National College of Ireland

Towards Trustworthy Healthcare AI: Multi-Method Explainable AI for Transparent Adverse Drug Reaction Detection from Patient-Generated Text

MSc Research Project
MSc Data Analytics

Alex Sunny
Student ID: x23304600

School of Computing
National College of Ireland

Supervisor: Arjun Chikkankod

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Alex Sunny
Student ID:	x23304600
Programme:	MSc. Data Analytics **Year:** 2024-2025
Module:	MSc Research Practicum
Supervisor:	Arjun Chikkankod
Submission Due Date:	11/08/2025
Project Title:	Towards Trustworthy Healthcare AI: Multi-Method Explainable AI for Transparent Adverse Drug Reaction Detection from Patient-Generated Text
Word Count:	7417 **Page Count** 20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Alex Sunny
Date:	11/08/2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Towards Trustworthy Healthcare AI: Multi-Method Explainable AI for Transparent Adverse Drug Reaction Detection from Patient-Generated Text

Alex Sunny
x23304600

Abstract

Adverse Drug Reactions (ADR) are a critical challenge in pharmacovigilance. While there are automated ADR detection ML systems, their opaque decision-making process stalls clinical usage. This research addresses how these ADR detection systems can be combined with explainable AI methods to enhance transparency. Extensive and comprehensive review of the state of the art and previous works and contributions have also been discussed. A framework integrating ML and DL models with five distinct XAI techniques (LIME, SHAP, Integrated Gradient, Tree Interpreter and PDP) is implemented for ADR detection from patient provided drug reviews. The LSTM model achieved optimal performance (91% precision) for ADR detection. The XAI methods are evaluated using quantitative metrics like faithfulness, fidelity and complexity. LIME and SHAP have the best balance between accuracy and interpretability for clinical applications and the Tree Interpreter showcased the highest faithfulness. This research contributes an evaluation framework for XAI methods in healthcare, showcases the practical integration of XAI in pharmacovigilance and provides evidence-based guidance for developing transparent ADR detection systems which improve clinical decision-making and maintain model performance.

1 Introduction

The classification and detection of Adverse Drug Reactions (ADR) is one of the most important challenges in pharmacovigilance and health informatics. There are significant number of patient mortality and morbidity cases due to ADRs. The World Health Organization (WHO) estimates that ADRs contribute approximately 5% of all hospital admissions annually in USA alone. (Komagamine, 2024). Conventional methods for detection of ADR like spontaneous reporting and clinical trial data suffer from huge limitations such as underreporting, inadequate coverage of rare and long-term effects and delayed recognition. With the coming of digital health platforms and patient review systems, we have large volumes of unstructured textual data with valuable insights on drug safety profiles. Patient- generated reviews on pharmaceutical review platforms together with clinical trial data provide real-time evidence of drug experiences. The extraction of meaningful insights from the unstructured vast data using the traditional techniques are not feasible, necessitating the development of automated Machine Learning (ML) approach. Natural Language Processing (NLP) and ML methods have excellent solutions for automated ADR detection from text data. Deep learning architectures also help in capturing sequential dependencies and relationship in the context. Given the sparse nature of NLP-processed medical data, ensemble methods like Random Forest and XGBoost are also really effective.

Though these ML techniques have promising performance for ADR detection, their sophisticated inherent processing creates a significant barrier for clinicians, pharmaceutical firms and medical professionals. These stakeholders need transparent and interpretable

decision-making processes. The increasing use of opaque AI applications in this domain shows the critical importance of using explainable AI (XAI). These XAI techniques will help convert the black box of ML models into more understandable and acceptable models, increasing the trust in using AI in healthcare. This research studies the application of XAI in ADR detection models and attempts to provide complementary views on model decisions. The study reviews local explanation methods like SHAP (Shapley Additive explanations) and LIME (Local Interpretable Model Agnostic Explanations), global methods like Partial Dependence Plots and Tree interpreter and advanced gradient-based methods like Integrated Gradients. This research tackles the following research question: How can Machine Learning approaches be combined with explainable AI techniques to design a transparent and trustworthy Adverse Drug Reaction detection system? Within the scope of this research question, the study also investigates how different XAI methods provide insights into the ADR decision making process. The main aim of this research is to design and evaluate a framework for ADR classification that combines different ML techniques with diverse XAI methods. This work has many significant contributions in the domain of health informatics. It gives a comparative analysis of ML and DL approaches to ADR detection, unlike previous works that focus on single algorithms. Moreover, this work presents a systematic application of five distinct XAI methods, bringing out their strengths and weaknesses. Also, this study gives practical guidance for implementing XAI systems in healthcare settings, thus addressing the gap between theoretical XAI and real-world healthcare applications.

2 Related Work

Latest advancements the field of Explainable AI (XAI) has a great potential in addressing the application of AI in pharmacovigilance tasks (Lee et al 2023). The paper describes that XAI has the capability to interpret the rationale behind pharmacovigilance predictions made by the ML systems, that increase the confidence of healthcare professionals in these developments. This section comprehensively reviews the existing systems and body of literature in the integration of XAI in Adverse Drug Reaction detection.

2.1 XAI Methods in Healthcare Applications

A rigorous application of SHAP based explainability for cardiac drug toxicity prediction is implemented and discussed in another study (Pramudito et al, 2024). The authors implemented SHAP across six ML models – ANN, RF, SVM, KNN, RBF and XGBoost. The resulting SHAP values quantified individual contributions of the biomarkers. Some limitations of this study include its exclusive focus on cardiac toxicity and their lack of guidance on when SHAP selection is helpful. The study could also be improved by including domain knowledge from healthcare practitioners for biomarkers.

The concept of Contraction Analysis LIME (CA-LIME) is discussed in a paper (Wolf et al, 2023). It is a specialized LIME method for explaining cardiomyocyte beating pattern classifications. They used LIME's super pixel perturbations to make the explanations more clinically understandable. The study also uses validation with known drug effects which makes this study remarkable. However, this research also focusses on cardiomyocyte beating

patters and cannot be used to generalize to diverse ADR detection. Also, CA-LIME works with only CNN architecture.

Another study (Muhammad et al, 2024) discusses an explainable deep learning design for genitourinary cancer prediction. The DNN model achieved an accuracy of 90.64% with integration of SHAP for global feature importance, LIME for local explanations and Partial Dependence Plots for feature interactions. The strength of the study is its practical clinical focus focussing healthcare professionals and patients. However, there are shortcomings in this research as there is a lack of quantitative evaluation of the explanation quality or user studies validating the clinical interpretability. Also, the feature set is relatively small. Future improvements for this methodology could include temporal modeling capabilities like how explanations may change over time.

A tree based XAI method for giving local explanations called AraucanaXAI is introduced in another paper (Parimbelli et al, 2023). It strongly discusses the technical details of the comparison between LIME and SHAP on multiple clinical and health datasets. It also quantitatively evaluates the XAI methods using metrics like fidelity, identity, separability and speed. AraucanaXAI works by creating local dataset of neighbouring instances and these instances are re-labelled using predictions. Though this is a new method, it might be computationally time intensive and the current implementation lacks detailed visualization capabilities.

Another research (Kim et al, 2025) classifies XAI methods across three dimensions, passive or active, explanation types and local or global scope. It also uses perturbation analysis for quantitatively comparing Integrated gradients, Gradient SHAP, DeepLIFT and LRP. However, the study focusses only on attribution methods without discussing about other rule- based approaches.

2.2 XAI Applications in ADR and Drug Safety

The use of AI to detect drug-herb adverse interactions is discussed in a study (Spanakis et al, 2025). It talks about integrating the supervised and unsupervised ML models with global XAI techniques like LIME and SHAP for interpretability in pharmacological contexts, to understand how models make decisions at the global level. This study also mentions about the use of SHAP values to explain the CYP3A4 (an enzyme in the liver) induction mechanism. However, the paper lacks validation of the proposed models.

AI driven frameworks for ADR detection along with integration of explainability are gaining momentum. In a study (Khemani et al, 2025) integration of SHAP and LIME in ADR detection is discussed. It implements a hybrid approach joining structured and unstructured data with an extensive pre-processing pipeline with advanced NLP. It goes on to use SHAP and LIME which identifies drug-drug similarity scores as a key feature in predicting neurological ADRs. This study could be enhanced with a deeper analysis of the quality of the explainability and its clinical interpretability. This work provides a very solid foundation for XAI and ADR integration and its feasibility.

A methodological overview is presented in a study (Chandre et al, 2023) that discusses developing explainable AI models for patient readmission prediction that takes into factor their experience with post-surgery drugs. The group created a ML pipeline using decision trees, random forests and neural networks. The paper also suggests the use of partial

dependence plots to understand the effect a single feature has on the output keeping the contribution of other features constant. However, this research work could benefit from more technical implementation and use of XAI techniques.

The application of AI and XAI in pharmacovigilance is really important given the modern developments in this field. A PARISMA based review that was conducted (Dimitsaki et al, 2024) revealed the evaluation strategies for AI in pharmacovigilance like Fairness, Universality and Explainability. It also speaks about the use of innovative ML techniques like ConvSCCS with self-controlled case series and says that only 11% of the reviewed studies implemented XAI methods, demonstrating the critical research gap in the implementation of XAI. The research could benefit by including unstructured data too, crucial for ADR detection.

A robust ensemble of ML and DL models was designed in a study of AIE-DRP framework for ADR prediction (Lambay et al, 2022). It integrated NLP along with TF-IDF vectorization. The strength of this paper is the multi-model approach in using LGBM, CNN and LSTM. However, this paper lacks extensive discussion and implementation of LIME, SHAP or other XAI techniques, but briefly mentions it.

2.3 XAI Evaluation and Assessment Frameworks

A systematic review of XAI evaluation metrics has been performed in another study (Kadir et al, 2023). The authors conducted a PRISMA analysis of 6122 studies and showed that there is a scarcity of evaluation metrics for XAI methods. They identified eleven XAI evaluation metrics some of which are sensitivity, monotonicity, faithfulness correlation, selectivity and sufficiency. A remarkable contribution of this study is the dual taxonomy approach, metric-based classification and application-based classification that helps researchers to choose the relevant evaluation strategies based on their context. The mathematical formulations for all these metrics have also been provided. Some of the limitations of this study include the focus of metrics like pixel flipping and region perturbation for image-based data. The study however reveals that most XAI research relies on qualitative assessment without evaluation metrics and highlights that much needs to be done in this space of XAI.

In another research (Markus et al, 2021) gives a framework for XAI in healthcare and the evaluation metrics relevant for such applications. It distinguishes between interpretability and fidelity for AI models. The study establishes that there should be explainability that is both understandable and faithful, for the stakeholders to use it with full trust. The paper also classifies XAI methods based on models, attribution, example-based and score (global and local). However, the paper remains theoretical mainly and it does not provide concrete solutions for measuring the quality of explanations.

A comprehensive study (Bommer et al, 2024) enlists and discusses evaluation strategies for XAI methods which are robustness, faithfulness, complexity, localization and randomization. The strength of the paper is the development of quantitative metrics for each property, the use of skill scores and the practical use of them with prediction models like MLP and CNN. It also demonstrates that different XAI methods excel in different properties like gradient methods show strong randomization and LRP-z is good at faithfulness.

However, the limitation to this approach is the difficulty to adapt to heterogenous data types and the research could benefit by handling the scalability of these methods.

Tree based XAI methods are also gaining prominence in their use for interpreting tree-based ML models. In a study (Sharma et al, 2020) the SHAP TreeExplainer (SHAP-TE) and the TreeInterprter (TI) are contrastively evaluated for anomaly detection. The study focusses on real time applications of TI which shows faster performance (0.058 seconds) over the SHAP-TE model (3.44 seconds) per instances. The study also introduces evaluation metrics like rank biased overlap (RBO) for comparing feature rankings. It demonstrates that TI is better than SHAP-TE for attribution accuracy. However, the study cannot be called complete as it focusses on tabular data and the metrics focus keenly on statistical accuracy rather than clinical medical cases.

2.4 Clinical Applications

An extensive and comparative analysis of SHAP and LIME applied on cancer patient data is performed on another research (Mujawar et al, 2024). In this paper, four ML algorithms, Random Forest, Bagging, Gradient Boosted Trees and Logistic Regression are used. The strength of this research is the practical case study showing SHAP value computations of each feature for a 65-year-old patient for the survival probability. However, there are some shortcomings, like there is absence of quantitative evaluation of the XAI methods and the dataset used is small.

The integration of XGBoost with SHAP for disease classification has been discussed in a study (Yi et al, 2023). It implements a model that handles class imbalance efficiently using weight adjustment preventing data leakage common in medical AI systems. The detailed SHAP analysis gives clinically interpretable results and identifies two neuropsychological tests – CDRSB and ADAS13 as key risk features aligning with medical knowledge. It also has SHAP plots and partial dependence plots giving multi-faceted explanations. The study could be enhanced by taking into account unstructured data also, but overall, it offers an excellent reference for implementing SHAP and other XAI methods for medical AI and ADR detection.

The integration of XAI in predicting drug responses within the IGF1R signalling pathways in Cancer treatment was explored in a study (Janiel et al, 2024). Its strength is in combining ensemble ML models with XAI explanations. This helped clinicians to understand how cellular features affect the efficacy of the drug. It used dimensionality reduction techniques like t-SNE to get better insights. However, this research is limited to experimental setup without clinical validation. Also, its focus is on a specific pathway and a specific drug and it narrows down the applicability of XAI in the diverse healthcare domain.

2.5 Conceptual Frameworks and Reviews

There is an urgent need of applying explainable AI in AI applications in the healthcare domain. A study (Nguyen et al, 2024) addresses the critical need of XAI in Prognostics and Health management (PHM). It presents a systematic approach to categorize XAI methods in the PHM field and addresses the gap between AI performance and transparency in safety critical systems. The paper also gives guidance on how to select a proper XAI technique

based on the PHM requirements. However, this paper does not address any evaluation framework for the XAI methods and focusses more on the conceptual organization.

A rigorous overview of XAI applications in drug discovery is presented in a study (Alizadehsani et al, 2024). It explains that XAI models increase the interpretability of tasks like target identification, compound design and toxicity prediction of drugs. It strength include the research done in showcasing how XAI overcomes the black-box of neural networks. However, there are no discussions of any interpretability metrics and quantitative assessment of XAI methods.

A comprehensive review (Wang et al, 2022) shows how ML models detecting ADRs, often lack transparency. This research goes onto to classify XAI methods into ante hoc and post hoc methods. It proposes a weighted CRF model for detecting ADR. Tools like LIME, SHAP and attention mechanisms would be very useful for this. This work very firmly states that explainability is not only optional but essential in healthcare AI.

Another research (Jimenez-Lune et al, 2020) discusses that deep learning models are very effective in medical AI, but many methods are opaque. The paper also discusses XAI methodologies and their applicability in drug discovery pipelines. However, this study remains mainly conceptual and could be improved by experimenting and evaluating the XAI methods in ADR detection systems.

The literature review reveals some gaps in the application and evaluation of XAI methods in ADR detection systems. Drawing from the methodological foundations presented by Pramudito et al.'s SHAP implementation and other works, this research was motivated to implement and evaluate the XAI methods like LIME, SHAP, Integrated Gradients, partial Dependence Plots and TreeInterpreter. The evaluation framework and quantitative metrics put forth by Kadir et al. and other groups are also used to address the gap in evaluation of XAI methods. This multi- method approach aims to address the identified research gaps and create a transparent and trustworthy system for AI – driven pharmacovigilance systems.

3 Research Methodology

This study seeks to perform an experimental methodology to carry out a comparative analysis of multiple machine learning approaches to predict ADR and their combination with multiple XAI techniques. This research follows established approaches in ADR prediction while also addressing the critical gap in interpretability. A mix of both ML and deep learning methods are analyzed through multiple XAI frameworks, determining their interpretability characteristics suitable for healthcare applications. A systematic data science approach using the CRISP-DM approach, with data preprocessing, feature engineering, model development and interpretability check is implemented as part of this research.

3.1 Data Collection

The Drug Adverse Events dataset of the US Food and Drug Administration is used for analyzing the application of XAI in ADR detection. It comprises of patient generated reviews of medications. The features of the dataset are drug name, patient reported rating (0-10) for the drug, medical condition treated, effectiveness (of the medication) and side effects review and comments provided by the patient. All the reviews in the dataset are anonymized and

lacks any personally identifiable information. The textual reviews were consolidated for a comprehensive study to detect ADR.

3.2 Data Preprocessing

A multi-stage data preprocessing methodology was undertaken. First, text was normalized using the established Natural Language Processing (NLP) methods. Given the nature of the reviews, advanced NLP processing was also conducted to handle the negation sentiment in the reviews. The categorical variables were encoded. To extract the linguistic features, methods like n-gram analysis and TF-IDF matrix were also employed.

3.3 Initial ADR Labelling

For identifying ADR in the patient reviews, a keyword-based approach was used based on the domain knowledge of experts. (Borchert et al, 2019). Using a predefined lexicon of ADR terms, binary classification labels were assigned to the records.

3.4 Model Development

Initially, two ensemble models, the Random Forest and XGBoost models were developed to predict and classify the records as ADR or no ADR.
To handle the complexities and dependencies in the long texts, deep neural networks, the bidirectional LSTM was implemented.

3.5 Explainability Framework

To examine the incorporation of explainable AI techniques in ADR detection models, multiple XAI techniques were implemented. Feature importance analysis for the ensemble methods was carried out using Tree Interpreter and Partial Dependence Plots. These methods, which are also Global XAI techniques help in understanding the working of the whole model.
To understand why a specific decision was predicted by the model, that is to handle the local explainability, Local XAI methods like SHAP and LIME were used.
To capture the intricacies of the neural networks in the LSTM model, the XAI technique of Integrated Gradients was also tried. It provides comprehensive mathematical explanations for the DL model.

3.6 Evaluation Methodology

The ADR detection models were evaluated using the classification metrics. Macro and weighted averages were also computed for accounting the imbalance effects. To analyse the performance of the implemented XAI methods, established quantitative metrics were used, that measure their readability and utility. Metrics like faithfulness, fidelity and complexity were measured to determine which XAI method offers the most stable, interpretable and reliable method.

4 Design Specification

The main objective of the research is the development of a machine learning pipeline that can detect ADR and explain how that decision was arrived at. Medical data and patient reviews are complex and sensitive, so the results must be accurate and interpretable by the stakeholders. The system developed to implement this work consists of five core layers. The high-level architecture of the system shown in figure 1 displays these layers.

- Data Layer: This layer is concerned with the ingestion of available data so that it can be processed and prepared for the models in the next phase of the pipeline.
- Processing Layer: In this stage, the pre-processed data is used to engineer features from the textual reviews. Since text cannot be directly used for processing by the models, the text is transformed into relevant vector format for smooth handling and for getting better insights.
- Model Layer: It is at this stage that the different ML and DL models specified earlier are developed using the processed data. The models are trained and tested on the dataset and their performance evaluated for detecting ADR.
- Explainability Layer: In this layer, explainability is incorporated into the developed models. The different XAI techniques specified undertaken as part of this investigation are used to explain and interpret the decision made by the ML model.
- Evaluation layer: The XAI methods are evaluated for their interpretability performance using established XAI performance metrics in this layer. This would help in determining the most reliable and stable technique for use in ADR detection and healthcare applications.

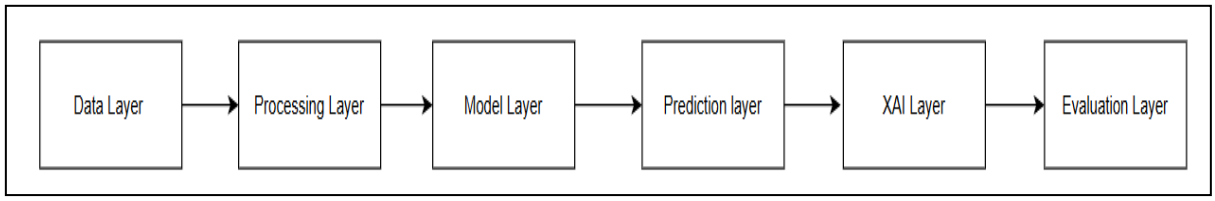

Fig 1. High Level Archtecture

4.1 Model Architecture

The model layer uses both ML and DL models to implement this project. All the models used in the process predict binary output that is 1 (ADR detected) or 0 (ADR not detected). The ML models used are the following. (Figure 2 displays the detailed architecture)

Random Forest - A random forest model uses the bootstrap aggregation to overcome overfitting and this is very useful for the high dimensional, text-derived features, the TF-IDF vectors generated from the patient reviews. Also, random forest models have been known to perform well in distinguishing true and false ADR reports (McMaster et al, 2019).

XGBoost - This model too addresses both overfitting and bias and is robust to noise and imbalanced data. XGBoost has been useful in efficiently performing in identifying risk factors from patient data.

To capture the long dependencies in the lengthy textual reviews more efficiently, deep learning networks, the Bidirectional Long Short-Term Memory (LSTM) network was used. LSTM – This model is a Bidirectional LSTM that takes in the word embeddings as input. This model has 64 LSTM units and the embedding dimension is 100. The maximum length of the sequence is 200. The model ran for 10 epochs with a batch size of 32.

4.2 Explainable AI Specifications

Five distinct XAI techniques are used in this research for explainability in ADR detection.

LIME (Local Interpretable Model agnostic Explanations) – This XAI method works by creating local approximations of complex models that aew linear, for individual predictions. It systematically removes features (words/TF-IDF terms in this case) and observed the impact on the prediction confidence. Through this and the feature importance scores that it generates, it can be determined which specific feature most influenced the decision of the model. LIME provides word level explanations and is model agnostic. Also, recent studies have shown that LIME is effective in medical applications, in detecting complex conditions (Gaspar et al, 2024), making it suitable in the ADR detection context.

SHAP (SHapley Additive exPlanations) – This method calculates the marginal contribution of every feature to the prediction outcome. For this, it uses the cooperative game theory principles. The values generated by the SHAP method quantify how much each feature contributes to the oredcition score. SHAP is particularly beneficial as it uses rigorous mathematical explanations which can be really useful in the high stakes medical setup. Research has also shown that SHAP has high potential for expalining predictions of ML models used in drug development (Bobadilla et al, 2024).

Integrated Gradients - This XAI method calculates feature contributions by integrating the gradients of the output produced by the models. It computes how much each feature token contributes to the final output by measuring the gradient information from the neautral baseline (in this context, zero embeddings) to the actual word embeddings. Integrated gradients is useful as it addresses the gradient saturation that is often a problem in deep neural networks. It is also helpful in giving detailed attributions to identify the linguistic patterns (Janiel et al, 2024) , given the context of textual patient reviews.

Tree Interpreter - This method is useful for interpreting the prediction made by tree based models, which are the Random Forest and XGBoost models in this study. For every prediction made, the TreeInterpreter traces the decision path along the tree. Then it computes how much each feature contributed to move the prediction farther from the average or baseline prediction. It is particularly helpful as with it one can detect the exact decision path taken by the model, so this becomes really beneficial for the medical professionals. Also, TreeInterpreters have been used and have been proven to be efficient in clinical decision making workflows in predicting hypertension (Elshawi et al, 2019).

Partial Dependence Plots (PDP) – PDPs are used to show the slightest effects one or two features have on the outcome of the model. They present the dependence between the target and a set of input features, marginalizing the values of the remaining features. PDPs provide a global insight to the model and can be very useful in revaling temporal patterns in the patient reviews.

4.3 Evaluation Framework

This research used the following quantitative metrics to evaluate the reliability of the implemented XAI techniques.

- Faithfulness: This measures how accurately the explanation demonstrates the internal reasoning of the model. It checks if the features flagged as important really influence the decision. (Bommer et al, 2024)
- Fidelity: This metric measures how well have the explanations approximated the actual model locally (Markus et al, 2021). It also represents the similarity between the original and explained model output.
- Complexity: It is the number of features that have non-zero explanation size or importance (Bommer et al, 2024). It helps in determining if the explanation is usable. Lower the value, more useful the explanation would be.

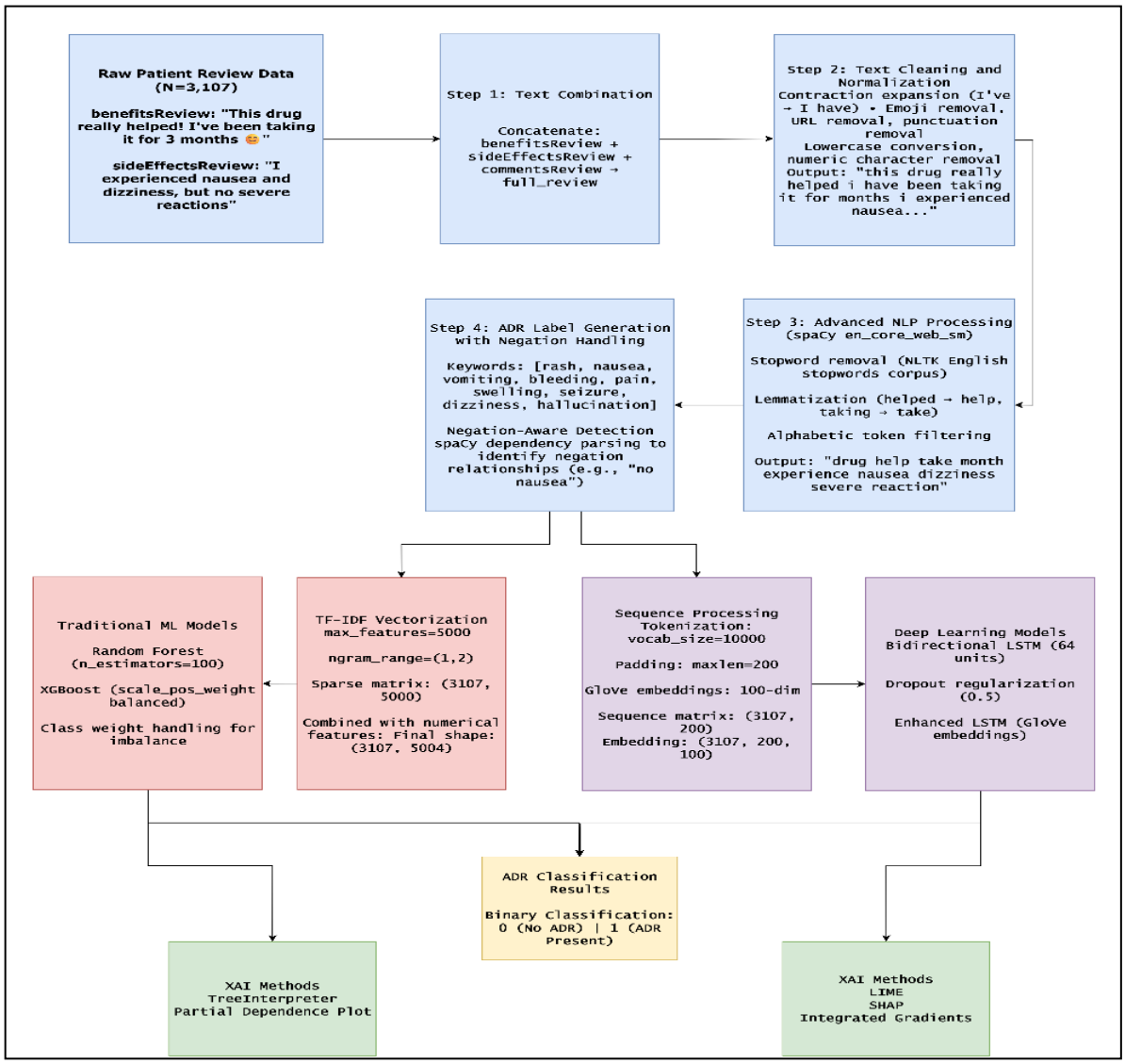

Fig 2. Detailed process architecture

5 Implementation

Detailed steps were taken to implement proper processing of the textual reviews so that they could be utilized efficiently to extract insights and make predictions by the models, develop the prediction models and integrate explainability in them.

5.1 Data Processing

The pre-processing pipeline began with transforming the raw review data into a suitable format for ML and DL analysis. The dataset consisted of 161,297 drug review records that included the review features in the form of benefits, side effects and general comments about the treatment by the patients. All the three different review features were unified by concatenation for easier handling. The missing values were handled using empty string substitution. The various NLP steps followed were: Expansion of contraction words using the contractions library, removal of emoticons, URL sanitization, elimination of hashtags and mentions, normalization of punctuation and removal of numbers, conversion to lowercase and whitespace standardization and spaCy library's English language model (en_core_web_sm) was used for lemmatization and stop word removal.

5.2 Label Engineering

The system to label ADR incorporating keyword-based detection and negation handling. A curated list of ADR indicative keywords - rash, nausea, vomiting, bleeding, pain, swelling, seizure, dizziness and hallucination was used (Leaman et al, 2010). Reviews containing any of these terms would be indicative of producing an adverse drug reaction.

To handle negation of terms in the reviews, the spaCy library was used. It was used to create dependency trees of each of the ADR key terms. Then a constant – 'neg' was defined that contains all the negative implying terms and words such as 'not', 'no', 'never' etc. The spaCy model then checks both the parent and child of the dependency tree of the ADR term to check for any negations. If negation is found, then the review would be flagged as 0 for ADR classification. An example of this is as follows:

Before implementing the spaCy negation function, the following review was flagged as 1 for ADR.

'use cramp badly would leave ball bed least day ponstel take pain away completely take edge much normal activity possible definitely miracle medication heavy bleeding clotting normal take pill onset menstrual cramp every hour take pill need day cramp cramp bad make sure take every hour dot medication stop work suddenly unfortunately take hour hour half kick back cramp moderate take every hour okay'

However, after using the spacy library for negation handling the same review was flagged as 0 for ADR.

'use cramp badly would leave ball bed least day ponstel take pain away completely take edge much normal activity possible definitely miracle medication heavy bleeding clotting normal take pill onset menstrual cramp every hour take pill need day cramp cramp bad make sure take every hour dot medication stop work suddenly unfortunately take hour hour half kick back cramp moderate take every hour okay'

This can be explained as follows: ‘ponstel take pain away completely’ – in this vector, ponstel takes the pain away, that is, it removes pain, so the spacy function detects this – ‘takes away’ from the dependency tree of the ADR keyword, pain.

5.3 Text Vectorization and Feature Engineering

Since text cannot be given as inputs directly to the LSTM model, conversion of the processed reviews into numerical representation was essential. This was done by using TensorFlow’s tokenizer using a vocabulary size of 10,000 words and out of vocabulary token handling that handles previously unseen text without failures. This feature representation was combined with the other numerical features like user ratings and effectiveness scores using sparse matrix concatenation methods for creating the input vectors for the ML models.

5.4 Model Development

To determine the effectiveness of Machine Learning and Deep Learning approaches and to provide a comparative context for XAI implementation, the ML and DL models were implemented. All these models predict 0 for no ADR and 1, if ADR is detected.

Random Forest Model - This model was designed with 100 decision trees by using balanced class weighting to address class imbalance. This model achieved moderate ADR detection with a precision of 85% and a recall of 21%.

XGBoost - This model used the gradient boosting method along with automatic scale_pos_weight calculation technique to handle class imbalance. Also, early stopping mechanism was used to prevent overfitting. This model performed better than the Random Forest with an overall accuracy of 93% and a significant recall of 77%.

LSTM - The bidirectional LSTM neural network was trained with regularization methods. It had an LSTM layer with 64 units, a dense hidden layer with 32 units with ReLU activation and a sigmoid output layer for the binary classification. This model also has dynamic sequence handling and a hierarchical processing.

The model performed significantly well with 92% precision and 88% recall, that it has a high No-ADR accuracy, so it can confirm the absence of adverse reactions efficiently. It has reasonable performance on the minority class too (ADR -1) with a 51% precision and 61% recall.

5.5 XAI Implementation

LIME Integration - The LIME-text library was used to implement this. A prediction wrapper, to interface with the LSTM model was also developed. This also provided probability distributions for both ADR and non-ADR classes. The prediction function converts the single-class sigmoid output to probability distributions for the ADR and no-ADR classes. The LIME explainer is designed with the [‘No ADR’, ‘ADR’] class names and it identifies the top 10 most influential features for each prediction.

Result - For the document/record at index 0, the LIME explainer revealed critical insights. The model assigned a probability of 82.3% to ‘No ADR’ and 17.7% to ‘ADR’, showing moderate confidence in classifying this review as not generating an ADR in the patient. LIME also showed what features actually contributed to the 17.7% of ADR class 1, the

presence of the term ‘pneumonitis’ (a lung condition) with a probability contribution of 0.005. The other terms in the review pushed up the class 0 (no ADR) prediction.

SHAP Integration - To implement the SHAP integration, a reference background distribution to calculate baseline expectations was developed. A representative sample of 100 reviews was randomly selected to serve as the background. Then a SHAP Kernel Explainer was implemented that takes padded tokens as input and returns probability predictions with which SHAP can evaluate the feature importance using coalition sampling.

Result - This revealed very detailed token level attributions for the reviews. Features that contributed towards ADR classification for the record at index 0 include terms like ‘dysfunction’ (0.001), ‘ventricular (0.001) and ‘progression’ (0.001) and ‘alone’ (0.001). The SHAP waterfall plot in figure 3 shows the contribution of individual features and how they additively combine. The baseline prediction values of 0.153681 is incrementally increased through the positive contributions to 0.177412, providing transparency to the cumulative decision-making process.

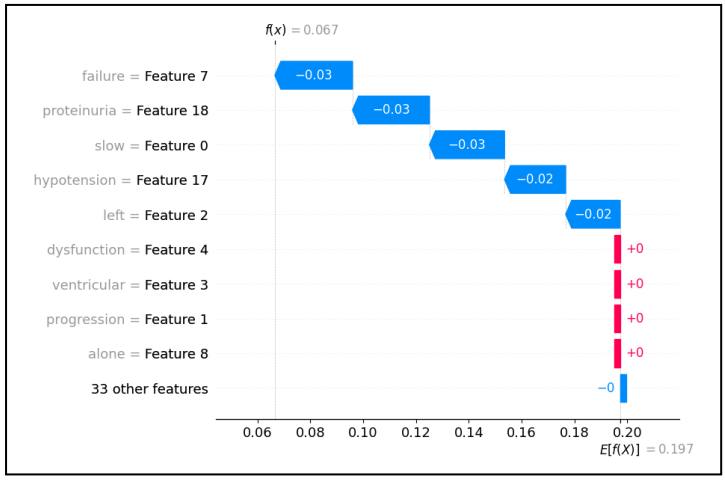

Fig 3. SHAP Waterfall Plot

Integrated Gradients - Integrated gradient uses the model’s internal gradient information to calculate the feature contributions unlike the model agnostic LIME. It divides the LSTM models into an embedding layer and the processing layers. It creates zero-filled embedding vectors to act as the neutral reference point. The TensorFlow’s GradientTape captures the gradients at each interpolation point and then the attribution of each feature is calculated.

Result - On applying this Integrated gradient explainer on the record at index 1, (give the record), features like ‘cramp’ (0.0015), ‘heavy’ (0.001) contributed positively to the probability prediction of ADR - 1 class. The ADR opposing features that contributed negatively to the prediction probability were ‘effective’ (-0.0015), ‘birth’ (-0.0042) and ‘control’ (-0.0022).

Tree Interpreter - An instance level feature contribution analysis was implemented using the treeinterpreter library. This provides transparency to the tree based Random Forest model by decomposing a certain prediction into its contributing components, giving interpretability to the decision. The tree interpreter calculated three quantities for each prediction of the model.

- Bias - This is the base value and represents the average output of all tree in the forest and is the prediction of the model without considering any features, just by looking at the dataset superficially.
- Feature Contributions - These are the computed values of the attribution of each of the features towards the output of that particular index.
- Final prediction – This is the sum of the bias and all the feature contributions resulting in the final class probability for that instance.

Result - The bias of the model without considering any of the features for the record at index 1 was 0.50 for No-ADR and 0.49 for ADR class. The features that contributed positively towards the ADR - 1 class include 'seizure' (0.0456), 'tired' (0.0156) and 'dose' (0.0107). The combination of all such positively contributing terms pushed the final prediction probability of ADR – 1 class to 15% and similarly the prediction probability for ADR – 0 class was 85%.

Partial Dependence Plots (PDP) - PDPs were developed to show the relationship and interpret how a selected feature affected the outcome. The PartialDependenceDisplay module of sklearn was used to generate the PDPs. These PDPs were especially useful for understanding the contribution of the numerical features like rating.

Results -

PDP for 'rating' - The plot in the figure 4 shows a clear negative relationship between rating provided by the patients and the model's prediction probability for ADR class 1. It means a higher rating corresponds to no adverse reaction produced. When the rating is low (1 or 2), the partial dependence is highest, around 0.205, contributing to ADR class 1. As the rating increases from 3 to 10, the partial dependence reduces to about 0.172. Thus, this PDP helps in interpreting the strength and direction of a feature's influence, here the rating score on the model's prediction.

PDP for 'effectiveness' - The label encoding of the effectiveness feature is as follows: - 0 (Considerably Effective), 1 (Highly Effective), 2 (Ineffective), 3 (Marginally Effective), 4 (Moderately Effective). The PDP plot for effectiveness in figure 5 shows that code 2 (Ineffective) corresponds to the highest partial dependence, that is, these drugs are most associated with higher chance of ADR 1 prediction. Code 0 and 1 have the lowest partial dependence and thus lower contribution towards ADR 1 class. Codes 3 and 4 have only moderate contribution to the ADR 1 prediction.

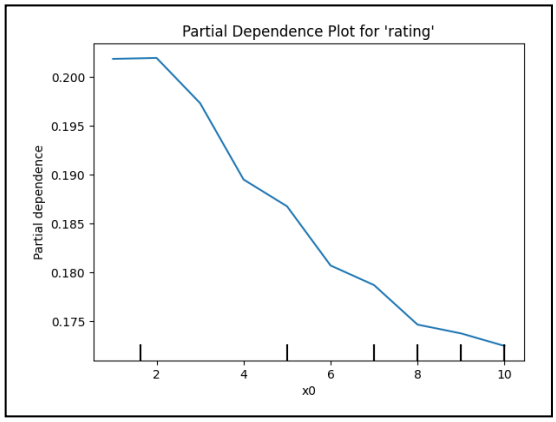

Fig 4. PDP for 'rating' feature

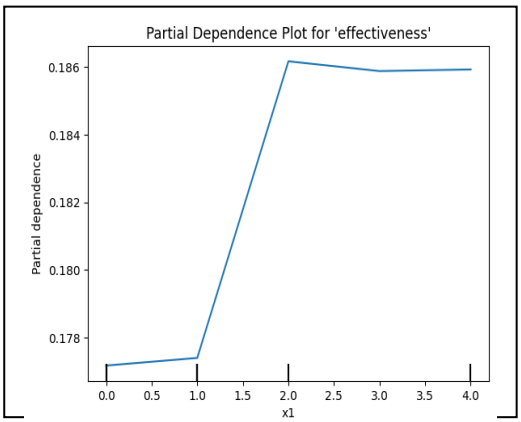

Fig 5. PDP for 'effectiveness' feature

6 Evaluation

The implementation of the ensemble ML techniques and the LSTM model for detecting ADR revealed important results. The Random Forest model, that was the baseline model had a moderate performance with 85% precision but a notably low recall of only 21%. This shows that the model failed to capture much of the actual ADR cases. The XGBoost model showed a superior performance with an overall accuracy of 93% and good recall score of 77%. The improvement on sensitivity is crucial for ADR detection where the missing positive cases can have serious clinical consequences. The bidirectional LSTM model achieved a balanced performance out of the three with 91% precision and 88% recall for ADR detection. The dynamic sequence handling and hierarchical processing helped in capturing the temporal dependencies and contextual relationships of the text data effectively.

The XAI methods surely helped in interpreting how these models arrived at the prediction. All the methods demonstrated how different features contributed to the prediction outcomes. It also helped in identifying the important features and the features that affected the output most.

The LIME implementation revealed the local contributions of the features towards the output. For a given record, the contribution of each and every feature (here TF-IDF terms) was computed and the terms contributing most towards positive ADR case (ADR class 1) were identified. The evaluation metrics of LIME were moderate with the LIME explainer having a faithfulness score of 0.0065, indicating very little deviation between the original and explained models' approximations. A fidelity score of 0.1612 suggests moderate accuracy in recreating the actual model's behaviour locally. The complexity score was 12 and it balances interpretability with comprehensiveness.

The SHAP explainer was able to provide detailed token-level attributions. The SHAP waterfall plot visualization was very useful in identifying the important terms and the additive nature of the feature contributions. SHAP had a superior faithfulness, 0.0482. showing it has better alignment with the actual model's decision-making process. It has the perfect fidelity score of 0.000 suggesting exact reproduction of the model behaviour. Its relatively lower complexity score of 10 indicates feature selection that is efficient. However, the SHAP explainer took significant time to produce the explanations, and was slower compared to the other methods. This could possibly be attributed to the intensive computations that SHAP internally performs.

The Integrated Gradients implementation utilized the LSTM models' internal gradient information to compute the feature contributions. Even this method was able to effectively calculate the contribution of each feature towards the predicted outcome. This method was useful for the model specific explanations directly derived from the LSTM's learned

representations. The faithfulness metric was moderate, 0.0344 and its fidelity was 0.0311, denoting reasonable approximation of the model behaviour. The particularly high complexity score of 107 suggests verbose explanations and it may have an impact on the interpretability for the end users.

The TreeInterpreter gave a detailed decomposition of the Random Forest predictions with the computation of the bias, feature contributions and the final predictions. This too helped in understanding how the prediction probability shifts for the ADR class 1 and ADR class 0 predictions. Through this, the features that contributed to this shift can be identified easily. This method had the highest faithfulness score of 0.0700 among all the methods, meaning it had the best alignment with the original model's decision-making process. Also, the perfect fidelity score of 0 strengthens the exact reproduction of the model's behaviour. However, the very high complexity score of 1606 suggests that the explanations it provided may be difficult for practical usability.

The Partial Dependence Plots correctly demonstrated the contribution and effect of a single feature on the model output, by holding the values of the remaining features constant. This helps in identifying how influential a feature can be, and how it affects the outcomes.

Table 1 shows the evaluation metrics of these XAI methods for the ADR detection system.

Method	Faithfulness	Fidelity	Complexity
LIME	0.0065	0.1612	12
SHAP	0.0482	0.0000	10
Integrated Gradient	0.0344	0.0311	107
TreeInterpreter	0.0700	0.0000	1606
PDP	N/A - Visual aid	N/A - Visual aid	N/A - Visual aid

Table 1. XAI methods evaluation metrics results

PDPs cannot be directly evaluated using these three metrics as they do not approximate local model behaviour, instead give a global visual representation of feature effects. PDPs serve as a complementary visual aid that increase the interpretability by displaying the influence of individual features. Though the TreeInterpreter method has the highest faithfulness and it efficiently explains the predictions made by the tree-based models, it has a high complexity. SHAP and LIME provide better balance for practical clinical use and also tackle the relationship among the features and ADR prediction. The evaluation addresses the primary research question of determining which XAI method provides the most reliable and interpretable explanations for deep learning-based ADR classification models. Statistical analysis and visual representations are employed to critically assess the performance, reliability, and practical utility of four XAI techniques: LIME, SHAP, Integrated Gradients, Tree Interpreter and PDP.

6.1 Discussion

Through this study, the experimental design was able to demonstrate the practical implementation of the use of multiple XAI methods for ADR detection and to determine their comparative effectiveness. The use of a significantly sized dataset that had unstructured and textual data along with a robust pre-processing pipeline validate the findings. However, this study is not complete and has few limitations too. The keyword-based ADR labelling, though practical, has the tendency to introduce bias by concentrating on predefined adverse reaction terms. This could then result in under representation of new or less common terms absent in the keyword list. With the negation handling implementation many previously ADR tagged reviews were rectified and correctly tagged as No ADR, however, there are chances of it missing complex linguistic structures that may change the meaning. Given the healthcare dataset, there is always a chance of imbalance in the dataset and there may arise the need to implement cross-validation methods.

The results of the implementation demonstrate differing suitability of XAI techniques for different requirements. The balanced performance across faithfulness, fidelity and complexity metrics of the model agnostic LIME and SHAP make them ideal for clinical decision support ADR systems. The TreeInterpreter and PDPs could be very beneficial in systems that use tree-based models. The findings also align with previous research in this field of applying XAI in healthcare AI and ADR detection. This implementation could be enhanced in the future by exploring ensemble XAI approaches that combine the best features of different methods. Also, making the SHAP method computationally less time intensive could also be explored as part of future works. More meaningful evaluation frameworks can be designed by developing domain specific faithfulness metrics tailored for healthcare applications. It would be invaluable if healthcare professionals could validate the practical utility of the explanations that have been produced.

7 Conclusion and Future Work

This research investigated the primary question – How can Machine Learning approaches be combined with explainable AI techniques to design a transparent and trustworthy Adverse Drug Reaction detection system? The study addressed three primary objectives: (1) to design and implement ML and DL models for ADR detection using NLP processing of patient reviews, (2) to integrate explainable AI techniques into these models and (3) to establish an evaluation framework comparing these methods. All the objectives were achieved through a systematic approach of processing 161,297 drug review records, developing sophisticated NLP preprocessing, implementing ML and DL models and integrating XAI techniques to them. The results demonstrated that the bidirectional LSTM model performed relatively better considering all the performance metrics and a higher recall. It was also able to capture temporal dependencies and long contextual relationships in the reviews. The integration of XAI methods proved very useful and effective, It helped in efficiently identifying the features that influence the outcome, features that crucially shift the prediction probabilities and the contribution of each feature to the final decision. SHAP and LIME provide optimal balance between accuracy and interpretability. Integrated Gradient also is effective, however this

technique along with SHAP took considerable time to produce the explanations. Some methodological innovations that were used and implemented include the sophisticated negation handling system using spaCy dependency parsing, the multi-metric evaluation framework for the XAI methods and the integration of both model agnostic and model specific XAI techniques. However, the research has some limitations, like the keyword-based ADR labelling approach, though is practical, can induce bias by considering only limited number of keywords. Also, the quality and clinical accuracy of the patient reported adverse reactions in the dataset remain uncertain without expert medical validation. Future research into this study should focus on the integration of more clinical data like laboratory results and medication history to make the ADR detection more comprehensive. The current methodology can be extended by incorporating temporal analysis of ADR using patient illness severity and recovery. Thus, XAI methods adapted for time-series data could be used to explain the patient's condition over extended treatment periods. Any practical use of this study should be accompanied by validation and verification by medical experts and healthcare regulatory agencies. This research successfully shows how explainable AI can provide interpretable and meaningful insights for ADR detection and how useful it is to increase the interpretability and trustworthiness of healthcare AI applications. As AI increasingly carves its way in healthcare, this study supports the development of AI systems that enhance rather than promote the use traditional black-box decision making models, ultimately contributing to patient safety and enhanced healthcare services.

References

Komagamine, J. (2024) *Prevalence of urgent hospitalizations caused by adverse drug reactions: a cross-sectional study*. Scientific Reports, 14, 6058. https://doi.org/10.1038/s41598-024-56855-z

Lee, S., Kim, S., Lee, J., Kim, J.-Y., Song, M.-H. & Lee, S., 2023. Explainable Artificial Intelligence for Patient Safety: A Review of Application in Pharmacovigilance. *IEEE Access*, 11, pp.50830–50840. https://doi.org/10.1109/ACCESS.2023.3271635

Pramudito, M.A., Fuadah, Y.N., Qauli, A.I. et al., 2024. Explainable artificial intelligence (XAI) to find optimal in-silico biomarkers for cardiac drug toxicity evaluation. *Scientific Reports*, 14, p.24045. https://doi.org/10.1038/s41598-024-71169-w

de Wolf, T.H., Snelders, M., Essers, J. & Smal, I., 2023. CA-LIME: An Algorithm that Explains Classification Results of Cardiomyocyte Beating. *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, Cartagena, Colombia, pp.1–4. https://doi.org/10.1109/ISBI53787.2023.10230622

Muhammad, D., Rafiullah, & Bendechache, M. (2024) *Improving Diagnostic Trust: An Explainable Deep Learning Framework for Genitourinary Cancer Prediction*. IET Conference Proceedings. ADAPT Centre, School of Computer Science, University of Galway, Ireland. doi:10.1049/icp.2024.3275

Parimbelli, E., Buonocore, T.M., Nicora, G., Michalowski, W., Wilk, S. and Bellazzi, R. (2023). *Why did AI get this one wrong? — Tree-based explanations of machine learning*

model predictions. Artificial Intelligence in Medicine, 135, 102471. https://doi.org/10.1016/j.artmed.2022.102471

Kim, S.G., Ryu, S., Jin, K. and Kim, H. (2025) *Quantitative comparison of explainable artificial intelligence methods for nuclear power plant accident diagnosis models*. Progress in Nuclear Energy, 180, 105605. https://doi.org/10.1016/j.pnucene.2025.105605

Spanakis, M., Tzamali, E., Tzedakis, G., Koumpouzi, C., Pediaditis, M., Tsatsakis, A. & Sakkalis, V., 2025. Artificial Intelligence Models and Tools for the Assessment of Drug–Herb Interactions. *Pharmaceuticals*, 18, p.282. https://doi.org/10.3390/ph18030282

Khemani, B., Malave, S., Shinde, S., Shukla, M., Shikalgar, R. and Talwar, H. (2025) *AI-driven pharmacovigilance: Enhancing adverse drug reaction detection with deep learning and NLP*. MethodsX, 15, 103460. https://doi.org/10.1016/j.mex.2025.103460

Chandre, P., Vanarote, V., Kuri, M., Uttarkar, A., Dhore, A. and Pathan, S. (2023) *Developing an Explainable AI Model for Predicting Patient Readmissions in Hospitals*. In: Proceedings of the 2023 2nd International Conference on Edge Computing and Applications (ICECAA), Namakkal, India, pp. 587–592. doi:10.1109/ICECAA58104.2023.10212152

Dimitsaki, S., Natsiavas, P. and Jaulent, M.-C. (2024) *Applying AI to Structured Real-World Data for Pharmacovigilance Purposes: Scoping Review*. Journal of Medical Internet Research, 26. https://www.jmir.org/2024/1/e57824

Lambay, M.A., Mohideen, S.P. & Rahman, B.S.A., 2022. AIE-DRP: Framework with Machine Learning and Deep Learning Models for Adverse Drug Reaction Prediction in Healthcare Use Case. *2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Coimbatore, India, pp.1330–1336. https://doi.org/10.1109/ICACCS54159.2022.9784963

Kadir, M.A., Mosavi, A. and Sonntag, D. (2023) *Evaluation Metrics for XAI: A Review, Taxonomy, and Practical Applications*. In: Proceedings of the 2023 IEEE 27th International Conference on Intelligent Engineering Systems (INES), Nairobi, Kenya, pp. 000111–000124. doi:10.1109/INES59282.2023.10297629

Markus, A.F., Kors, J.A. and Rijnbeek, P.R. (2021) *The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies*. Journal of Biomedical Informatics, 113, 103655. https://doi.org/10.1016/j.jbi.2020.103655

Bommer, P.L., Kretschmer, M., Hedström, A., Bareeva, D. and Höhne, M.M.-C. (2024) *Finding the Right XAI Method—A Guide for the Evaluation and Ranking of Explainable AI Methods in Climate Science*. Published online 25 June 2024; print publication 1 July 2024. https://doi.org/10.1175/AIES-D-23-0074.1

Sharma, P., Grossmann, G. and Ram, S. (2020) *Evaluating Tree Explanation Methods for Anomaly Reasoning: A Case Study of SHAP TreeExplainer and TreeInterpreter*. In: Grossmann, G. and Ram, S. (eds) *Advances in Conceptual Modeling. ER 2020*. Lecture Notes in Computer Science, vol. 12584. Springer, Cham. https://doi.org/10.1007/978-3-030-65847-2_4

Mujawar, S., Sayed, M. and Chaware, A. (2024) *Empowering Decision Making in Healthcare – A Comparative Analysis of eXplainable AI Techniques Using SHAP and LIME for Cancer Patients Data*. In: Proceedings of the 2024 3rd Edition of IEEE Delhi Section Flagship Conference (DELCON), New Delhi, India, pp. 1–5. doi:10.1109/DELCON64804.2024.10867049

Yi, F., Yang, H., Chen, D., Qin, Y., Han, H., Cui, J., Bai, W., Ma, Y., Zhang, R. and Yu, H. (2023) *XGBoost-SHAP-based interpretable diagnostic framework for Alzheimer's disease*. BMC Medical Informatics and Decision Making, 23, 137. https://doi.org/10.1186/s12911-023-02238-9

Janiel, J. & Rajendran, P.S., 2024. Leveraging Explainable AI for Improved Understanding and Prediction of Drug Responses in IGF1R Signaling Pathways. *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, Chennai, India, pp.1–5. https://doi.org/10.1109/ADICS58448.2024.10533479

Nguyen, D.A., Nguyen, K.T.P. and Medjaher, K. (2024) *Enhancing Trustworthiness in AI-Based Prognostics: A Comprehensive Review of Explainable AI for PHM*. In: Tran, K.P. (ed.) *Artificial Intelligence for Safety and Reliability Engineering*. Springer Series in Reliability Engineering. Springer, Cham. https://doi.org/10.1007/978-3-031-71495-5_6

Alizadehsani, R. et al., 2024. Explainable Artificial Intelligence for Drug Discovery and Development: A Comprehensive Survey. *IEEE Access*, 12, pp.35796–35812. https://doi.org/10.1109/ACCESS.2024.3373195

Wang, J., Yu, L.C. & Zhang, X., 2022. Explainable detection of adverse drug reaction with imbalanced data distribution. *PLoS Computational Biology*, 18(6), e1010144. https://doi.org/10.1371/journal.pcbi.1010144

Jiménez-Luna, J., Grisoni, F. & Schneider, G., 2020. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence*, 2, pp.573–584. https://doi.org/10.1038/s42256-020-00236-4

Borchert, J., Wang, B., Ramzanali, M., Stein, A., Malaiyandi, L. and Dineley, K. (2019) *Adverse events due to insomnia drugs reported in a regulatory database and online patient reviews: comparative study*. *Journal of Medical Internet Research*, 21(11), e13371. Available at: https://www.jmir.org/2019/11/e13371

McMaster, C., Liew, D., Keith, C. *et al.* (2019) *A machine-learning algorithm to optimise automated adverse drug reaction detection from clinical coding*. *Drug Safety*, 42, pp.721–725. https://doi.org/10.1007/s40264-018-00794-y

Gaspar, D., Silva, P. and Silva, C. (2024) 'Explainable AI for intrusion detection systems: LIME and SHAP applicability on multi-layer perceptron', *IEEE Access*, 12, pp. 30164–30175. https://doi.org/10.1109/ACCESS.2024.3368377

Ponce-Bobadilla, A.V., Schmitt, V., Maier, C.S., Mensing, S. and Stodtmann, S. (2024) 'Practical guide to SHAP analysis: Explaining supervised machine learning model

predictions in drug development', *Clinical and Translational Science*, 17(11), e70056. https://doi.org/10.1111/cts.70056

Elshawi, R., Al-Mallah, M.H. and Sakr, S. (2019) 'On the interpretability of machine learning-based model for predicting hypertension', *BMC Medical Informatics and Decision Making*, 19, 146. https://doi.org/10.1186/s12911-019-0874-0

Sokol, K. and Flach, P. (2019) 'Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches', *arXiv preprint* arXiv:1912.05100. Available at: https://doi.org/10.48550/arXiv.1912.05100

Leaman, R., Wojtulewicz, L., Sullivan, R., Skariah, A., Yang, J. and Gonzalez, G. (2010) 'Towards Internet-Age Pharmacovigilance: Extracting Adverse Drug Reactions from User Posts in Health-Related Social Networks', *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, Uppsala, Sweden, July 2010, pp. 117–125. Association for Computational Linguistics. Available at: https://aclanthology.org/W10-1915/