validation set. This calibration allowed model outputs to be interpreted as absolute risks and enabled patient stratification into low-, moderate-, and high-risk categories. Probability calibration is essential in clinical decision-support systems, as the usefulness of predictions depends not only on classification accuracy but also on the reliability of estimated probabilities [25].

Performance evaluation metrics. The performance of the ADR prediction model was assessed using metrics spanning discrimination, calibration, and clinical utility. For discrimination, we used the area under the receiver operating characteristic curve (AUROC), which summarizes the model's ability to distinguish between patients with and without ADRs, and the area under the precision–recall curve (AUPRC), which is more suitable for imbalanced data as it emphasizes performance on the positive class.

To evaluate probability calibration, we used the Brier Score, defined as the mean squared error between predicted probabilities and actual outcomes. This metric jointly reflects aspects of discrimination and calibration; a low value indicates that predicted probabilities can be interpreted clinically as absolute risks [26]. We also used calibration plots and derived metrics such as expected calibration error (ECE). The confusion matrix metrics are calculated using four fundamental classifications: True Positives (TP) represent patients correctly identified as having ADR risk; False Positives (FP) are patients incorrectly flagged as at-risk when they are not; True Negatives (TN) are patients correctly identified as not at-risk; and False Negatives (FN) are patients incorrectly classified as not at-risk when they actually are. These fundamental measures form the basis for calculating sensitivity (TP/(TP + FN)), specificity (TN/(TN + FP)), positive predictive value (TP/(TP + FP)), and negative predictive value (TN/(TN + FN)) described in Table 1.

From a clinical perspective, we calculated confusion matrix–based metrics: sensitivity (recall), which reflects the system's ability to detect most at-risk patients;

specificity, which measures the ability to correctly identify non-risk patients; positive predictive value (PPV), which measures the proportion of alerts that truly correspond to ADRs; and negative predictive value (NPV), which indicates the safety of considering a patient not at risk when no alert is generated [27]. Together with trade-off analyses between sensitivity and specificity, these metrics allow tuning of decision thresholds according to clinical priorities (e.g., prioritizing sensitivity in severe cases).

Finally, to guarantee interpretability, we applied the SHAP (SHapley Additive exPlanations) framework, which decomposes each prediction into contributions from individual features. This enabled us to validate that the most influential factors were clinically coherent and to provide patient-level explanations to healthcare professionals, increasing trust in the system [28].

Table 1. Metrics and criteria used for evaluation and validation of the ADR prediction model

Heading Level	Formula	Clinical Utility
AUROC	Area under the TPR vs. FPR curve	Model comparison; threshold-independent
AUPRC	Area under the Precision–Recall curve	More informative with rare classes (infrequent ADRs)
Sensitivity (Recall)	$\frac{TP}{TP+FN}$	Minimizes false negatives; key for patient safety
Specificity	$\frac{TN}{TN+FP}$	Reduces false positives and alert fatigue
PPV (Precision)	$\frac{TP}{TP+FP}$	Reliability of alerts; physician confidence
NPV	$\frac{TN}{TN+FN}$	Clinical reassurance when ruling out risk
Brier Score	$\left(\frac{1}{N}\right)\Sigma\left(\hat{y}_i - y_i\right)^2$	Evaluates calibration and probabilistic accuracy
SHAP (explainability)	–	Global and local explanations; increases clinical trust

4 RESULTS

4.1 Performance of the predictive model

The XGBoost-based predictive model was evaluated on a stratified dataset with 2,356 records for training and 531 for testing. The test set included 237 serious adverse events and 294 non-serious cases, maintaining a balanced distribution through synthetic oversampling techniques. On the test set, the model achieved an AUROC of 0.9726, substantially surpassing the 0.90 threshold commonly considered indicative of excellent discrimination in clinical contexts. This value reflects an exceptional capacity to distinguish between patients who will experience serious ADRs and those who will not. The AUPRC of 0.9596 confirms robust performance in identifying the minority (positive) class, a particularly relevant metric given the inherent imbalance of pharmacovigilance data (refer to Table 2).

Table 2. Comparison of model performance metrics between training and test sets

Metric	Training	Test	Difference
AUROC	0.9962	0.9726	0.0236
AUPRC	0.9948	0.9596	0.0352
Accuracy	98.34%	93.60%	4.74%
F1-score	0.9836	0.9274	0.0562
Sensitivity	99.32%	91.56%	7.76%
Specificity	97.37%	95.24%	2.13%
Precision (PPV)	97.42%	93.94%	3.48%

In Table 2, moderate differences between training and test (4.74% in accuracy, 2.36 percentage points in AUROC) indicate adequate generalization of the model. The confusion matrix for the test set shows 280 true negatives, 217 true positives, 14 false positives, and 20 false negatives. A sensitivity of 91.56% indicates that the model correctly identifies approximately nine out of ten serious ADRs, while a specificity of 95.24% reflects a strong ability to correctly discard non-serious cases. The PPV of 93.94% provides high confidence in generated alerts, and the NPV of 93.33% offers robust reassurance that patients classified as low risk indeed have a reduced probability of serious events (see Figure 4).

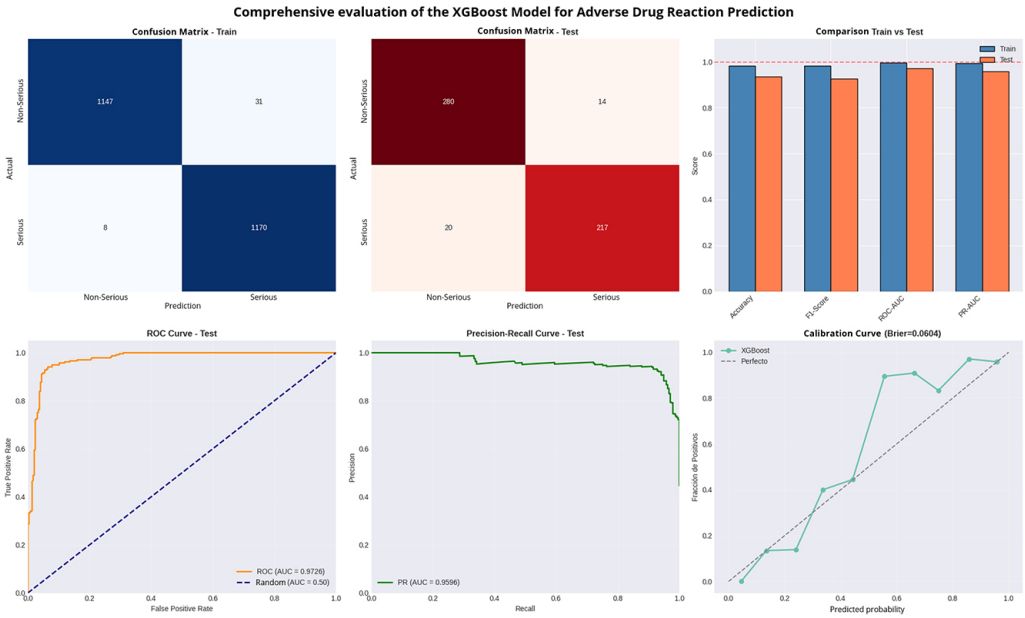

Fig. 4. Comprehensive evaluation of the XGBoost model for ADR prediction: confusion matrices, ROC–PR curves, and calibration analysis

In Figure 4, the top-left panel shows the confusion matrix for the training set (TN = 1147, FP = 31, FN = 8, TP = 1170). The top-center panel shows the confusion matrix for the test set (TN = 280, FP = 14, FN = 20, TP = 217). The top-right panel compares metrics between training and test. The bottom panels display ROC (AUC = 0.9726), precision–recall (AUC = 0.9596), and calibration (Brier = 0.0604) curves.

4.2 Probability calibration

The calibrated model achieved a Brier Score of 0.0604, a notably low value indicating excellent concordance between predicted risks and observed event rates. The calibration curve in Figure 4 shows that predictions consistently approximate the ideal diagonal, particularly in the 0.1–0.6 probability range where most predictions are concentrated.

Table 3. Calibration of the ADR prediction model by deciles of estimated probability

Bin	Predicted Prob.	Observed Frequency	Difference	Status
1	0.044	0.000	0.044	☑
2	0.135	0.135	0.000	☑
3	0.239	0.400	–0.161	⚠
4	0.317	0.438	–0.121	⚠
5	0.443	0.444	–0.002	☑
6	0.556	0.895	–0.339	⚠
7	0.663	0.909	–0.246	⚠
8	0.749	0.833	–0.084	☑
9	0.839	1.000	–0.161	⚠
10	0.958	0.958	0.000	☑

In Table 3, ☑ indicates excellent calibration (|difference| < 0.10); ⚠ indicates moderate deviation. Calibration is strongest at low and high probability extremes, with deviations in the intermediate range. Three risk categories were defined: low (probability < 0.30), moderate (0.30–0.70), and high (>0.70). Patients classified as high risk had a 70.0% incidence of serious ADRs, more than five times higher than the low-risk group (12.8%), validating the model's ability to effectively stratify clinical risk (see Table 4).

Table 4. Risk stratification for ADRs and observed event frequency by probability category

Category	Probability Range	Patients (n)	Percentage	ADRs Observed	Incidence
Low	< 0.30	187	35.2%	24	12.8%
Moderate	0.30 – 0.70	194	36.5%	108	55.7%
High	> 0.70	150	28.3%	105	70.0%

4.3 Feature importance and clinical interpretability

The feature importance analysis from XGBoost revealed that BioBERT semantic embeddings (prefix drug_emb_) dominate the ranking, with 18 of the 20 most important features. The dimensions drug_emb_504, drug_emb_597, and drug_emb_480 showed the highest contributions. Although these dimensions lack direct semantic interpretation, they represent latent combinations of pharmacological properties learned by BioBERT during pre-training on biomedical literature. Among the most relevant structured variables are impaired renal function, history of previous ADRs,

advanced age (≥75 years), polypharmacy (>5 drugs), and multiple documented allergies (see Figure 5).

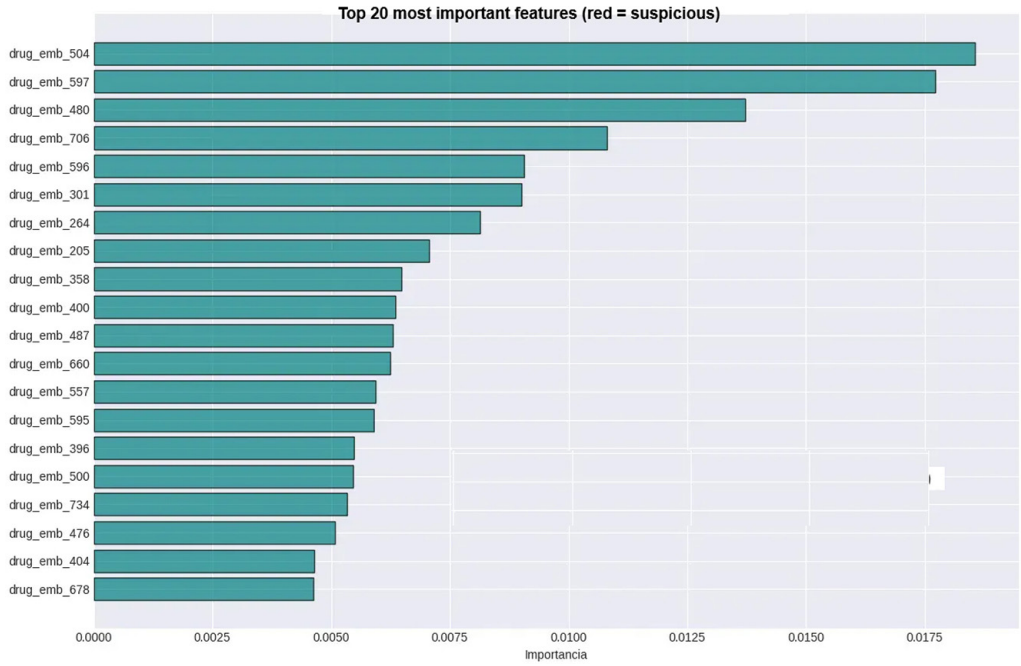

Fig. 5. Top twenty features with highest relative importance in the XGBoost ADR prediction model

In Figure 5, bars represent relative importance according to XGBoost's gain metric. Features prefixed with drug_emb_ are BioBERT semantic embedding dimensions.

4.4 Clinical impact demonstration

Three representative cases illustrate clinical applicability: Case 1 (low risk): Woman, 25 years, no comorbidities, paracetamol. Predicted probability: 23.3%. Case 2 (High risk): Man, 82 years, multiple comorbidities, reduced renal function (eGFR 22 mL/min), four prior ADRs, seven medications, warfarin. Probability: 100.0%. Case 3 (Moderate risk): Woman, 55 years, diabetes, hypertension, metformin, and additional medications. Probability: 33.9%. These cases demonstrate how the system generates personalized risk estimates to support clinical decisions.

5 DISCUSSION

5.1 Key findings and comparison with previous work

The hybrid BioBERT-XGBoost model achieved an AUROC of 0.9726, placing it at the upper end of ADR prediction systems (typically 0.75–0.88). The AUPRC of 0.9596 is especially relevant given pharmacovigilance data imbalance. While previous studies report sensitivities around 78% [4], this model achieved 91.56% sensitivity with 95.24% specificity. BioBERT embeddings dominating feature importance (90% of top 20, see Figure 5) confirms that linguistic representations encode substantial predictive information. The Brier Score of 0.0604 indicates superior calibration versus typical values (0.10–0.25).

5.2 Clinical implications and implementation considerations

The model generates calibrated probabilistic estimates that enable personalized clinical decisions. High specificity (95.24%) and PPV (93.94%) suggest the system can be integrated without excessive alert fatigue. Implementing tiered alerts (interruptive for probabilities >0.70, passive for 0.30–0.70) could optimize usability. Explainability through feature importance analyses (see Figure 5) allows clinicians to understand high-risk predictions. High NPV (93.33%) enables safe reduction of unnecessary monitoring in low-risk patients.

5.3 Unexpected findings and methodological considerations

The exceptionally high AUROC (0.9726) warrants critical discussion. Although moderate differences between training (0.9962) and test (0.9726) suggest adequate generalization (a gap of 2.36 percentage points), the absolute magnitude remains unusually high. Possible explanations include (1) artificial rebalancing that simplifies discrimination, (2) BioBERT embeddings capturing associations from biomedical literature and introducing prior-knowledge bias, and (3) differential verification bias derived from retrospective ADR labeling. Prospective validation in independent cohorts constitutes the standard prior to clinical implementation. Results should therefore be interpreted as optimistic estimates that require rigorous external validation.

6 LIMITATIONS AND FUTURES SOPES

This work has important limitations. The retrospective design implies underreporting of moderate ADRs (50–70%), potentially underestimating true risks. Reliance on international data sources limits immediate generalizability. BioBERT's 2019 training cutoff [22] restricts representation of newly approved drugs. We did not incorporate pharmacogenomic variables or evaluate performance across demographic subgroups. Evaluation focused on statistical metrics without measuring clinical impact. Future work should prioritize: (1) multicenter prospective validation; (2) implementation studies; (3) randomized trials on clinical outcomes; (4) algorithmic fairness analyses; (5) pharmacogenomic data; and (6) dynamic risk models.

7 CONCLUSION

This study shows that the hybrid BioBERT-XGBoost model achieves outstanding performance in predicting serious adverse drug reactions, with notable metrics (AUROC = 0.9726, AUPRC = 0.9596, and Brier = 0.0604). The importance of semantic features confirms that biomedical language representations provide predictive information beyond traditional structured variables. The obtained risk stratification (70.0% incidence in the high-risk group vs. 12.8% in the low-risk group) demonstrates the model's potential to support personalized preventive interventions. Likewise, the balance between sensitivity (91.56%) and specificity (95.24%) suggests that it is feasible to integrate the system into clinical environments without triggering excessive alert fatigue. The main limitations include the retrospective study

design, lack of prospective external validation, and absence of real-world clinical impact evaluation. Therefore, the exceptional performance reported here requires confirmation through multicenter prospective studies to validate its generalizability. The advancement toward intelligent and responsible pharmacovigilance demands interdisciplinary collaboration among data science, medicine, and pharmacology experts. The true impact of this line of research will be measured by its contribution to patient safety: reductions in preventable ADRs, fewer hospitalizations, and optimization of pharmacological therapies. This work represents a significant step toward clinically interpretable and ethically implementable predictive systems in medical practice.

8 ACKNOWLEDGMENT

The authors are grateful to the Dirección de Investigación de la Universidad Peruana de Ciencias Aplicadas (UPC) for the support provided for this research work through the UPC-EXPOST-2025-2 incentive.

9 REFERENCES

[1] Ministerio de Salud del Perú, "Resultados de la evaluación de indicadores de farmacovigilancia y tecnovigilancia 2024," *DIGEMID*, Lima, Peru, 2024. [Online]. Available: https://www.digemid.minsa.gob.pe/Archivos/PortalWeb/Informativo/Farmacovigilancia/Indicadores/resultados_tecnovigilancia_2024.pdf [Accessed: Sep. 21, 2025].

[2] B. Poojar *et al.*, "A prospective study of the medication regimen complexity index and hospitalization due to adverse drug reactions among people living with HIV," *Medicina*, vol. 60, no. 10, p. 1705, 2024. https://doi.org/10.3390/medicina60101705

[3] Q. Hu, Y. Chen, D. Zou, Z. He, and T. Xu, "Predicting adverse drug event using machine learning based on electronic health records: A systematic review and meta-analysis," *Front. Pharmacol.*, vol. 15, p. 1497397, 2024. https://doi.org/10.3389/fphar.2024.1497397

[4] V. Dsouza, L. Leyens, J. R. Kurian, A. Brand, and H. Brand, "Artificial Intelligence (AI) in pharmacovigilance: A systematic review on predicting adverse drug reactions (ADR) in hospitalized patients," *Res. Social Adm. Pharm.*, vol. 21, no. 6, pp. 453–462, 2025. https://doi.org/10.1016/j.sapharm.2025.02.008

[5] O. Chandraumakantham, S. Srinivasan, and V. Pathak, "Detecting side effects of adverse drug reactions through drug-drug interactions using graph neural networks and self-supervised learning," *IEEE Access*, vol. 12, pp. 93823–93840, 2024. https://doi.org/10.1109/ACCESS.2024.3407877

[6] L. Zhang and T. Liu, "PreAlgPro: Prediction of allergenic proteins with pre-trained protein language model and efficient neutral network," *International Journal of Biological Macromolecules*, vol. 280, no. Pt 3, p. 135762, 2024. https://doi.org/10.1016/j.ijbiomac.2024.135762

[7] L. T. Silva, A. C. F. Modesto, R. A. de Oliveira, R. G. Amaral, and F. M. Lopes, "Mortality and years of life lost related to adverse drug events in Brazil," *Revista de Saúde Pública*, vol. 58, no. 1, p. 20, 2024. https://doi.org/10.11606/s1518-8787.2024058005458

[8] H. K. Kim, K. S. Jang, and D. W. Kim, "Comparative analysis of adverse drug reactions associated with new antiseizure medications from the Korea adverse event reporting system database," *Epilepsy and Behavior*, vol. 154, p. 109784, 2024. https://doi.org/10.1016/j.yebeh.2024.109784

[9] A. Z. Al Meslamani, "Adverse drug event reporting among women: Uncovering disparities in underserved communities," *Expert Opinion on Drug Safety*, vol. 23, no. 5, pp. 543–545, 2024. https://doi.org/10.1080/14740338.2024.2337745

[10] Z. Mitkova, A. Dimova, G. Petrova, and M. Dimitrova, "Adverse drug reactions of cardiovascular classes of medicines—Data for Bulgarian population," *Biomedicines*, vol. 12, no. 10, p. 2163, 2024. https://doi.org/10.3390/biomedicines12102163

[11] R. Daunt, S. McGettigan, L. Kelly, D. Curtin, and D. O'Mahony, "Detection of potential prescribing cascades in multimorbid older patients hospitalised with acute illness—An observational prospective prevalence study," *Drugs and Aging*, vol. 42, pp. 535–546, 2025. https://doi.org/10.1007/s40266-025-01201-9

[12] A. Z. Al Meslamani, "Underreporting of adverse drug events: A look into the extent, causes, and potential solutions," *Expert Opinion on Drug Safety*, vol. 22, no. 5, pp. 351–354, 2023. https://doi.org/10.1080/14740338.2023.2224558

[13] Y. Song, Z. Wang, N. Wang, X. Xie, T. Zhu, and Y. Wang, "A real-world pharmacovigilance study of omalizumab using disproportionality analysis in the FDA adverse drug events reporting system database," *Scientific Reports*, vol. 15, no. 1, p. 8045, 2025. https://doi.org/10.1038/s41598-025-91463-5

[14] P. Das and D. H. Mazumder, "Inceptionv3-LSTM-COV: A multi-label framework for identifying adverse reactions to COVID medicine from chemical conformers based on Inceptionv3 and long short-term memory," *ETRI Journal*, vol. 46, no. 6, pp. 1030–1046, 2024. https://doi.org/10.4218/etrij.2023-0288

[15] X. Xu *et al.*, "In-silico approaches to assessing multiple high-level drug–drug and drug–disease adverse drug effects," *Expert Opinion on Drug Metabolism and Toxicology*, vol. 20, no. 7, pp. 579–592, 2024. https://doi.org/10.1080/17425255.2023.2299337

[16] A. Azizi, M. Azizi, and M. Nasri, "Artificial Intelligence techniques in medical imaging: Systematic review," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 19, no. 17, pp. 66–97, 2023. https://doi.org/10.3991/ijoe.v19i17.42431

[17] Y. Yang and N. Xia, "Enhancing students' metacognition via AI-driven educational support systems," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 18, no. 24, pp. 133–148, 2023. https://doi.org/10.3991/ijet.v18i24.45647

[18] D. Mauricio, C. M. Flores-Cortegana, A. J. Shuan-Arias, P. Castañeda, L. Rojas-Mezarina, and J. L. Castillo-Sequera, "Sedentary: A healthy lifestyle app for home office workers," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 19, no. 8, pp. 188–209, 2025. https://doi.org/10.3991/ijim.v19i08.49147

[19] U.S. Food and Drug Administration, "FDA Adverse Event Reporting System (FAERS): Latest quarterly data files," openFDA, 2024. [Online]. Available: https://open.fda.gov/data/faers/ [Accessed: Nov. 3, 2025].

[20] D. S. Wishart *et al.*, "DrugBank 5.0: A major update to the DrugBank database for 2018," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1074–D1082, 2018. https://doi.org/10.1093/nar/gkx1037

[21] Congreso de la República del Perú, "Ley N° 26842, Ley General de Salud," Diario Oficial El Peruano, Lima, Perú, 1997. [Online]. Available: https://www.gob.pe/institucion/congreso-de-la-republica/normas-legales/26842 [Accessed: Nov. 3, 2025].

[22] J. Lee *et al.*, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020. https://doi.org/10.1093/bioinformatics/btz682

[23] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785–794. https://doi.org/10.1145/2939672.2939785

[24] "XGBoost documentation: Scale_Pos_Weight," XGBoost. [Online]. Available: https://xgboost.readthedocs.io/ [Accessed: Nov. 3, 2025].

[25] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in *Proc. 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, AB, Canada, 2002, pp. 694–699. https://doi.org/10.1145/775047.775151

[26] E. W. Steyerberg, *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*, 2nd ed. Cham, Switzerland: Springer, 2019. [Online]. Available: https://link.springer.com/book/10.1007/978-3-030-16399-0 [Accessed: Nov. 3, 2025].

[27] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, 2020. https://doi.org/10.1186/s12864-019-6413-7

[28] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, 2017. https://doi.org/10.48550/arXiv.1705.07874

10 AUTHORS

Alexandra Ramirez is an undergraduate student in Information Systems Engineering at Universidad Peruana de Ciencias Aplicadas (UPC). Her research interests include machine learning, biomedical natural language processing, and pharmacovigilance systems. This work represents her contribution to hybrid model development and clinical decision support systems as part of her undergraduate research project (E-mail: u20211g190@upc.edu.pe).

Raul Pingo is an undergraduate student in Information Systems Engineering at Universidad Peruana de Ciencias Aplicadas (UPC). His research interests encompass explainable artificial intelligence, predictive modeling in healthcare, and biomedical data analysis. This paper represents his contribution to the development of interpretable models for adverse drug reaction prediction (E-mail: u202120632@upc.edu.pe).

Sandra Wong-Durand has a master's degree in Artificial Intelligence; a master's degree in Business Administration from ESAN with a mention in Advanced Project Management; a Systems Engineer degree from UNIFE with specialization studies in Innovation and Leadership at the Escuela Superior de Administración y Dirección de Empresas (ESADE)—Spain, Process Improvement Management with CMMI at the Software Engineering Institute; Software Quality at UNIFE; Strategic Project Management at PM Certifica; SOA Architectures at IBM and Oracle (E-mail: pcsiswon@upc.edu.pe).

Pedro Castañeda obtained his Ph.D. from Universidad Nacional Mayor de San Marcos (UNMSM), Lima, Peru. He is a full-time professor at the Faculty of Information Systems Engineering at Universidad Peruana de Ciencias Aplicadas (UPC), Lima, Peru. He is a RENACYT researcher certified by CONCYTEC. His research interests include machine learning, big data, health technologies, and software engineering. He has extensive experience in project management and serves as thesis advisor for undergraduate and graduate students. He has the following certifications: Project Management Professional (PMP), Scrum Certified Developer (CSD), IBM Certified Professional in Rational Unified Process, and ORACLE Certifications. Areas of Interest: Artificial Intelligence, Software Productivity, Business Intelligence, Data Analytics, Machine Learning, Software Engineering (E-mail: pedro.castaneda@untrm.edu.pe).

Alejandra Oñate-Andino holds a degree in computer systems engineering from Escuela Superior Politécnica de Chimborazo (Ecuador), a master's in network

Interconnectivity from Escuela Superior Politécnica de Chimborazo (Ecuador), and a PhD in systems engineering and computer science from Universidad Mayor de San Marcos (Peru). Currently she is the coordinator of the software career at the Escuela Superior Politécnica de Chimborazo (Ecuador). In addition, she is a research professor, with more than 15 years of experience leading teaching, research, and management processes. She has directed and participated in several research and community outreach projects. Author of several scientific articles in the area of information technology governance, business intelligence, information technology management, and others (E-mail: monate@espoch.edu.ec).