A SUPERVISED MACHINE LEARNING AND NLP FRAMEWORK FOR PREDICTING FDA PREGNANCY DRUG SAFETY CATEGORIES USING STRUCTURED AND UNSTRUCTURED DRUG DATA

GROUP-11: AUTHORS: SAHITHI KALAPALA, MOHAN NAIK PALITHYA, VENKATA SAI PALLAVI PALLAPOLU, HARIKA PAMULAPATI

SAINT LOUIS UNIVERSITY — EST. 1818

SAINT LOUIS UNIVERSITY — DEPARTMENT OF HEALTH AND CLINICAL OUTCOMES RESEARCH

Background

- Up to 90% of pregnant women take at least one medication during pregnancy, yet safety data is often lacking due to the exclusion of pregnant women from clinical trials.[1,2,3]
- The FDA's 1979 A-X category system was widely criticized as oversimplified and has been replaced by the PLLR narrative labeling, which still lacks structured, trimester-specific guidance.[4,5,6]
- Over 90% of approved drugs since 1980 lack sufficient human pregnancy safety data, and post-marketing surveillance suffers from underreporting and fragmentation.[2,7]
- Prior ML models have relied mostly on chemical structure data, overlooking real-world context like side effects and patient-reported outcomes, limiting clinical relevance.[1,3,8]
- Explainability and scalability have hindered adoption of earlier AI models, while recent NLP approaches focus more on information retrieval than prediction.[3,9]
- This study proposes a supervised ML + NLP framework using structured drug metadata and unstructured side-effect narratives from Drugs.com to fill this critical gap in risk classification.[4,10]

Methods

- **Data Source:** Utilized the publicly available Drugs, Side Effects, and Medical Condition data(n = 2,931; final n = 2,591 after cleaning), containing drug names, structured attributes (e.g: Rx/OTC, CSA Schedule), and free-text side-effect narratives.
- **Preprocessing:** Applied NLP techniques (lowercasing, stopword removal, lemmatization) to side-effect texts using SciSpaCy.[11,12] Structured fields were cleaned, missing values imputed, and inconsistent entries standardized. Transformed textual data into TF-IDF vectors and structured categorical variables via one-hot encoding.[13,14]
- **Statistical Analysis:** Chi-square and ANOVA tests were performed to identify associations between predictors (e.g: Drug Class, Medical Condition) and pregnancy category outcome.
- **Target & Predictors:** The target variable was the FDA Pregnancy Category (A,B,C,D,N,X). Predictors included structured fields, Drug Class, Medical Condition, Rx/OTC, CSA Schedule, Alcohol Interaction, and unstructured Side Effects text, which were processed for use in ML models.[15]
- **Model Development:** Trained and tuned multiple classifiers (Logistic Regression,[16] SVM,[17] KNN,[18] Decision Trees,[19] Random Forest,[20] Gradient Boosting,[21] XGBoost,[22] ANN[23]) using a unified pipeline with stratified 80/20 train-test split.
- **SMOTE Application:** To address class imbalance, SMOTE was selectively applied only to the training data of the best-performing model.[24,25]
- **Interpretability:** SHAP values were used to evaluate the influence of structured and text-derived features on model predictions. Topic modeling using NMF was applied to uncover latent themes within side-effect texts.
- **Evaluation:** Multiclass ROC-AUC curves, confusion matrices, and prediction confidence plots were used to assess and visualize model performance across six pregnancy categories.[26]

Table 1: Chi-Square Test Results for Categorical Features and Their Association with Pregnancy Risk Categories

Variable	Chi2	df	p-value	95% CI (Chi2)	Significant
Csa	157.47	25	0.0	[13.12, 40.65]	Yes
Rx_Otc	793.72	10	0.0	[3.25, 20.48]	Yes
Alcohol_Interaction	67.01	5	0.0	[0.83, 12.83]	Yes
Drug_Classes	9729.81	1315	0.0	[1216.39, 1417.39]	Yes
Medical_Condition	5046.65	230	0.0	[189.89, 273.9]	Yes

Figure 2: SHAP Plot of Top Predictors in the Final SMOTE-Enhanced ANN Model

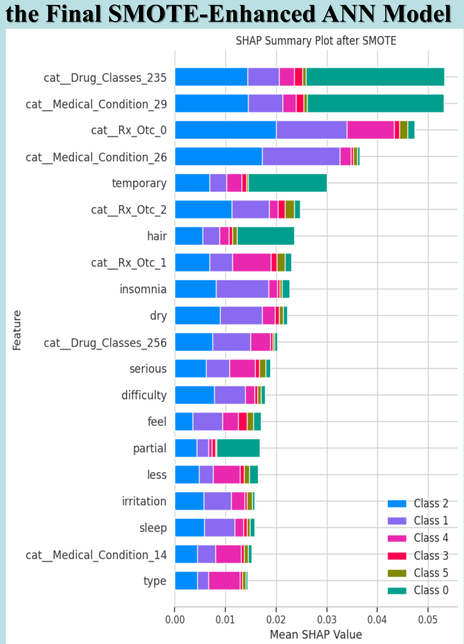


Figure 3: Confusion Matrix for SMOTE-applied Tuned ANN Model

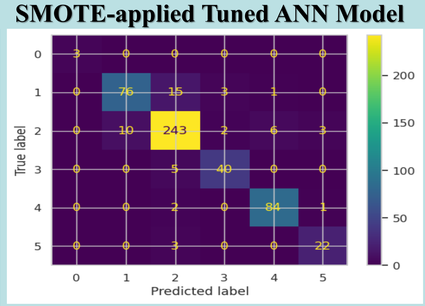


Figure 4: ROC-AUC Curve for Final SMOTE-applied Tuned ANN Model

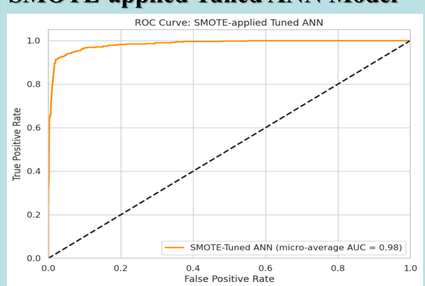


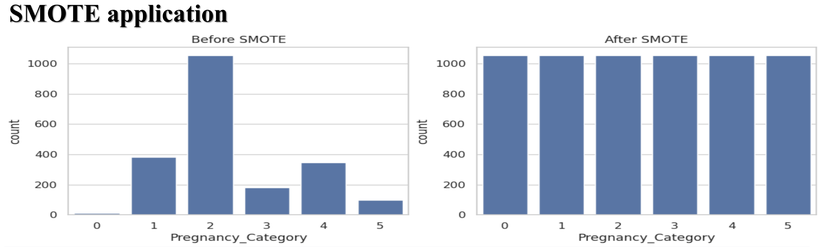
Table 2: Post-Tuning Performance Metrics of Machine Learning Models

Model	Test Accuracy	Macro Precision	Macro Recall	Macro F1-Score	AUC-ROC
Logistic Regression	0.8940	0.8951	0.9219	0.9072	0.9748
SVM	0.8843	0.8960	0.9095	0.9012	0.9792
KNN	0.8786	0.8920	0.8749	0.8819	0.9316
Decision Tree	0.8362	0.8661	0.8163	0.8372	0.8861
Random Forest	0.8766	0.9317	0.8469	0.8823	0.9863
Gradient Boost	0.8227	0.8730	0.8188	0.8404	0.9629
XGBoost	0.8670	0.9256	0.8336	0.8696	0.9811
ANN	0.8959	0.9102	0.9098	0.9053	0.9825
SMOTE applied ANN	0.9094	0.9239	0.9273	0.9242	0.9791

Results

- **Statistical Significance:** Chi-square and ANOVA tests revealed significant associations ($p < 0.001$) between pregnancy category and all structured predictors, especially Drug Class and Medical Condition ($\chi^2 = 9729.81$ and 5046.65, respectively), confirming their predictive value.
- **Topic Modeling Insights:** NMF on side-effect narratives extracted 20 coherent clinical themes (e.g: respiratory, dermatologic, neurologic), many of which were more prevalent in low to moderate-risk drug categories (B, C, N), enhancing interpretability.
- **Model Performance:** Among 8 baseline and tuned ML models, the SMOTE-applied ANN achieved the highest accuracy (90.94%), macro F1-score (0.9242), and AUC-ROC (0.9791), outperforming Logistic Regression, SVM, XGBoost, and Random Forest.
- **Confusion Matrix:** Class-specific results showed high true positive rates, with most misclassifications between adjacent categories.
- **SMOTE Impact:** Selective SMOTE application to the ANN's training set improved minority-class predictions without degrading overall performance, increasing test accuracy from 89.59% to 90.94% and stabilizing recall and precision across all six classes.
- **Interpretability:** SHAP analysis highlighted key predictors including thyroid and respiratory drugs, hypothyroidism, hay fever, and keywords like "insomnia", "hair", and "temporary," validating the hybrid model.

Figure 1: Class distribution of pregnancy safety categories before and after SMOTE application



Conclusion

- **Integrated ML Framework:** Combining structured drug attributes and unstructured side-effect narratives, the SMOTE-enhanced ANN model achieved superior performance (90.94% accuracy, F1 = 0.92, AUC = 0.98), showcasing the value of real-world, patient-centered data.
- **Interpretability & Clinical Insight:** SHAP analysis and topic modeling revealed clinically relevant predictors and side-effect clusters aligned with pregnancy risk, supporting model transparency & potential clinical relevance.
- **Future Directions:** To improve clinical applicability, future work should incorporate EHR data, trimester-specific risk factors, and the updated PLLR labeling system, moving toward personalized, evidence-based maternal safety tools.

References

1. Peng J, Fu L, Yang G, Cao D. Advanced AI-Driven Prediction of Pregnancy-Related Adverse Drug Reactions. Journal of Chemical Information and Modeling. 2024;64(24):9286-9298. doi:https://doi.org/10.1021/acs.jcim.4c01657
2. Kennedy D, Batagol R. Drug safety in pregnancy. Australian Prescriber. 2025;48(1):5-9. doi:https://doi.org/10.18773/austprescr.2025.008
Doc link that provides the remaining references used in the poster: https://docs.google.com/document/d/1T3Rxvx1su_9VznQM5qOpAFdQolCPw_FFu7p8oveDd6A/edit?usp=sharing

Application link: https://pregnancy-drug-safety-prediction-project-by-saipallavi.streamlit.app/