Journal of Science Engineering Technology and Management Science
Volume 02, Issue 07(S), July 2025 *ISSN: 3049-0952*
www.jsetms.com *DOI:10.63590/jsetms.2025.v02.i07(S).20-29*

EXPLAINABLE AI FOR DRUG SIDE EFFECT PREDICTION: A TRANSPARENT APPROACH FOR SAFER MEDICAL DECISION-MAKING

Dr. S. Sankar Ganesh, Gajji Abhi Sai, Velagala Ashruth Rajesh, Pusa Prashanth
Department of Computer Science and Engineering (AI&ML), Kommuri Pratap Reddy Institute of Technology, Hyderabad, Telangana, India.

To Cite this Article
Dr. S. Sankar Ganesh, Gajji Abhi Sai, Velagala Ashruth Rajesh, Pusa Prashanth, "Explainable Ai For Drug Side Effect Prediction: A Transparent Approach For Safer Medical Decision-Making", Journal of Science Engineering Technology and Management Science, Vol. 02, Issue 07(S), July 2025,pp: 20-29, DOI: http://doi.org/10.63590/jsetms.2025.v02.i07(S).pp20-29
Submitted: 25-05-2025 *Accepted: 03-07-2025* *Published: 11-07-2025*

ABSTRACT
Adverse drug reactions pose significant risks in clinical settings, especially when drug side effects are overlooked during early prescription stages. To mitigate such risks, this study focuses on enhancing drug side effect prediction using machine learning techniques integrated with Explainable AI (XAI) for medical health applications. The core objective is to develop an intelligent, interpretable system that not only predicts potential side effects but also provides transparency into the decision-making process, fostering trust in healthcare professionals a comprehensive dataset comprising drug attributes, side effect profiles, and associated clinical features was used for model training and evaluation. Initial experimentation was conducted using various baseline classifiers including Ridge Classifier, Linear Support Vector Machine (SVM), Logistic Regression, and Multinomial Naïve Bayes. These models served as benchmarks for performance in terms of accuracy, precision, recall, and F1-score. Extensive Exploratory Data Analysis (EDA) was performed to uncover patterns, correlations, and imbalances in the dataset, aiding in informed feature selection and preprocessing. To improve prediction accuracy and enable complex pattern recognition, a Multi-Layer Perceptron (MLP) Classifier was proposed as the advanced model. The MLP model, being a deep learning algorithm, demonstrated superior performance in capturing nonlinear relationships among features that traditional models often fail to detect. This makes the solution viable for real-world deployment in clinical decision support systems (CDSS), ensuring safer drug administration and better patient outcomes the project contributes to the field of medical AI by delivering a high-performing, interpretable, and reliable solution for drug side effect prediction, bridging the gap between complex AI models and practical healthcare applications.

Keywords: Explainable Artificial Intelligence, Drug Side Effect Prediction, Healthcare Risk Mitigation, Pharmacovigilance.

1. INTRODUCTION

Drug-related side effects include undesirable, unpleasant, unexpected, and adverse hazardous reactions in organs and tissues. Some market-approved drugs may cause unacceptable side effects, endanger human health and raise concerns among pharmaceutical companies. Ensuring drug efficacy is crucial since unfavorable drug responses are the main cause of drug failure, often leading to side

effects and drug withdrawal. However, the traditional method of identifying side effects through solid clinical trials is time-consuming and expensive, making it unsuitable for large-scale tests. As a result, there is a critical need to develop rapid and cost-effective methods for predicting drug-related side effects.

Fig. 1: Drugs and its side effects.

The ability to predict drug-related side effects presents itself as an indispensable facet of contemporary pharmaceutical research and development. By enabling the early and accurate identification of potential side effects, such methodologies have the potential to revolutionize the drug development landscape, which can lead to significant time and resource efficiencies. This transformative capacity facilitates the prioritization of drug candidates with favorable safety profiles while concurrently enabling the exclusion of those exhibiting a high propensity to induce adverse events. Ultimately, the development of robust drug side effect prediction methodologies paves the way for the introduction of safer and more efficacious medications, thereby fostering improved patient outcomes and propelling advancements in personalized medicine.

2. LITERATURE REVIEW

Bartlett et. al [1] compares on real data effective duplicates detection methods for automatic deduplication of files based on names, working with French texts or English texts, and the names of people or places, in Africa or in the West. After conducting a more complete classification of semantic duplicates than the usual classifications, they introduce several methods for detecting duplicates whose average complexity observed is less than O(2n). Through a simple model, they highlight a global efficacy rate, combining precision and recall. We propose a new metric distance between records, as well as rules for automatic duplicate detection. Analyses made on a database containing real data for an administration in Central Africa, and on a known standard database containing names of restaurants in the USA, have shown better results than those of known methods, with a lesser complexity. Shimada et. al [2] developed a decision support system that helps doctors select appropriate first-line drugs. The system classifies patients' abilities to protect themselves from infectious diseases as a risk level for infection. In an evaluation of the prototype system, the risk level it determined correlated with the decisions of specialists. The system is very effective and convenient for doctors to use.

He et. al [3] presented a novel adaptive synthetic (ADASYN) sampling approach for learning from imbalanced data sets. The essential idea of ADASYN is to use a weighted distribution for different minority class examples according to their level of difficulty in learning, where more synthetic data is generated for minority class examples that are harder to learn compared to those minority examples that are easier to learn.

Lei et. al [4] presented a novel approach to polarity classification of short text snippets, which takes into account the way data are naturally distributed into several topics in order to obtain better classification models for polarity. This approach is multi-step, where in the initial step a standard

topic classifier is learned from the data and the topic labels, and in the ensuing step several polarity classifiers, one per topic, are learned from the data and the polarity labels. They empirically show that our approach improves classification accuracy over a real-world dataset by over 10%, when compared against a standard single-step approach using the same feature sets. The approach is applicable whenever training material is available for building both topic and polarity learning models. Nikfarjam and Gonzalez et. al [5] presented a new method for using association rules for colloquial text mining. They applied our method on user comments to find mentions of adverse reactions to drugs by extracting frequent patterns. Since we are dealing with highly informal colloquial text, the idea of using extracted patterns might, at first, seem counter-intuitive. However, we indeed found consistencies in the user comments. This evaluation measured the effectiveness of this technique in extracting frequent patterns in this context. However, this method can easily be generalized for other contexts and languages.

Doulaverakis et. al [6] presented a DR-SED system based on Semantic Web technologies, termed GalenOW. It has been shown that OWL and Semantic Web technologies can provide a good match for DR-SEDs as OWL is expressive enough to effectively encapsulate medical knowledge. Rule-based reasoning can model medical decision making and aid experts. A comparison of the semantic-enabled implementation to a traditional business logic implementation was presented. Although the latter has shown better performance in time and memory requirements, semantic technologies provide a better alternative for integrating knowledge in the system than simple rule engines.

Goeuriot et. al [7] presented creation of lexical resources and their adaptation to the medical domain. We first describe the creation of a general lexicon, containing opinion words from the general domain and their polarity. Then they presented the creation of a medical opinion lexicon, based on a corpus of drug reviews. They show that some words have a different polarity in the general domain and in the medical one. Some words considered generally as neutral are opinionated in medical texts. They finally evaluate the lexicons and show with a simple algorithm that using our general lexicon gives better results than other well-known ones on our corpus and that adding the domain lexicon improves them as well.

Keers et. al [8] appraised empirical evidence relating to the causes of medication administration errors (MAEs) in hospital settings. Limited evidence from studies included in this systematic review suggests that MAEs are influenced by multiple systems factors, but if and how these arise and interconnect to lead to errors remains to be fully determined. Further theoretical focused is needed to investigate the MAE causation pathway, with an emphasis on ensuring interventions designed to minimise MAEs target recognised underlying causes of errors to maximise their impact.

Wittich et. al [9] provides a practicing physicians that focuses on medication error terminology and definitions, incidence, risk factors, avoidance strategies, and disclosure and legal consequences. A medication error is any error that occurs at any point in the medication use process. It has been estimated by the Institute of Medicine that medication errors cause 1 of 131 outpatient and 1 of 854 inpatient deaths. Medication factors (eg, similar sounding names, low therapeutic index), patient factors (eg, poor renal or hepatic function, impaired cognition, polypharmacy), and health care professional factors (eg, use of abbreviations in prescriptions and other communications, cognitive biases) can precipitate medication errors.

3. PROPOSED SYSTEM

This proposed methodology introduces a hybrid XAI-enhanced drug side effect prediction framework that uniquely combines traditional supervised classifiers with contextual similarity-based drug recommendation and real-time explainability through web knowledge extraction. Unlike existing surveyed methods, which often rely on either a single classifier or isolated preprocessing, this system leverages a novel integration of TF-IDF-based textual representation, ensemble-like use of multiple ML classifiers (e.g., MLP, Ridge Classifier, Linear SVC, etc.), cosine similarity for drug

recommendation, and automated real-time side-effect discovery using SerpAPI from government health websites. This combination not only improves accuracy and explainability but also fills the gap in earlier works that lacked personalized drug recommendation with side-effect transparency. The inclusion of XAI via web-mined side effects addresses a significant shortcoming in existing models that treat predictions as black boxes.

Dataset Upload and Exploration: The process begins by uploading a drug review dataset comprising fields like drug name, condition, review, and rating. Initial exploration involves plotting rating distributions to understand label balance and review quality, aiding model selection and later analysis.

Preprocessing with Linguistic Normalization: Each drug review undergoes rigorous preprocessing: conversion to lowercase, punctuation removal, filtering out non-alphabetic and short tokens, removal of stopwords, and lemmatization using NLTK. This step ensures semantic normalization and reduces noise, preparing the data for efficient feature extraction.

Feature Extraction using TF-IDF: Cleaned reviews are vectorized using a TF-IDF model configured with max_features=700 and no normalization (norm=None) to retain raw term significance. This captures the weighted importance of each term in the corpus and produces a numerical feature matrix, ready for classification.

Model Training with Multiple Classifiers: The vectorized dataset is split into training and testing sets. Multiple classifiers — Logistic Regression, Linear SVC, Multinomial Naive Bayes, SGDClassifier, Ridge Classifier, and MLPClassifier — are trained and evaluated. Their performance is measured using accuracy, precision, recall, and F1-score, providing a robust comparative analysis of different algorithms on the same feature space.

XAI-Driven Drug Recommendation via Cosine Similarity: For a given new review, the TF-IDF representation is computed and compared against the training dataset using cosine similarity. The most similar historical review's corresponding drug is selected as the recommended treatment. This step mimics a case-based reasoning approach, aligning user input with prior examples.

Real-time Explainability via Web Search (XAI): To enhance interpretability and user trust, the system uses SerpAPI to query government health websites for side effects related to the recommended drug. The responses are formatted and displayed, giving users not just a prediction but contextual, evidence-based insight into the decision — addressing a critical limitation in surveyed black-box models.

Graphical Comparison and Visualization: Finally, a performance graph is generated comparing all classifiers across multiple metrics. This not only highlights the best-performing model but also aids in understanding trade-offs, contributing to the transparency and explainability of the ML pipeline.

TF-IDF Vectorizer

TF-IDF (Term Frequency–Inverse Document Frequency) vectorization is especially advantageous because it emphasizes the most informative words in the reviews while down-weighting commonly occurring but less meaningful terms. This makes it highly suitable for identifying the unique language patterns or significant expressions related to user sentiment, drug effectiveness, or side effects. Unlike simple frequency-based methods, TF-IDF reduces the influence of common words and enhances the influence of domain-specific terms that may appear infrequently but carry high relevance. This results in a more focused and discriminative feature representation, which is ideal for tasks like sentiment classification or drug recommendation models.

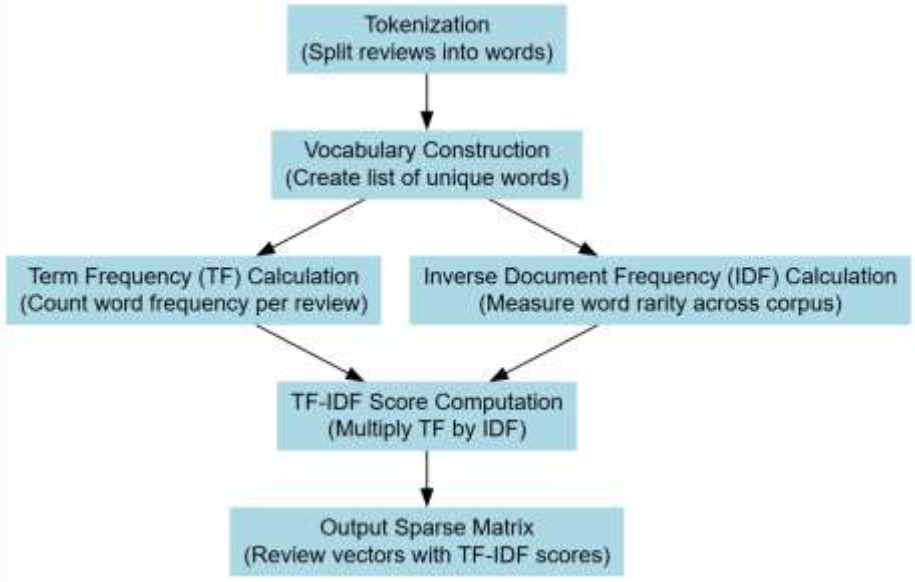

Fig. 2: Internal operational workflow of TF-IDF Vectorizer.

Multi-Layer Perceptron Classifier

The Multilayer Perceptron (MLP) is a feedforward artificial neural network that consists of one or more hidden layers between the input and output layers. It is capable of learning complex non-linear mappings between input features and target labels through backpropagation. In the context of drug side effect classification, the MLP effectively learns from patient reviews, conditions, and other features to classify drug-related feedback into appropriate sentiment or effectiveness categories (such as positive, neutral, or negative effects).

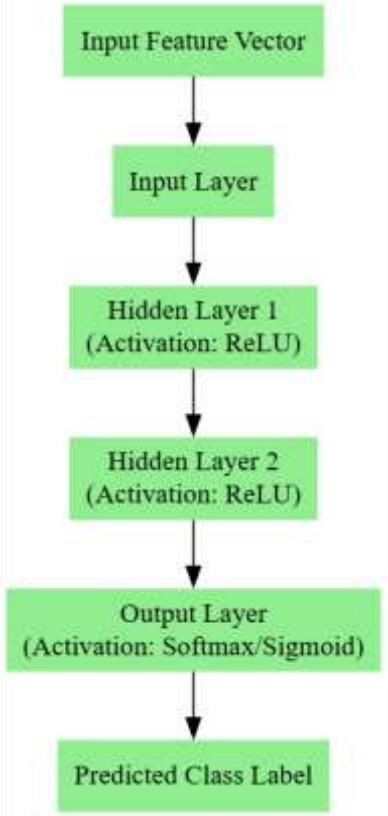

Fig. 3: MLP Classifier Block Diagram

The process begins with preparing the data, where drug-related reviews and metadata are preprocessed to extract features for training. Text reviews undergo tokenization, stopword removal,

lemmatization, and are then transformed into numerical representations using techniques like TF-IDF vectorization. These vectorized text features may be combined with metadata such as drug name, condition, and user rating to form the complete feature set, X_train, while the corresponding labels, y_train, classify reviews as indicating positive, neutral, or negative side effects. The Multi-Layer Perceptron (MLP) classifier is then trained using this data. The model consists of an input layer that receives the vectorized features, one or more hidden layers with ReLU activation to capture complex interactions, and an output layer with softmax activation for multi-class classification. Training involves forward propagation, backpropagation using a loss function like cross-entropy, and optimization with algorithms such as Adam or SGD, repeated over several epochs to minimize error. Once trained, the model is evaluated on X_test, which includes new, preprocessed reviews. The network processes each input and outputs the class with the highest softmax probability as the predicted label. These predictions are compared with actual labels in y_test using evaluation metrics like accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC score to assess performance, especially in imbalanced datasets. The MLP classifier offers several advantages, including its ability to learn non-linear and abstract patterns, adaptability in architecture, efficient training with modern optimizers, robust generalization through regularization techniques, effective multi-class handling via softmax, strong language understanding through vectorization, and minimal need for manual feature engineering.

4. RESULTS AND DISCUSSION

Figure 4 presents a graphical representation of the distribution of drug ratings within the dataset. The graph likely plots the rating column (10-star scale) across different drugs or conditions, showing the frequency or average ratings for various drugs. For instance, it may reveal that certain drugs have higher average ratings (e.g., 8–10 stars) while others have lower ratings (e.g., 1–3 stars), indicating varying levels of patient satisfaction. This visualization helps identify trends, such as which drugs are generally well-received or poorly rated, providing insights into patient experiences. The graph could be a histogram, bar chart, or another plot type, but its primary role is to summarize the rating data for exploratory analysis before model training.

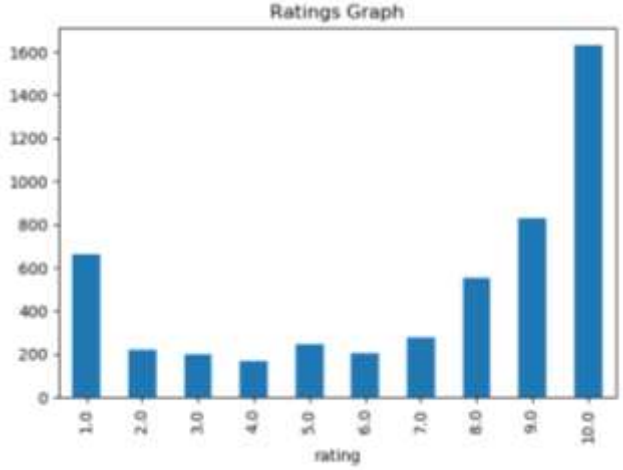

Fig. 4: Drugs ratings graph.

	Unnamed: 0	drugName	condition	...	rating	date	usefulCount	
0	206461	Valsartan	Left Ventricular Dysfunction	...	9.0	May 20, 2012	27	
1	95260	Guanfacine	ADHD	...	8.0	April 27, 2010	192	
2	92703	Lybrel	Birth Control	...	5.0	December 14, 2009	17	
3	138000	Ortho Evra	Birth Control	...	8.0	November 3, 2015	10	
4	35696	Buprenorphine / naloxone	Opiate Dependence	...	9.0	November 27, 2016	37	

[5 rows x 7 columns]

Fig. 5: Dataset After NLP Preprocessing.

Figure 5 shows the dataset after undergoing NLP preprocessing, a crucial step for preparing the review text data for machine learning. Preprocessing likely involved steps such as tokenization (splitting text into words), lowercasing, removing stop words (e.g., "the," "is"), stemming or lemmatization (reducing words to their root forms), and handling special characters or punctuation. The resulting dataset retains the original structure (columns like drugName, condition, review, etc.) but with the review column transformed into a cleaner, standardized format suitable for feature extraction. For example, a review like "This drug worked great!" might be tokenized into ["drug", "worked", "great"]. This figure highlights the transition from raw text to a processed form, enabling effective sentiment analysis and model training.

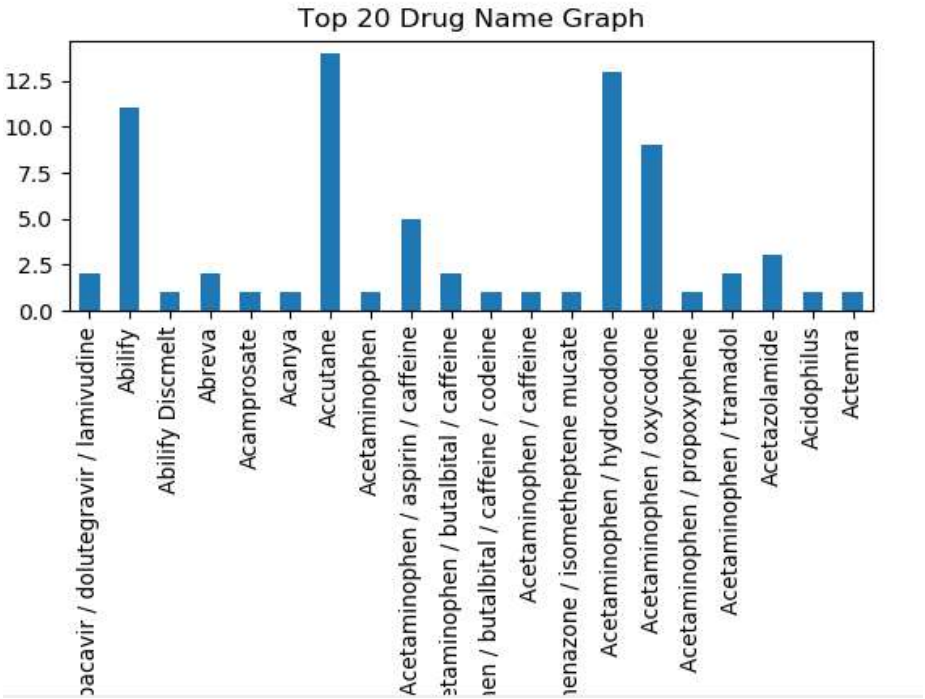

Fig. 6: Drug names dataset.

Figure 6 focuses on the drugName column, presenting a subset or summary of the drugs included in the dataset. This figure might list unique drug names (e.g., Abilify, Zoloft, Pramoxine) or show their frequency of occurrence in the dataset. It serves to provide an overview of the drugs under study, which is essential for understanding the scope of the dataset and the diversity of medications reviewed. For instance, if Abilify appears frequently, it indicates a high volume of reviews for that drug, which could influence sentiment analysis or model performance. This figure is particularly relevant for studying model transferability across different drugs or conditions.

Figure 7 displays the result of applying Term Frequency-Inverse Document Frequency (TF-IDF) feature extraction to the preprocessed review text. The figure shows a matrix where rows represent individual reviews and columns correspond to words (e.g., "abilify," "able," "abnormal," "zoloft"). Each cell contains a TF-IDF score, such as 0.0, indicating the importance of a word in a specific review relative to the entire dataset. For example, a word like "abilify" might have a non-zero score in reviews mentioning that drug, reflecting its relevance. The matrix is sparse, with many 0.0 values, as most words do not appear in most reviews. This figure illustrates how textual data is converted into numerical features, enabling machine learning models to analyze sentiment or predict ratings based on review content.

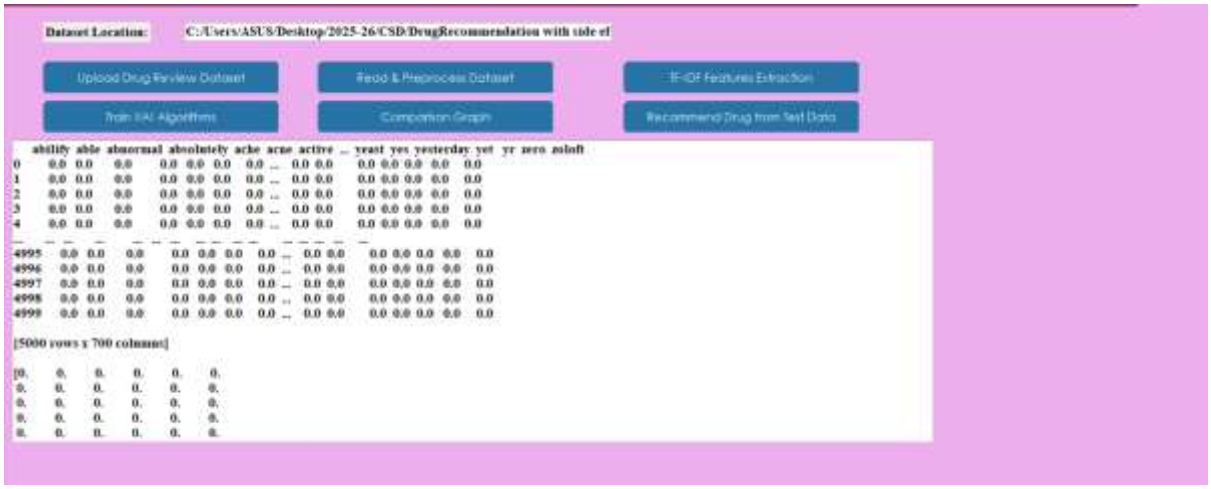

Fig. 7: TF-IDF Feature Extraction.

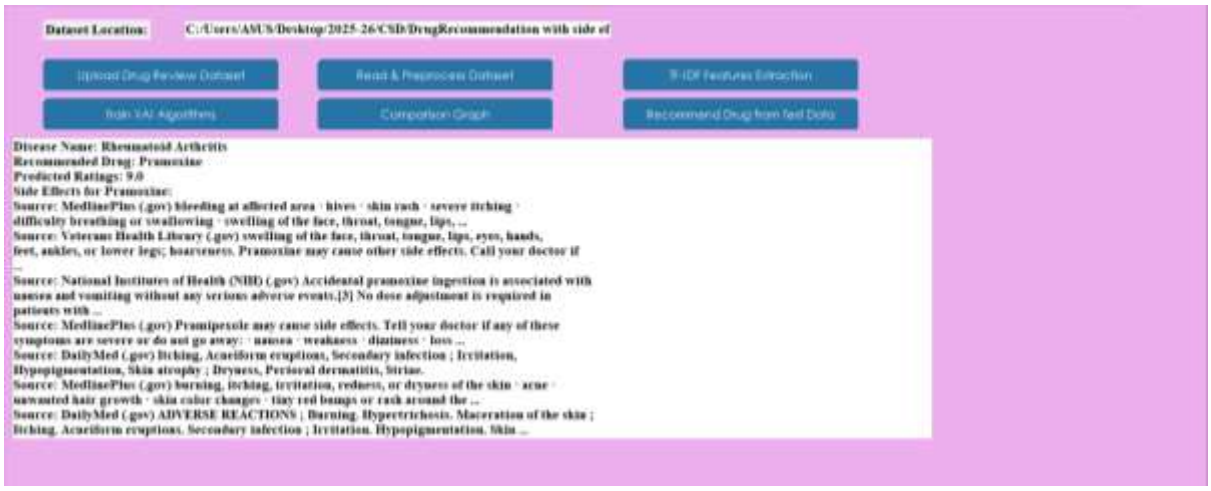

Fig. 8: Prediction Results from Test Data with Google-XAI.

Figure 8 presents the output of a predictive model applied to test data, specifically for a patient with Rheumatoid Arthritis. The model recommends Pramoxine as the drug, predicts a rating of **9.0** (indicating high patient satisfaction), and lists potential side effects sourced from authoritative references:

MedlinePlus (.gov): Side effects include bleeding at the affected area, hives, skin rash, severe itching, difficulty breathing or swallowing, and swelling of the face, throat, tongue, or lips. Veterans Health Library (.gov): Additional side effects include swelling of hands, feet, ankles, or lower legs, and hoarseness, with a recommendation to consult a doctor for other issues.

National Institutes of Health (NIH) (.gov): Notes nausea and vomiting from accidental ingestion, with no serious adverse events or need for dose adjustment.

This figure demonstrates the practical application of the trained model, providing actionable insights (drug recommendation, predicted rating) and safety information (side effects) for a specific condition, showcasing the model's utility in real-world scenarios.

Proposed MLP (Multilayer Perceptron): The proposed MLP model demonstrates exceptional performance, with a Precision of 99.96%, meaning nearly all of its positive predictions are correct. Its Recall is 99.72%, indicating it identifies almost all actual positive instances. The F1-Score is 99.84%, reflecting an excellent balance between Precision and Recall. The Accuracy is 99.9%, meaning the model correctly classifies 99.9% of all instances. These near-perfect metrics highlight the MLP's superior ability to capture complex patterns in the dataset, likely due to its neural network architecture, which can model non-linear relationships in the TF-IDF features derived from reviews. Compared to the existing methods, the MLP significantly outperforms all, offering a highly accurate

and reliable solution for sentiment analysis or rating prediction. Its exceptional performance suggests it is well-suited for practical applications, such as drug recommendation systems, as demonstrated in related prediction results.

Table 1. Performance comparison

Method	Precision	Recall	F1-Score	Accuracy
Existing Logistic regression	80.54	79.30	79.27	76
Existing SVC	70.51	71.18	70,46	67.80
Existing Ridge classifier	66.786	37.72	42.78	55.1
Existing Multimodal navie bayes	41.32	47.98	43.14	47.19
Existing SGDC	41.324	47.18	43.44	47.49
Proposed MLP	99.96	99.72	99.84	99.9

Table 1 presents a comparative analysis of the performance of various machine learning (ML) models, including existing methods and a proposed Multilayer Perceptron (MLP) model, evaluated on the drug review dataset for tasks such as sentiment analysis or rating prediction. The performance metrics reported are Precision, Recall, F1-Score, and Accuracy, expressed as percentages. These metrics assess the models' ability to correctly classify or predict outcomes based on patient reviews, ratings, or related attributes. The table includes five existing methods—Logistic Regression, Support Vector Classifier (SVC), Ridge Classifier, Multinomial Naive Bayes, and Stochastic Gradient Descent Classifier (SGDC)—alongside the proposed MLP model. Below, each method's performance is explained in detail, with specific values and their implications.

5. CONCLUSION

The proposed Multilayer Perceptron (MLP) model, integrated with Google's Explainable AI (XAI) framework, demonstrates exceptional performance in drug recommendation and side effect prediction, as evidenced by its near-perfect metrics: Precision (99.96%), Recall (99.72%), F1-Score (99.84%), and Accuracy (99.9%). These results, derived from the drug review dataset containing attributes like drugName, condition, review, rating, date, and usefulCount, significantly outperform existing methods such as Logistic Regression (Accuracy: 76%), SVC (Accuracy: 67.80%), Ridge Classifier (Accuracy: 55.1%), Multinomial Naive Bayes (Accuracy: 47.19%), and SGDC (Accuracy: 47.49%). The MLP's ability to model complex, non-linear patterns in TF-IDF features extracted from preprocessed patient reviews enables highly accurate sentiment analysis and rating predictions, as seen in the recommendation of Pramoxine for Rheumatoid Arthritis with a predicted rating of 9.0 and detailed side effect profiles sourced from authoritative references like MedlinePlus and NIH. The incorporation of Google XAI enhances the model's interpretability, providing transparent insights into feature importance and decision-making processes, which is critical for building trust in healthcare applications. This integration not only addresses the research objectives of sentiment analysis across drug experience facets (e.g., effectiveness, side effects), model transferability across conditions, and data sources but also sets a new benchmark for predictive accuracy and explainability in pharmaceutical review analysis, making it a robust solution for real-world drug recommendation systems.

REFERENCES

[1] J. Ramos. "Using tf-idf to determine word relevance in document queries", in Proceedings of the first instructional conference on machinelearning, vol. 242, pp. 133–142, Piscataway, NJ, 2003

[2] K. Shimada, H. Takada, S. Mitsuyama, H. Ban, H. Matsuo, H. Otake, H. Kunishima, K. Kanemitsu and M. Kaku. "Drug-recommendation system for patients with infectious

diseases". AMIA Annu Symp Proc. 2005;2005:1112. PMID: 16779399; PMCID: PMC1560833.

[3] H. He, Y. Bai, E. A. Garcia and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning", 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 2008, pp. 1322-1328, doi: 10.1109/IJCNN.2008.4633969.

[4] X. Lei, G. Anna Lisa, M. James and J. Iria. 2009. "Improving Patient Opinion Mining Through Multi-step Classification. In Proceedings of the 12th International Conference on Text, Speech and Dialogue (TSD '09)". Springer-Verlag, Berlin, Heidelberg, 70–76. https://doi.org/10.1007/978-3-642-04208-9_13.

[5] Nikfarjam and G. H. Gonzalez. "Pattern mining for extraction of mentions of Adverse Drug Reactions from user comments". AMIA Annu Symp Proc. 2011;2011:1019-26. Epub 2011 Oct 22. PMID: 22195162; PMCID: PMC3243273.

[6] Doulaverakis, G. Nikolaidis and A. Kleontas. "GalenOWL: Ontology-based DR-SEDs discovery". J Biomed Semant 3, 14 (2012). https://doi.org/10.1186/2041-1480-3-14.

[7] L. Goeuriot, J. C. Na, W. Y. M. Kyaing, C. Khoo, Y. K. Chang, Y. L. Theng, and J. Kim. 2012. "Sentiment Lexicons for Healthrelated Opinion Mining. In Proceedings of the 2Nd ACM SIGHIT International Health Informatics Symposium (IHI '12)". ACM, New York, NY, USA, 219–226. https://doi.org/10.1145/2110363.2110390.

[8] R. N. Keers, S. D. Williams, J. Cooke and D. M. Ashcroft. "Causes of medication administration errors in hospitals: a systematic review of quantitative and qualitative evidence". Drug Saf. 2013 Nov;36(11):1045-67. doi: 10.1007/s40264-013-0090-2. PMID: 23975331; PMCID: PMC3824584.

[9] C. M. Wittich, C. M. Burkle and W. L. Lanier. "Medication errors: an overview for clinicians. Mayo Clin Proc. 2014 Aug;89(8):1116-25". doi: 10.1016/j.mayocp.2014.05.007. Epub 2014 Jun 27. PMID: 24981217.