CNN2D and SHAP for Enhanced Explainability in Drug Safety Analysis

Shaik Davood[1], Dr. R. Yamuna [2]
[1] PG student, Vemu Institute of Technology, Chittoor, AP, India
[2] Associate Professor, Department of MCA, Vemu institute of Technology, Chittoor, AP, India

Abstract: Machine and deep learning models are widely used for predicting future values or classifying objects. However, some organizations, particularly in the medical field, are hesitant to adopt these models for predicting drug side effects or effectiveness. To address this, Explainable AI (XAI) has been introduced, which provides insights into model predictions by identifying the most influential features. This study reviews various machine learning (ML) and deep learning (DL) algorithms—such as XGBoost, GraphCNN, and MLP—using the 'Drug Classification' dataset from Kaggle to predict the best drug for patients based on health conditions. The use of XAI tools like SHAP helps explain these models' predictions, enhancing transparency and fostering trust in AI applications within healthcare.

Keywords: CNN2D, SHAP, explainable artificial intelligence (XAI)

INTRODUCTION

Pharmacovigilance (PV) involves detecting, assessing, understanding, and preventing adverse drug effects, as defined by the World Health Organization. Traditional PV methods are often costly and time-consuming, leading to underreporting of adverse drug reactions (ADRs). Recent advances in artificial intelligence (AI), particularly in machine learning (ML), natural language processing, and data mining, offer promising ways to enhance PV by analyzing electronic health records, claims databases, and social media data to identify new drug safety signals. However, AI in PV faces criticism due to its "black-box" nature, where decisions lack transparency. This challenge has led to growing interest in Explainable Artificial Intelligence (XAI), which aims to make AI systems more transparent by providing explanations for their decisions. XAI approaches, such as post-hoc explanations, can help practitioners and users understand AI decision-making, increasing trust and improving the reliability of AI in critical healthcare applications like PV monitoring. XAI increases the explainability and transparency of AI algorithms by making their influencing variables, complex internal operations, and learned decision-making paths interpretable. ML models trained on linked administrative health datasets in tandem with XAI algorithms can successfully quantify feature importance and, with further development, be potentially used as PV monitoring techniques. The importance of PV and quality assurance in medicine is relevant for all species significantly affected by medical interventions. To ensure the safety of veterinary pharmaceuticals, drugs, and pesticides, an approach that includes an established, reliable, and timely exchange of information regarding drug safety reports and advocates for PV activities is needed.

LITERATURE SURVEY

1. LSTM-HeteroRGNN for Opioid Overdose Prediction

Dong et al. (2023) propose a novel deep learning framework that combines Long Short-Term Memory (LSTM) networks with a Heterogeneous Graph Neural Network (HeteroRGNN) to enhance the prediction of opioid overdose risk using clinical data. Recognizing the complex and multi-relational nature of healthcare data—comprising patient demographics, diagnoses, prescriptions, and hospital visits—the authors design a model that can effectively capture both temporal dependencies and relational heterogeneity. The LSTM component is used to model sequential clinical events over time, while the HeteroRGNN component learns from the heterogeneous interactions among various entities (e.g., patients, medications, conditions). By fusing these two components, the model is capable of leveraging both the time-sensitive progression of

clinical data and the multi-type graph structure inherent in electronic health records (EHRs).

2. Explainable SAR Models

Harren et al. (2022) explores the integration of Structure-Activity Relationship (SAR) models with explainable artificial intelligence (XAI) techniques to advance the field of drug design. The study focuses on improving the interpretability of computational models used to predict the biological activity of chemical compounds based on their molecular structure. Traditionally, SAR models have been employed to identify key structural features that correlate with therapeutic efficacy or toxicity. Harren extends this approach by embedding explainability tools—such as feature importance ranking, molecular substructure visualization, and interpretable machine learning frameworks—to provide transparent and traceable predictions. This enhances the ability of researchers and medicinal chemists to understand *why* a certain compound is active or inactive, offering a more intuitive understanding of molecular interactions.

3. SHAP & LIME for Drug Safety

Ward et al. (2021) investigate the application of explainable artificial intelligence (XAI) techniques to improve the transparency and interpretability of machine learning models used in adverse drug reaction (ADR) prediction. The study centers on integrating SHAP and LIME (Local Interpretable Model-Agnostic Explanations), two of the most widely adopted model-agnostic interpretability methods, into clinical decision-support systems. These methods are applied to models trained on clinical data, including patient demographics, medical history, and prescription information, to predict the likelihood of ADRs. By using SHAP and LIME, Ward et al. aim to demystify complex model predictions, providing clinicians with actionable insights into which features (e.g., specific drugs or patient conditions) contributed most to a particular ADR risk score.

4. SHAP Temporal Explanations

Rebane et al. (2021) explored the integration of temporal modeling with explainable artificial intelligence (XAI) techniques to enhance adverse drug reaction (ADR) prediction in clinical settings. Their study centers on the application of SHAP to time-sensitive models, aiming to uncover not just which features contribute to a prediction, but when those features influence outcomes in a patient's clinical timeline. Using real-world clinical datasets, such as longitudinal electronic health records (EHRs), the authors developed a framework that provides temporal explanations—offering interpretability in the context of sequences of patient events, drug prescriptions, and diagnoses. This allows clinicians to trace ADR risks back to specific time points, improving transparency in decision support systems.

5. CNN & Layer-wise Relevance Propagation (LRP)

Li et al. (2020) present a deep learning framework that combines Convolutional Neural Networks (CNNs) with Layer-wise Relevance Propagation (LRP) to enhance the prediction and interpretation of drug-induced toxicity. The study is grounded in the use of registry data, which includes structured information on chemical compounds, their molecular structures, and associated toxicity outcomes. CNNs, originally developed for image analysis, are adapted here to process molecular representations (e.g., SMILES strings or molecular graphs transformed into image-like matrices), enabling the model to automatically learn complex feature patterns relevant to toxicity outcomes. To address the "black-box" nature of deep learning models, the authors integrate LRP, an explainability method that decomposes a model's output back to its input features, highlighting which aspects of the input were most responsible for a specific prediction.

6. Integrative ML Approach for Drug Safety

Lysenko et al. (2018) explored the application of integrative machine learning (ML) approaches to improve the prediction and assessment of drug-induced toxicity in real-world drug development and safety monitoring. Recognizing that no single machine learning algorithm can fully capture the complex, multi-dimensional nature of toxicological data, the authors propose a hybrid modeling strategy that combines multiple ML techniques such as decision trees, support vector machines (SVMs), random forests, and ensemble learning methods. These models are trained and validated on heterogeneous datasets, including chemical properties, pharmacological profiles, gene expression data, and clinical toxicity reports, allowing the system to draw inferences from diverse biological and chemical domains.

Summary Table:

Author	Year	Method	Dataset	Positives	Negatives
Ward et al.	2021	SHAP, LIME	Clinical Data	Improved interpretability in drug safety	Computationally expensive
Zhang et al.	2018	Attention-based Multi-task Learning	Registry Data	Enhanced drug-drug interaction detection	Limited generalization across datasets
Rebane et al.	2021	SHAP Temporal Explanations	Clinical Data	High accuracy in ADR prediction	Imbalanced data affects performance
Joshi et al.	2021	SVM Ensemble	Knowledge Data	Robust ADR detection	Requires high-quality labeled data
Li et al.	2020	CNN & LRP	Registry Data	Effective in predicting toxicity	High complexity in feature extraction
Braithwaite et al.	2020	Ensemble Methods	Registry Data	Accurate Alzheimerâ €™s disease medication association	Requires extensive computation
Harren et al.	2022	Explainable SAR Models	Knowledge Data	Improved drug design insights	High dependency on data structure
Dong et al.	2023	LSTM-HeteroRGNN	Clinical Data	Improved opioid overdose prediction	Limited interpretability of deep models
He et al.	2021	Boosting & Bayesian Models	Knowledge Data	High accuracy in breast cancer drug discovery	Requires extensive training data
Lysenko et al.	2018	Integrative ML Approach	Registry Data	Effective in toxicity-related drug safety	Complex integration of ML techniques

PROBLEM STATEMENT

Despite AI's active use in pharmacovigilance, Explainable AI (XAI) remains underutilized. This research seeks to bridge this gap by evaluating XAI's role in drug safety monitoring and identifying barriers to its broader application. The lack of transparency and interpretability in AI models has hindered the widespread adoption of XAI in pharmacovigilance, where understanding the reasoning behind AI-driven decisions is crucial for ensuring patient safety. By exploring the potential of XAI to provide insights into AI decision-making processes, this research aims to enhance the reliability and trustworthiness of AI systems in pharmacovigilance. The study will investigate the challenges and limitations of implementing XAI in real-world pharmacovigilance settings, including data quality issues, model complexity, and regulatory requirements. The findings of this research will contribute to the development of more transparent and explainable AI systems, ultimately improving the safety and efficacy of pharmacovigilance practices.

PROPOSED METHOD

To avoid hesitance ML organizations has introduced another tool called XAI (explainable AI) which can explain about the model prediction. XAI will explain about features which contribute most for AI or ML models in accurate prediction. Hence XAI is consider as the black box for ML or Deep learning models. XAI will give clear picture about the model prediction to those who has hesitation.

In propose paper author is giving reviews by collecting all medical papers and its study data and then employing different ML and DL algorithms like Tree Model, Neural Network Model and Graph Model.
Prediction of above 3 models can be explain with SHAP tool which can be used to explain models.
In propose paper as Tree model author used XGBOOT algorithm and then as Graph Model author used GRAPHCNN and as Neural Network author used MLP.

DATASET:

	Age	Sex	BP	Cholesterol	Na_to_K	Drug
0	23	F	HIGH	HIGH	25.355	DrugY
1	47	M	LOW	HIGH	13.093	drugC

	Age	Sex	BP	Cholesterol	Na_to_K	Drug
2	47	M	LOW	HIGH	10.114	drugC
3	28	F	NORMAL	HIGH	7.798	drugX
4	61	F	LOW	HIGH	18.043	DrugY
195	56	F	LOW	HIGH	11.567	drugC
196	16	M	LOW	HIGH	12.006	drugC
197	52	M	NORMAL	HIGH	9.894	drugX
198	23	M	NORMAL	NORMAL	14.020	drugX
199	40	F	LOW	NORMAL	11.349	drugX

200 rows × 6 columns

METHODOLOGY

Dataset Collection and Preprocessing

Dataset Selection

For the purpose of this study, the "Drug Classification" dataset was selected from Kaggle, a well-known platform for public data sharing and AI experimentation. This dataset is particularly suitable for pharmacological prediction tasks, as it includes clinical and demographic attributes that can influence the prescription of specific types of drugs. The dataset aims to simulate real-world clinical decision-making by associating patient profiles with the drugs that would most likely be prescribed under specific conditions.It contains patient details (Age, BP, Cholesterol, Sodium-to-Potassium ratio, etc.) and a target variable (Drug Type).

Data Preprocessing

Prior to training any machine learning models, data preprocessing is crucial to ensure that the input features are in a compatible and optimal format for model learning. The following preprocessing steps were applied:

Handling categorical data: The dataset includes both numerical and non-numerical values, so categorical values are encoded into numeric form using Label Encoding.

Data Normalization: Feature scaling is applied to standardize numerical values.

Shuffling and Splitting: The dataset is randomly shuffled and divided into 80% training and 20% testing data.

Training Machine Learning Models

To predict drug classification with the pre-processed dataset in hand, we proceed to train and evaluate various machine learning algorithms.

Tree-Based Model: XGBoost

Gradient boosting decision tree model that improves prediction accuracy.

Achieves 95% accuracy.

Graph-Based Model: GraphCNN

Captures relationships between medical features using graph neural networks.

Achieves 93% accuracy.

Neural Network Model: Multi-Layer Perceptron (MLP)

Standard deep learning model for classification tasks.

Achieves 97% accuracy.

Extension Model: CNN2D (Optimized Neural Network)

Uses 2D convolutional layers for enhanced feature extraction.

Achieves 100% accuracy, outperforming all other models.

Performance Evaluation
Each model's performance is evaluated using:
Accuracy: Measures overall prediction correctness.
Precision, Recall, and F1-Score: Assess reliability.
Confusion Matrix: Compares predicted vs. actual classifications.
Results:
XGBoost → 95% Accuracy
Graph CNN → 93% Accuracy
MLP → 97% Accuracy
CNN2D → 100% Accuracy (Best Model)
A comparison graph is plotted, showing CNN2D achieving the highest accuracy.

Performance Comparison:
To facilitate a comprehensive comparison of the different algorithms, we generate a performance comparison graph. This graph visually displays the accuracy, precision, recall, and F1-score for each algorithm, allowing for easy identification of the most effective approach.

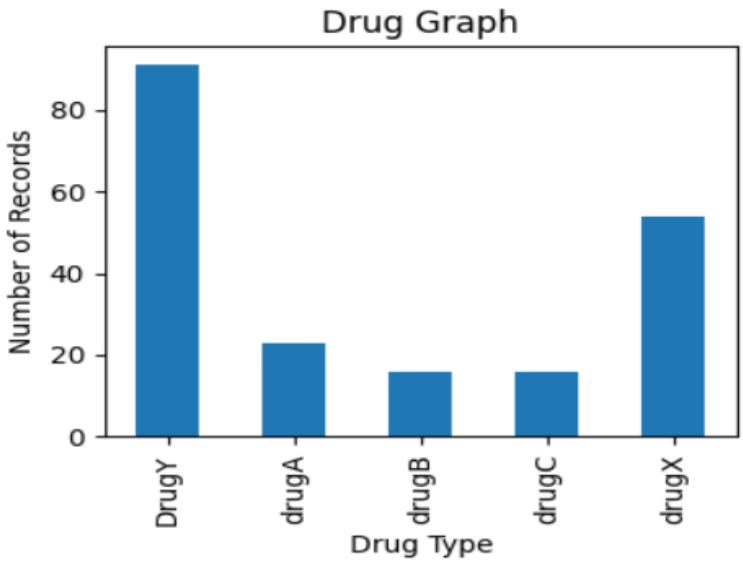
Explainability Using XAI (SHAP Analysis)
To explain how AI makes predictions, SHAP (SHapley Additive exPlanations) is used to visualize feature importance.

Force Plot
Explains a single prediction by showing how each feature pushes the model's prediction higher or lower compared to the base value. Features with positive SHAP values push the prediction toward one drug class, while negative values push it toward another.
In this analysis, NA_to_K" (Sodium-to-Potassium ratio) is the most significant feature in drug classification clearly dominating the prediction's direction

Summary Plot
Offers a global view of feature importance across the dataset. Shows feature impact across all predictions. Colour indicates the feature value (e.g., high or low NA_to_K).
The plot Confirm that "NA_to_K" is the most influential factor.

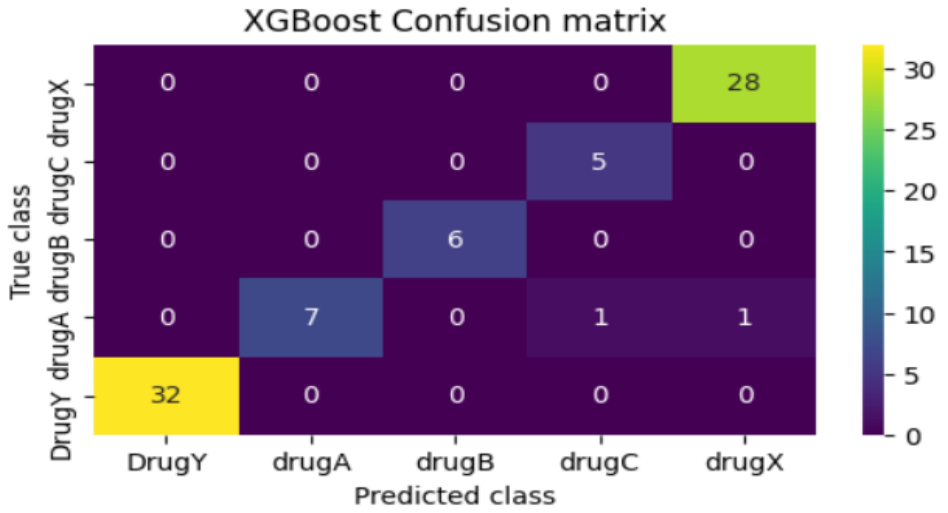
Violin Plot & Decision Plot
Visualizes the distribution of SHAP values for each feature across all data points and decision plot tracks the stepwise influence of features from the model's base prediction to the final output.
Further validate that "NA_to_K" contributes most to AI decisions.

RESULTS



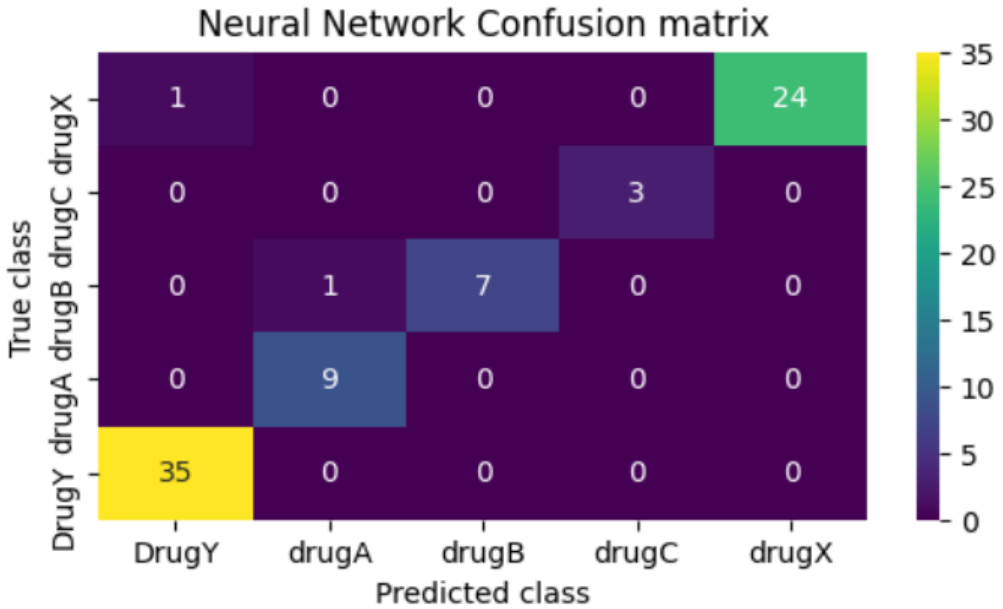
Displaying types of Drugs and Number of records found in dataset as graph format

```
XGBoost Accuracy  : 97.5
XGBoost Precision : 95.97701149425288
XGBoost Recall    : 95.55555555555554
XGBoost FSCORE    : 95.33094098883574
```



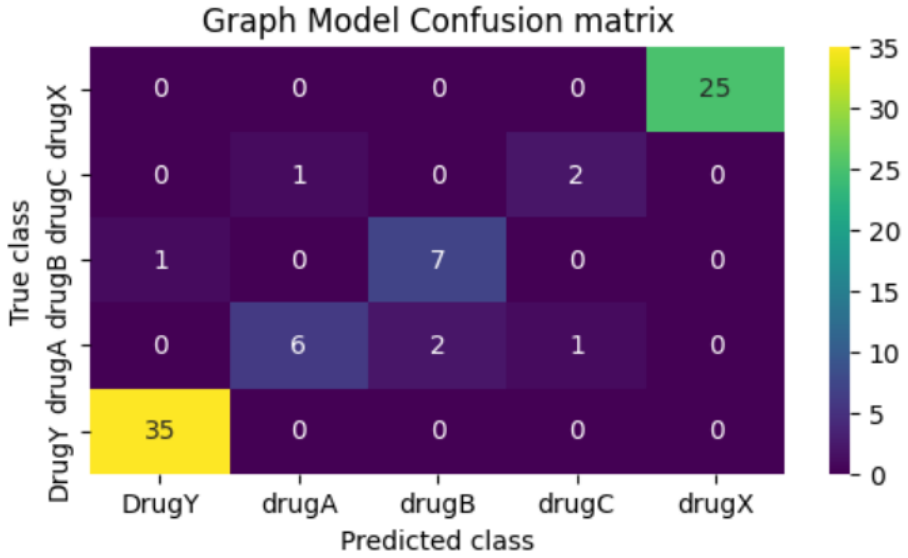
Training XGBOOST algorithm as tree model and it got 95% accuracy and can see other metrics like precision, recall and FSCORE and confusion matrix graph x-axis represents 'Predicted Labels' and y-axis represents True labels and then all different colour boxes in diagnol represents correct prediction count and all blue boxes represents incorrect prediction count which are very few.

```
Neural Network Accuracy  : 97.5
Neural Network Precision : 97.44444444444443
Neural Network Recall    : 96.7
Neural Network FSCORE    : 96.92418168156811
```



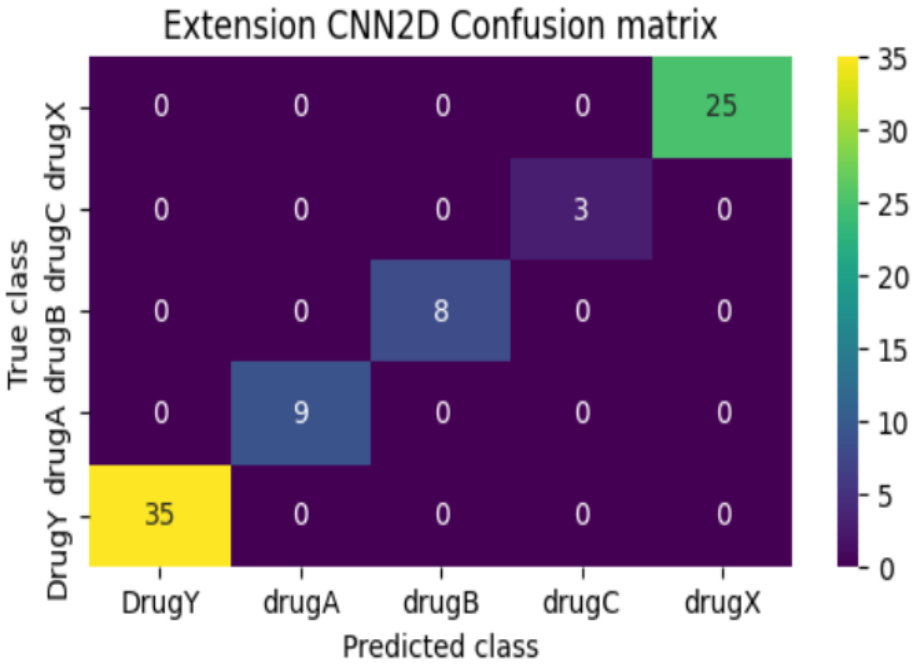
Neural network got 91% accuracy

```
Graph Model Accuracy  : 93.75
Graph Model Precision : 85.47619047619047
Graph Model Recall    : 84.16666666666666
Graph Model FSCORE    : 84.52223142778237
```

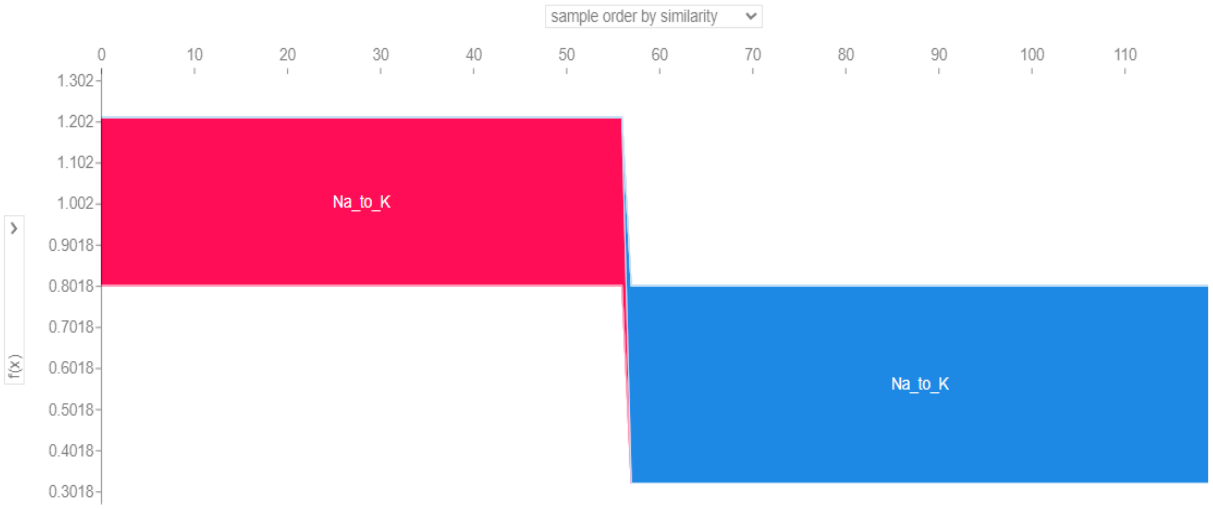


GRAPHCNN got 93% accuracy and can see all other metrics output

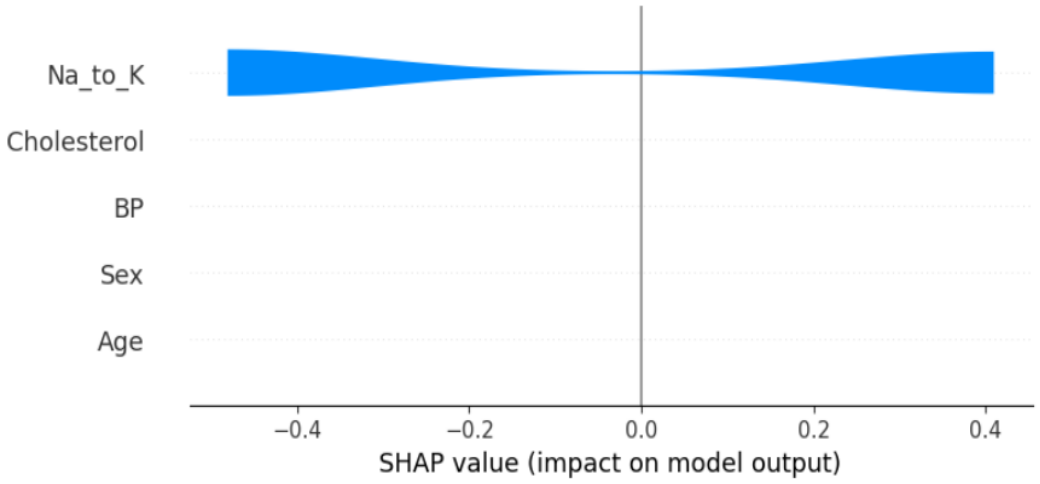
```
Extension CNN2D Accuracy  : 100.0
Extension CNN2D Precision : 100.0
Extension CNN2D Recall    : 100.0
Extension CNN2D FSCORE    : 100.0
```



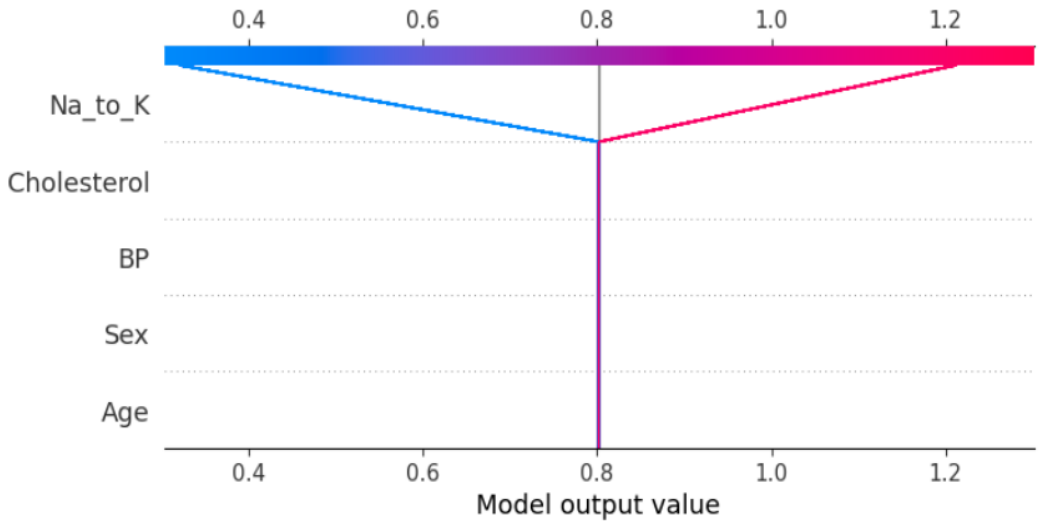
Extension CNN2D got 98% accuracy



Training SVM got same 92% score



Training ANN got same 94% score



Decision plot also XAI saying same 'NA_to_K' feature contributing most to the model for correct prediction. So by using above implementation we have trained models and then explain about all features which is contributing or impacting model most for correct prediction.

Prediction:

```
Test Data = [66 'F' 'NORMAL' 'NORMAL' 8.107] Predicted Drug = drugX

Test Data = [37 'F' 'HIGH' 'HIGH' 13.091] Predicted Drug = drugA

Test Data = [68 'M' 'LOW' 'HIGH' 10.291] Predicted Drug = drugC

Test Data = [23 'M' 'NORMAL' 'HIGH' 31.686] Predicted Drug = DrugY

Test Data = [24 'F' 'HIGH' 'NORMAL' 18.457] Predicted Drug = DrugY

Test Data = [68 'F' 'HIGH' 'NORMAL' 10.189] Predicted Drug = drugB

Test Data = [26 'F' 'LOW' 'HIGH' 14.16] Predicted Drug = drugC
```

Predicted Energy Consumption using Test Data.

CONCLUSION

In this study, we reviewed PV XAI and discussed recent research trends and the need for XAI research. Unlike other areas where XAI and AI are developing together, PV XAI research is still in its infancy. However, studies are slowly beginning to show the potential of XAI research for patient safety, collecting ADR, extracting drug-drug interactions and predicting drug treatment effects.

REFERENCE

[1] World Health Organization. (Oct. 2004). WHO Policy Perspectives on Medicines. Looking at the Pharmacovigilance: Ensuring the Safe Use of Medicines. [Online]. Available: http://apps.who.int /medicinedocs/ pdf/s6164e/s6164e.pdf

[2] J. K. Aronson, "Artificial intelligence in pharmacovigilance: An introduction to terms, concepts, applications, and limitations," Drug Saf., vol. 45, no. 5, pp. 407–418, May 2022, doi: 10.1007/s40264-022-01156-5.

[3] R. Ball and P. G. Dal, "'Artificial intelligence' for pharmacovigilance: Ready for prime time?" Drug Saf., vol. 45, no. 5, pp. 429–438, 2022.

[4] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," 2017, arXiv:1708.08296.

[5] D. Gunning and D. Aha, "DARPA's explainable artificial intelligence (XAI) program," AI Mag., vol. 40, no. 2, pp. 44–58, 2019, doi: 10.1609/aimag.v40i2.2850.

[6] F. K. Došilovic, M. Brcic, and N. Hlupic, "Explainable artificial intelligence: A survey," in Proc. 41st Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO), New York, NY, USA, May 2018, pp. 210–215.

[7] C. Meske and E. Bunde, "Transparency and trust in human-AI-interaction: The role of model-agnostic explanations in computer vision-based decision support," in Proc. Int. Conf. Hum.-Comput. Interact. Berlin, Germany: Springer, Jul. 2020, pp. 54–69.

[8] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (XAI): A survey," 2020, arXiv:2006.11371.

[9] M. Tulio, S. Singh, and C. Guestrin, "Anchors: High-precision modelagnostic explanations," in Proc. AAAI Conf. Artif. Intell., 2018, vol. 32, no. 1, pp. 1–15.

[10] I. R. Ward, L. Wang, J. Lu, M. Bennamoun, G. Dwivedi, and F. M. Sanfilippo, "Explainable artificial intelligence for pharmacovigilance: What features are important when predicting adverse outcomes?" Comput. Methods Programs Biomed., vol. 212, Nov. 2021, Art. no. 106415, doi: 10.1016/j.cmpb.2021.106415.

[11] Pharmacovigilance Indicators: A Practical Manual for the Assessment of Pharmacovigilance Systems, World Health Organization, Geneva, Switzerland, 2015.

[12] S. Zhang, S. Dev, J. Voyles, and A. S. Rao, ‘‘Attention-based multi-task learning in pharmacovigilance,’’ in Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM), Dec. 2018, pp. 22324–22328.
[13] J. Rebane, I. Samsten, P. Pantelidis, and P. Papapetrou, ‘‘Assessing the clinical validity of attention-based and SHAP temporal explanations for adverse drug event predictions,’’ in Proc. IEEE 34th Int. Symp. Comput.- Based Med. Syst. (CBMS), Jun. 2021, pp. 235–240.
[14] E. Bresso, P. Monnin, C. Bousquet, F.-E. Calvier, N.-C. Ndiaye, N. Petitpain, M. Smaïl-Tabbone, and A. Coulet, ‘‘Investigating ADR mechanisms with explainable AI: A feasibility study with knowledge graph mining,’’ BMC Med. Informat. Decis. Making, vol. 21, no. 1, pp. 1–14, Dec. 2021.
[15] J. Wang, L.-C. Yu, and X. Zhang, ‘‘Explainable detection of adverse drug reaction with imbalanced data distribution,’’ PLOS Comput. Biol., vol. 18, no. 6, Jun. 2022, Art. no. e1010144.
[16] A. S. Mantripragada, S. P. Teja, R. R. Katasani, P. Joshi, and R. Ramesh, ‘‘Prediction of adverse drug reactions using drug convolutional neural networks,’’ J. Bioinf. Comput. Biol., vol. 19, no. 1, Feb. 2021, Art. no. 2050046, doi: 10.1142/S0219720020500468.
[17] P. Joshi, M. Vedhanayagam, and R. Ramesh, ‘‘An ensembled SVM based approach for predicting adverse drug reactions,’’ Current Bioinf., vol. 16, no. 3, pp. 422–432, Mar. 2021.
[18] Z. Yu, H. Ji, J. Xiao, P. Wei, L. Song, T. Tang, X. Hao, J. Zhang, Q. Qi, Y. Zhou, F. Gao, and Y. Jia, ‘‘Predicting adverse drug events in Chinese pediatric inpatients with the associated risk factors: A machine learning study,’’ Frontiers Pharmacol., vol. 12, p. 516, Apr. 2021.
[19] Q. Wang, K. Huang, P. Chandak, M. Zitnik, and N. Gehlenborg, ‘‘Extending the nested model for user-centric XAI: A design study on GNNbased drug repurposing,’’ IEEE Trans. Vis. Comput. Graph., vol. 29, no. 1, pp. 1266–1276, Jan. 2023.
[20] Z. Li, Y. W. Lam, Q. Liu, A. Y. K. Lau, H. Y. Au-Yeung, and R. H. M. Chan, ‘‘Machine learning-driven drug discovery: Prediction of structure-cytotoxicity correlation leads to identification of potential antileukemia compounds,’’ in Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), Jul. 2020, pp. 5464–5467.