International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.3 (2025) pp. 3896-3909
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

Enhancing Drug-Drug Interaction Prediction with Explainable AI: Integrating GANs for Improved Clinical Transparency

Bareq Kadhim Faraj[1]*, Amir Lakizadeh[2]

[1]Computer Engineering and Information Technology Department, University of Qom, Qom, Iran
* **Corresponding Author Email:** b.alghanimi@stu.qom.ac.ir - **ORCID:** 0009-0005-0557-4980

[2]Computer Engineering and Information Technology Department, University of Qom, Qom, Iran
Email: amir3@gmail.com - **ORCID:** 0000-0001-9870-3676

Article Info:

DOI: 10.22399/ijcesen.2551
Received : 12 March 2025
Accepted : 17 May 2025

Keywords

Drug-Drug Interactions (DDIs),
Artificial Intelligence (AI),
Explainable AI (XAI),
Generative Adversarial Networks (GANs),
XGBoost,
SHapley Additive exPlanations (SHAP)

Abstract:

Drug-drug interactions (DDIs) are critical in polypharmacy, where the concurrent use of multiple drugs can lead to synergistic effects or adverse drug events (ADEs). The latter can significantly impact patient morbidity and mortality. The rapid introduction of new drugs further complicates the prediction of DDIs, making traditional wet-lab verification methods both time-consuming and resource-intensive.Problem Statement: While artificial intelligence (AI) models have been employed to predict DDIs, the development of highly complex "black-box" models poses challenges in terms of interpretability and trust in clinical settings. There is a pressing need for explainable AI (XAI) approaches to ensure these models are both accurate and transparent. This study utilizes a comprehensive dataset from DrugBank, encompass- ing various drug interactions. We implemented data preprocessing steps, including handling missing values, encoding categorical variables, and normalizing the data. To address data scarcity, we employed Generative Adversarial Networks (GANs) to generate synthetic data, which was combined with real data to enhance the training dataset. The augmented dataset was then used to train an XGBoost model, optimized for binary classification. To ensure interpretability, we integrated SHapley Additive exPlanations (SHAP) to analyze feature im- portance and model decision-making processes. The XGBoost model demonstrated high predictive accuracy with a validation accuracy of 99.06%, precision of 98.73%, recall of 99.03%, and an F1 score of 98.88%. SHAP analysis provided clear insights into feature importance, highlighting the most influential fea- tures in the model's predictions and enhancing the transparency of the decision-making pro- cess. The combination of advanced machine learning techniques and explainable AI methods effectively addresses the challenges of DDI prediction. The proposed approach not only achieves high predictive performance but also ensures model interpretability, fostering trust and adoption in clinical applications. This methodology offers significant potential for improving patient safety and treatment outcomes.

1. Introduction

Multiple drugs have been described whereby the presence of one drug may affect another in a patient using multiple medications. In the best-case scenarios these interactions elicit additive effects and therapeutic outcomes. However, in treatments for multiple diseases, cases of ADEs which lead to toxicity or reduced treatment effectiveness harm patients and contribute to morbidity/mortality [1, 2, 3]. Further, an active introduction and approval of new drugs and Indications leave a high propensity of achieving DDIs [4, 5]. Experiments to confirm DDIs are often performed in wet labs, which are costly and require a lot of time; therefore, they cannot be used frequently or in large numbers. As such, there have been uses of Artificial Intelligence (AI) models to forecast DDIs [6, 7, 8, 9]. These models have grown along with the development and enhancement of drug-database resources for aiding in clinical decision-making.

However, further improvement of AI based DDI prediction has resulted in the creation of com- plicated 'AI black-box models. Such high-level

models even though have increased performances are less easy to explain or interpret and hence less transparent to the users [10]. However, the less complex models are easier to understand, but they perform poorly most among the time [11]. Another critic came from the group's realization that in certain spheres such as medicine 'a fine line between getting the function right and interpretability has to be drawn'. The technical aspects of some models include such problems as opacity, which can include both false and accurate results, therefore, opacity can narrow the trust of clinicians and patients in highly complex models utilized in the medical field. Hence, explainable AI (XAI) has emerged, which deals with techniques for understanding the working of machine learning algorithms. Therefore, XAI seeks to develop safe, reliable, and easily explainable DDI prediction models to support clinical practice and ensure that the generated messages also offer cogent reasons for the produced predictions.

2. Literature Review

One of the most frequently used methods to predict Drug-Drug Interactions (DDIs) is Support Vector Machine (SVM), which has shown a very strong predictive power with AUCs that range from 0.565 to 0.985 across different applications [12, 13, 6, 14, 15, 16, 17]. The number of features that the predictive model contains affects its capability to predict. For example, an increase of 0.02 in the F-measure score (from 0.5786 to 0.5965) was reported by a study that employed feature reduction techniques [16]. SVMs are simply one of the most famous members of a class kernel machines, which in turn is used for pattern analysis. MethodsKernel classifiers, e.g., all- paths graph (APG) [18], kBSPS for short k-band shortest path spectrum kernels can extract local information from the paths that connect to a pair of vertices in each other using up to l edges magnifiers concatenated with at most one edge parameter which has minimum weight per vertex classification on real-world data and compound similarity [19] related contexts.Syntax-tree kernel also exists as shallow linguistic (SL)-kernel based methods have commonly been applied in drug pair classification context[20, 21]. Notably, Thomas et al. SL, APG kernels were not only the best F1-Score (0.606) out of all other statistical features like case-based reasoning and ensemble learning
[18] Additionally, Zhang et al. [22], label propagation algorithms were used for the small-labeld nodes problems in undirected weighted network.

LR Logistic regression (LR) has commonly been used for DDI prediction models. However, Xie et al. [23] combined active learning, random negative sampling and uncertainty sampling in clinical safety DDI information retrieval (DDI-IR) analysis for SVM LR. An alternative strategy, the Drug-Entity-Topic (DET) model used Bayes-rules to capitalize on augmented text-mining features for improved performance in discrimination and calibration of predictions [24]. In response to the growing need for adverse DDIs (ADDI) signal detection, a Bayesian network framework supplemented with specific domain knowledge was applied in order to find relationships between drug combinations and target ADEs directly [25].
In addition, the gradient boosting algorithm XGBoost has been applied with success in making robust DDI predictions even for drugs whose interaction profiles that been seen at all during training [26]. We have demonstrated XGBoost yielded competitive or even better predictive power compared to support vector machine, random forest and classical gradient boosting in terms of both prediction accuracy as well speed [27, 26].
Consistency of the results across different methods has been reported to increase the accuracy of the prediction of DDIs compared to individual models [28, 29, 30, 18, 31, 32, 33] . For example, integrating ML algorithms LibLINEAR that SVM linear, Naïve Bayes and voting Perceptron classifiers as an example returned a higher F-score than the unbalanced training model of 70.4% against 69.0% [34]. Likewise, in a study to predict the unknown DDIs, an HNAI frame which comprises of five diversified algorithms; NB, DT, k-NN, LR, and SVM were developed. This framework obtained the AUC of 0. 67, surpassing the performance of individual algorithms (NB: In the results it has been found that DT has shown an accuracy of 0.565, k-NN has got an accuracy of 0. 6, LR has got 0. 655 and SVM 0. 666 while the average accuracy is 0.66 [6].
Some other ensemble methods that have been applied to the field include those that involve using a combination of GA and LR in classifier ensemble rules for DDI predictions; such methods have recorded an AUC of 1 and accuracy higher than 90 percent regardless of the approved or unapproved drug pairs [30]. Due to the fact that different drugs consist of multiple descriptors or features, the integration of these features into the model is key to predicting reliable results in DDI prediction systems. Due to this, Zhang et al. [35] developed a multi-modal deep auto- encoders based drug representation learning method (DDI-MDAE) that works for predicting DDIs in large-scale, noisy, and sparse data. DDI-MDAE is a positive-unlabeled

learning model where Deep Learning framework is expressed as a random forest (RF) classifer.

Also, computational experiments proposed a sparse feature learning ensemble method with lin- ear neighborhood regularization (SFLLN) to predict the DDIs, including novel DDIs. Even though the authors show that SFLLN is highly accurate and even surpasses benchmark methods in certain aspects, it must be pointed out that it took a very long time to run [36].

That is why over the past decades, the growing number of drugs has created interactions that are beyond the capacity of conventional ML techniques [37]. DL, in particular, the neural networks possess multiple processing levels and due to that the proposed model applies them for DDIs pre- diction [38]. Following the architecture of the human brain DL has outperformed the conventional methods for classification [39] and hence the increased utilization in the prediction of DDIs. While using ML techniques, feature extraction was often a separate process from feature learning; how- ever, in DL, these both procedures are united. This makes DL particularly suitable for solving the centrality, vagueness, and highly nonlinearity of the problems of predicting DDIs. Most types of DL can be considered as representation learning since the DL system, within a multi-layered consecutive architecture, learns how to build its own features. Since the introduction of DL in the current field, this section attempts to discuss the leading DL frameworks used in the extraction and prediction of DDIs.

Artificial Neural Network (ANN) is a computational model based on the biological neural network that helps to learn functional relationships in the data. Thus, in ANN a large number of neurons are linked to utilize linear and non-linear models to give an answer. Previous work has also used ANN models in the prediction of DDIs[40, 41]. For example, Rohani et al [42] used a two-layered neural network model to classify a feature set obtained from a number of similarity matrices of five different data sources. Masumshah et al. [43] used feed forward neural network with all the layers connected where ReLU was used as an activation between the layers while Sigmoid was used in the output layer. Furthermore Shtar et al to the graph nodes of DDI and propagation of ANN methods, an XGBoost classifier was employed with the matrix of adjacency calculating whether drug pairs have an interaction.

As mentioned before Recurrent Neural Networks (RNN) are widely meant for natural language processing (NLP)[44, 45] as well as for processing sequential data. RNNs are different from CNNs because of the former's memory mechanism through which it retains information from previous inputs to help in processing of current inputs and generation of current outputs. When it comes to the process of relation extraction for identifying DDIs, which is a type of relation extraction task in NLP, Long Short Term Memory (LSTM) networks have successfully been used to extract DDIs from literature [46, 47, 48]. Char-RNNs although are generally employed for the morphologically rich languages [49] and text classification [50] have also been adopted for the DDIs extraction. For instance, Kavuluru et al. [51] have proposed character level embeddings in the DDIs extraction and utilised LSTM on character embeddings to evacuate word vectors.

Luo et al. [52] proposed a model based on LSTM for the prediction of DDIs in diabetes regarding the embedded drug-induced transcriptome data. LSTM, described by Hochreiter and Schmidhuber in [53], is aimed to resolve the problem of long-term dependencies and have specific cells in the hidden layers, namely input, output and forget gates. Gated Recurrent Units or GRU which is a way of addressing the short term memory problems that are inherent in the conventional RNN model[54] works with a combination of states and gates, specifically reset and update gates which help in controlling the information that is stored for the use at the time of making predictions.

Regarding the DDIs extraction task, Zhang et al. [29] proposed the hierarchical RNN model with feature representation of the shortest dependency path (SDP) between the two entities. This model uses Recurrent Neural Networks to learn the features of the given consecutive sentences and SDP to extract the DDIs. It was separately proposed by Zhou et al. [55] to encode biomedical text sentences using an attention-based BiLSTM model. Also, Jiang et al. [56] employed a skeleton structure to define DDIs instances and utilized LSTM in processing this structure (skeleton-LSTM). In the context of their framework, the specific workload is divided into sentences, then into units, skeleton units, distances until the first and the second drug, respectively are introduced to the embedding layer of a skeleton-LSTM.

RNN or LSTM based architectures of Encoder-Decoder can run into the loss of information, especially in the situations when the sequence contains extended sentences. This problem is un-tangled by the attention mechanism as applied in this work [57]. Yi et al. [58] used a bidirectional RNN layer to generate the sentence matrix related to word semantics and an attention layer for fusing similar sentences of the same drug pairs into the final representation. A softmax classifier was then used to classify specific DDIs. Zheng et al. [59]

also came up with a model that incor- porates an attention mechanism with an RNN and LSTM units for the classification of the DDIs from texts.

This means that the improvement in the predictive superiority of AI tools is realized by enhancing the model's complexity that makes them opaque black box systems with unclear modes of func- tioning. Such a state of affairs can slow down the usage of AI models in such important spheres as health care. Therefore, the method called eXplainable Artificial Intelligence or XAI for short, has appeared to try to provide for the demand for explained predictions. Interpretability approaches are often classified in terms of the algorithm, the scale, and the type of data [60]. Moreover, the existing interpretability approaches can be divided into groups based on the objectives, which include making creation of white-box models, explaining black-box models, improving model's fairness, and carrying out the predictive sensitivity analysis [61].

As for the techniques designed to account for DL-based predictions, the gradient-based attri- bution method [62] tries to shed light on the network's input elements. This method is normally used on predictions generated by deep neural networks (DNN) and can be a prospective technique for some of the undisclosed DNN models used in DDIs prediction [63, 64]. Also, the DeepLIFT algorithm, which implies significant benefits as compared to other gradient-based methods, can be used in DNN models [65]. Lastly, the Guided BackPropagation method can also be used for network structures and what replaces the max–pooling layer in CNNs is a convolution layer with an improved step size to solve the problem of accuracy loss during the training process Thus, it has certain enlightenment in developing more effective CNN-based DDIs prediction models [66, 67].

In case of NLP based neural networks, one approach [68] suggested is the use of rationales (small portion of the document) that results in the same decision as the entire document input. This method consists of two stages: generator and encoder, which will help to discover the text subsets in direct relation to the result of the prediction. Since DDIs extraction tasks are carried out under the models of NLP [69, 70], it is possible to improve the clarity of these models.

Aside from the above methods, other methods recommended in XAI are trying to build the white-box models like linear models, decision trees, or rule based models; or building complex but interpretable models at the same time. However, these approaches are not powerful in terms of predicting the potential results and mostly concentrate on the few results, therefore they have less interest in the NLP-based domains like DDIs extraction. It has also been suggested that there are different approaches to the question of AI's fairness, and very limited literature has taken into account fairness under conditions of DDIs in non-tabular data such as text-based data. It has been found out that text vectorization, which is commonly employed in the setup of many DDIs studies with the use of word embedding methods [35, 69], can contain strong bias [71]. Hence, it would be understandable for fairness to be of increased concern in DDIs examinations.

It is equally important to develop the methods for sensitivity calculation to increase the relia- bility of AI models. The Adversarial Example-based Sensitivity Analysis utilised by Zugner et al.

[72] to investigate graph-structured data pertains to alteration of nodes' connection or attributes to target node classification models. Since graph-based methods are popular in DDIs researches [73, 74] it can be useful to apply the same to DDIs prediction models. Interference to word embed-dings [75] in RNNs should also be discussed; for input reduction method, the paper by Feng et al.

[76] that uncovers oversensitivity in NLP models can be utilized in the DDIs extraction research. Despite the fact that there is a vast number of publications about the deficiencies of DL models in NLP tasks, the use of DDIs-NLP models is still rather limited.

Schwarz et al [77] conducted study on the basis of DDIs and tried to improve the model interpretability with the help of Attention scores at all the layers of modeling. Thus, the impact of similarity matrices on drug representation vectors was revealed, as well as the features of drugs that contribute to their better encoding. This strategy uses all levels of the network since this structure offers an understanding on the workings of the model.

3. Proposed Methodology

In this section, the authors describe the proposed method for the prediction of drug-drug interac-tions (DDI) through a Data preprocessing, Data augmentation with GANs, machine learning & interpretability with SHAP. A flowchart of the proposed method can be depicted as follows.

The starting step is to load the required libraries and data set into the working environment. This step is crucial to make sure that all the necessary tools as well as data that will be needed for the further analysis are provided. After that, data preparation is initiated to deal with missing values and to encode categories as well as normalize the data. These preprocessing steps are very necessary to ready the data for feeding to the machine learning algorithms

so that the models can make proper interpretations of the data.

Since there is a scarcity of data, the current work applies a technique known as data augmen- tation using a Generative Adversarial Network or GAN. The GAN creates synthetic data that completes the training data along with the original data, in other words it creates a series of data that is diverse. Hence, this augmented dataset aids in the enhancement of both the reliability and efficiency of the machine learning model due to the increased variety of data provided.

The merged data set which is containing both, the real and the synthetic data is used to build a machine learning model. This model then uses the newly enhanced database to train and determine the patterns which are characteristic of potential DDIs. The trained model should give precise prediction results of DDIs with the help of the more diverse and expanded data.

Meanwhile, the issue of interpretability is solved through the usage of SHAP SHapley Additive exPlanations. SHAP values aid in understanding the outcome predicted by the machine learn- ing model by establishing which features had the most bearing on the model's decision-making processes. This interpretability is significant in various essential applications such as predicting patients' health state, understanding the model's decision-making is critical to trust and trans- parency.

In the following sub-sections, the logical flow of these research activities will be described in greater detail to give the reader a clearer picture of the overall proposed method.

3.1 Dataset Overview

We obtained a dataset from DrugBank to identify drug-drug interactions reported in the literature. The highly accurate and computationally efficient data supplied by DrugBank is presented in a broad range of scientific, clinical and unstructured sources. The database, containing the data on medicines, protein sequences and target sequences is being updated regularly with all these information under continuous verification by medical experts. These highly validated datasets are then used in many areas of research.

DrugBank database includes various types of data, which are indispensable for many tasks in biomedical science It displays in-depth data about clinical trials to reveal more information on drugs, including what is currently being trialed and repurposed as well as those that got stuck at the trial steps. Each record of a clinical trial is complete with date, ID number and descriptive title along eligibility criteria-a sine qua non for comparing results in evaluations of different drugs under study. This information is invaluable for the monitoring of drug development and the identification of areas where drugs can be re-purposed.

DrugBank has also a lot of information about drug interactions, crucial for the early iden- tification of possible adverse effects due to the interaction between medications. This property of the dataset is especially valuable for healthcare professionals and researchers investigating pa- tient safety, as well as treatment effectiveness. In addition, this dataset contains pharmacological analysis-required higher-level parameters like metabolism, pharmacokinetics and drug proteine interactions. These are vital in being able to place bets and carry out a deep risk analysis.

Data Description : The particular dataset utilized in this work is tabular and characterized with ID, Drug1, Drug2. Cell line,ZIP - Northing Classification etc Exactly, an example being that the dataset would include interactions such as "5-FU and ABT-888 in cell line A2058 with a ZIP score of 5.88 inferred synergy" or "5-FU and DASATINIB in cell line A2058 with aZIPscoreof -5.79 classifiedasantagonism". This structured data generates a solid substrate for meta-analyzing the drug-interactions and its response, hopes in developing predictive models to predict synergy or antagonistic nature of drugs.

3.2 Preprocessing

As the data preprocessing is a very important but boring part of preparing the dataset for both analysis and model training. For example, in the case of missing values, initially any rows with nan were dropped to maintain consistency. This was followed with the selection of a subset of data for further analysis. We further subset the total data by randomly sampling 40% of all cross- sectional image-context pairs to guarantee (1) sufficient diversity in measurement, while keeping computational load at a reasonable level for model training and evaluation; and also intended as well defined surrogate that can be used interchangeably.

The next step was encoding the 'classification' target variable to binary: 1 for synergy and 0 for antagonism. So this transformation was required so that the machine learning models can understand our label in a proper way. The features (X) was then separated out from the target variable (y), dropping 'classification' as part of features.

Categorical variables (Drug1, Drug2 and Cell line) were encoded with one-hot-encoding. This encoding converted that categorical data into the form which may be provided to algorithm to predict

better. Afterwards, the feature matrix was normalized with StandardScaler which does a z-score scaling therefore all features will contribute equally to training. This process of normalization centered the features to have a zero mean with unit variance.
Finally the type of feature matrix was converted to float for compatibility with subsequent machine learning algorithms. This all produced a clean, formatted and regularized data suited for the next steps in analysis as well as modeling.

3.3 GAN for Data Augmentation

The Generative Adversarial Network (GAN) is utilized in this study to augment the training dataset by generating synthetic data. The GAN framework consists of two primary components: the generator and the discriminator. The generator aims to produce realistic synthetic data, while the discriminator evaluates the authenticity of the data, distinguishing between real and synthetic samples.

3.4 Generator Model

The generator model in our GAN is designed to map a random noise vector **z** to the data space. This mapping is achieved through a series of linear transformations and nonlinear activation functions. The generator can be described by the following set of equations:

$$\boldsymbol{h}_1 = ReLU(\boldsymbol{W}_1\boldsymbol{z} + \boldsymbol{b}_1)$$

$$\boldsymbol{h}_2 = ReLU(\boldsymbol{W}_2\boldsymbol{h}_1 + \boldsymbol{b}_2)$$

$$\boldsymbol{h}_3 = ReLU(\boldsymbol{W}_3\boldsymbol{h}_2 + \boldsymbol{b}_3)$$

$$\boldsymbol{x}_{fake} = Tanh(\boldsymbol{W}_4\boldsymbol{h}_3 + \boldsymbol{b}_4)$$

The synthetic data generation process uses z as input noise vector along with hidden layers h1 h2 h3 and produces xfake. Each weight matrix Wi and bias term bi belongs to a layer within the neural network while ReLU and Tanh represent the activation functions.
The input noise vector z is drawn from a standard normal distribution, defined as:
$\mathbf{z} \sim N(0, 1)$

3.5 Data Flow Description

The image beneath displays the data movement inside GAN. The beginning of GAN execution starts with input tensors t1,t2,...tn+1 that contain training samples. The generator uses noise vector z together with input tensors to create synthetic data. The process of generator network data transformation works by processing each input tensor ti according to the following sequence:

$$t_i = Generator(\boldsymbol{z})$$

The variable i runs from number 1 until n + 1. The generator produces its output which joins real data to build an augmented training dataset.
The generator model generates realistic synthetic data by using input tensors together with random noise vectors. Using this method leads to larger training dataset diversity as well as higher quantity which results in better predictive model performance and resilience.

3.6 Discriminator Model

The discriminator in the GAN framework is designed to distinguish between real and synthetic data. It acts as a binary classifier, predicting the probability that a given input is real. The discriminator model processes the input data through a series of linear transformations and nonlinear activation functions to make its predictions.
The architecture of the discriminator model can be described by the following equations:

$$\boldsymbol{h}_1 = LeakyReLU(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1, \alpha = 0.2)$$

$$\boldsymbol{h}_2 = LeakyReLU(\boldsymbol{W}_2\boldsymbol{h}_1 + \boldsymbol{b}_2, \alpha = 0.2)$$

$$\boldsymbol{h}_3 = LeakyReLU(\boldsymbol{W}_3\boldsymbol{h}_2 + \boldsymbol{b}_3, \alpha = 0.2)$$

$$P_{real} = Sigmoid(\boldsymbol{W}_4\boldsymbol{h}_3 + \boldsymbol{b}_4)$$

In these equations:

- **x** is the input data (either real or synthetic).
- $\mathbf{h}_1$, $\mathbf{h}_2$, $\mathbf{h}_3$ are the hidden layers.
- P_{real} is the output probability that the input data is real.
- $\mathbf{W}_i$ and $\mathbf{b}_i$ represent the weight matrices and bias vectors of the respective layers.
- LeakyReLU is used as the activation function with a negative slope coefficient $\alpha = 0.2$.
- Sigmoid is used as the activation function for the output layer to ensure the output is a probability between 0 and 1.
-

4. Generate Synthetic Data

For augmentation of the training dataset a synthetic data was obtained by generation using the generated model. This was done by sampling the scalar from a standard normal distribution and then

generate new noise vectors by passing them through the generator. These synthetic samples were created to resemble the characteristics of the real data and as such, served to increase the size of the initial set of training data as well as its variety and solidity.

In figure 1, three histograms show the distributions of three different features from the real and generated data. The histograms of the Feature 0, Feature 1 and Feature 2 illustrate the distribution of the values within these features in both the train and test sets. In the case of Feature 0, the histogram distribution of real data and generated data is fairly close to each other, though slight deviations can be observed in the fact that the generated data peak near one. In Feature 1 and 2, the distributions of the generated appear less diverse compared to the real data, pointing to directions that may require the generator's improvement to mimic the real data distribution comprehensively.

By and large, these visualizations are useful for evaluating the fidelity of synthetic data and-end of –the synthetic data's capability to augment the real data in future machine learning tasks.

4.1 Combine Real and Synthetic Data

Thus, the real and synthetic data were merged to improve the training dataset. Firstly, the real training data in numpy array form was converted into DataFrame for convinience and to optimized the data handling. After that, the synthetic data generated in the previous step was vertically combined with the real data. This process gave cumulative augmented dataset which contains both real samples and synthetic samples.

The labels associated with the augmented data set were generated through a combined method; the real data set labels, and the synthetic data set labels were randomly generated. This way, it is possible to add the synthetic data into the training process without violating or altering the aspect of labeling. Finally, since the augmented data now was made of both the real and their respective synthetic samples and their labels, it was transformed back into a DataFrame. This DataFrame was kept in such format to hold the feature names and keep the structure of the combined dataset intact.

Using synthetic data combined with real data results in the augmentation of the original dataset reaching a larger and more complex pool of data for developing the machine learning algorithms. This dampened dataset is important to successfully increase the generality of the model and com-putation error on new data.

4.2 Machine Learning: XGBoost

In the machine learning phase, the above enlarged dataset was utilised to train an XGBoost model. The first transformations of data were conducted in the DMatrix format, which is suitable for working with XGBoost and saves resources. This format directs various functionalities including the feature names which were retained to enhance equal pattern in displaying data.

The values used in the model's parameterization are specifically chosen to better suit the binary classification problem. The 'objective' parameter was set to 'binary:'logistic' to ensure that the model chosen is binary logistic regression. For decision trees, the 'max_depth' parameter was set to 6, which restricts the number of levels or the depth of the trees, so as not to over-train the models. Eta, as is set at 0. 1, it decided the learning rate, the rate through which the boost process is carried and effectively preventing the over-reaction of the entire model update process. Also, the initialization of the regressor was given as XGBRegressor, with the 'eval_metric' parameter defined as 'logloss'; this involves the logistic loss function, thus offers a measure of the model's performance.

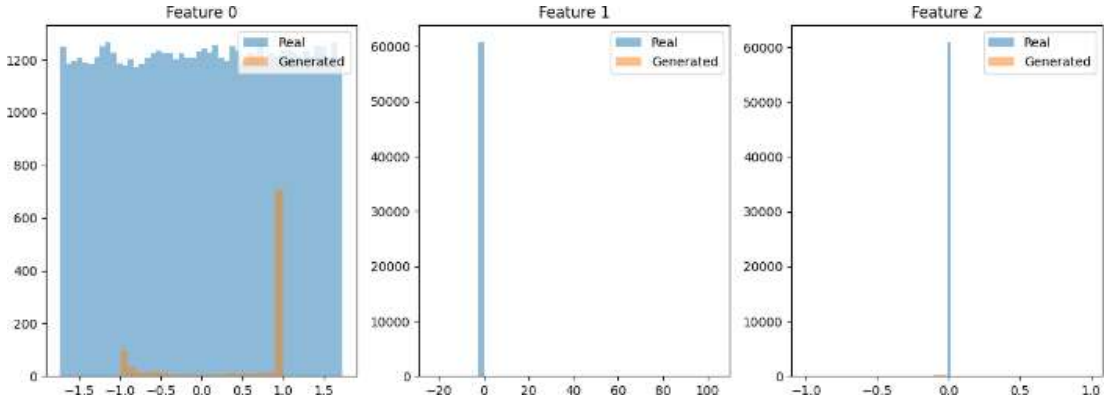


Figure 1. *Synthetic Data*

With regard to those parameters, the XGBoost model was trained with the number of boosting rounds set as 'num_boost_round' = 100. In order to avoid overfitting, the technique of early stopping was applied and 'early_stopping_rounds' was set to 10. This feature is to stop the training process when

the evaluation metric does not increase in the next 10 rounds. With respect to this, both the training as well as the validation datasets were used in this process in order to make it learn from the data that has been augmented while, at the same time checking on its efficacy on a different one.

This trained XGBoost model is set up with these parameters to make use of the diverse and abundant augmented data set which will make the model generalized and fine-tuned for drug-drug interaction prediction.

4.3 Explainable AI

SHapley Additive exPlanations (SHAP) served to make the XGBoost model's predictions more understandable. SHAP serves as a single approach for model prediction interpretation by identifying which input features affect model outputs. This approach reveals how individual features influence the final model prediction which stands as an advantage for explaining the model's determination processes in a simpler way.

A SHAP explainer was established to run on the XGBoost model after its training completion. The explainer produces SHAP values for each feature of the validation dataset by establishing their impact on the model predictions. The produced SHAP values served as inputs to create a summary plot which showed every feature's importance weight. The summary plot reveals which variables most substantially affect the model output and therefore demonstrates important predictors for model predictions.

XGBoost provided multiple feature importance evaluation tools that included the implementation of built-in plotting features alongside the SHAP summary plot. The assessment of feature importance happened through three metrics: weight, gain and cover. The weight metric reveals how often features appear in model trees while gain represents the model improvement from features and cover demonstrates the number of affected samples. The plots present feature importance data which complements the information obtained through SHAP analysis by showing different perspectives of the data.

The study generated SHAP dependence charts for the key supporting variables. The dependence plots reveal how each feature affects the output predictions by showing the impact of altering feature values on prediction results. The dependence plots enable researchers to visualize essential non-linear effects which develop between features inside predictive models.

The process of evaluating single data records involved generating SHAP waterfall summaries. These plots entirely explain how one prediction occurs by showing how each feature component influences the model output decision. Each case receives thorough interpretation through waterfall plots that demonstrate how the model makes progressive steps toward a decision.

The research utilizes SHAP analysis because it supplies XGBoost model predictions with both explanatory power and precision capability. The development of trust depends on Reality-Based Interpretability of model predictions since this functionality plays a crucial role in essential tasks like drug-drug interaction prediction operations. SHAP provides a model structure interpretation system that enhances human understanding of complex machine learning methods leading to wider practical implementation in industrial applications.

5. Experiment Results

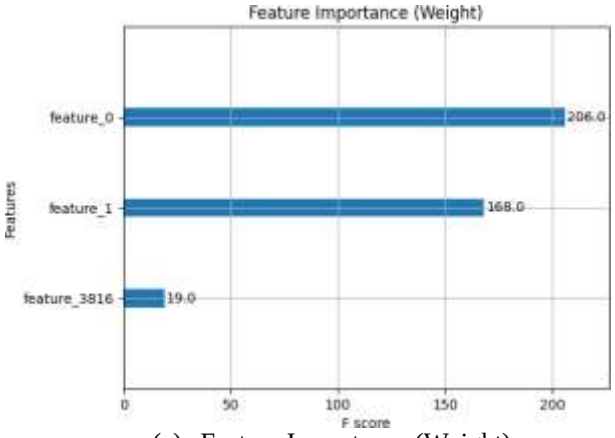
Testing the XGBoost model against the validation data set enabled a high level of performance measurement. The model achieved an outstanding validation accuracy of 0.9906 which indicates it properly classified close to 99% of validation examples. The F1 score as well as precision and recall values achieved high scores which demonstrate the model's capability to detect original and incorrect data correctly.

The model achieved precise accuracy which amounted to 98.73% successful positive prediction outcomes. The recall parameter measured at 0.9903 represented the successful identification rate of 99.03% for actual positive interactions through model prediction. The F1 score attained 0.9888 through the harmonic mean calculation between precision and recall levels.

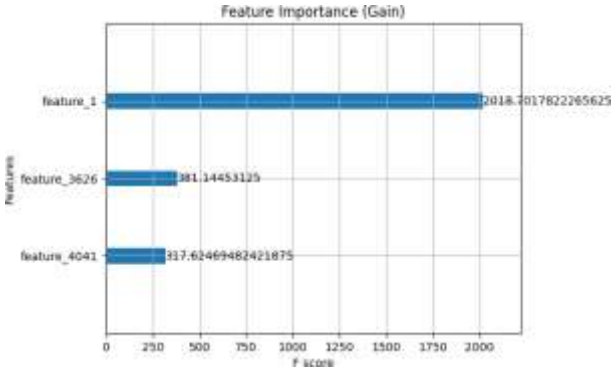
Through detailed reporting the classification process shows how the model performed between antagonism and synergy classes. The XGBoost model processed 7211 samples in the antagonism (class 0) with a precision of 0.99 and recall at 0.99 and F1 score at 0.99. виконання modelu na klasę 1 (synergy) otrzymoło dokładność 0.99 oraz ochronę 0.99 oraz skorzykodol 0.99 dla 5168 próbek w tej klasie.

Results show that the model performed exceptionally well according to macro and weighted average calculations of precision, recall, and F1 score at 0.99. The combination of real and synthetic data during XGBoost training results in significant improvement of drug-drug interaction prediction accuracy as demonstrated by these results.

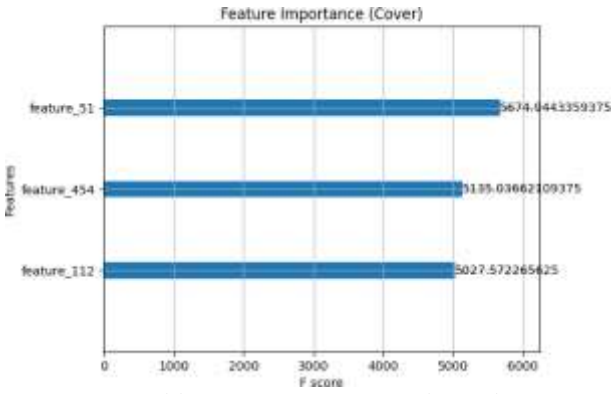
Figure 2. This figure consists of three subfigures (a), (b), and (c), which display different aspects of feature importance in the XGBoost model:



(a) Feature Importance (Weight)



(b) Feature Importance (Gain)



(c) Feature Importance (Cover)

Figure 2. *Feature Importance in XGBoost Model*

Figure 2 (a). The feature importance plot based on weight shows the number of times each feature is used to split the data across all trees in the model. Features with higher scores are used more frequently.

Figure 2 (b). The feature importance plot based on gain indicates the average gain of splits that include the feature, reflecting the feature's contribution to the model's performance improvement. Features with higher gain values have a greater impact on the model's predictive power.

Figure 2 (c). The feature importance plot based on cover shows the relative number of obser- vations related to each feature. Features with higher cover values affect more samples.

Figure 3. The SHAP bar plot shows the average impact of the top features on the model's output. Each bar represents the mean absolute SHAP value of a feature, indicating its overall contribution to the model's predictions. Feature 1 has the highest average impact, suggesting it is the most influential feature in the model.

Figure 3. *Average impact of the top features*

Table 1. *Experiment Results for XGBoost Model*

Precision		Recall	F1-Score	Support
Class 0 (Antagonism)	0.99	0.99	0.99	7211
Class 1 (Synergy)	0.99	0.99	0.99	5168
Accuracy			0.9906	
Macro Avg	0.99	0.99	0.99	12379
Weighted Avg	0.99	0.99	0.99	12379

Figure 4. The SHAP waterfall plot illustrates the contributions of individual features to a specific prediction. The plot breaks down the model's output (shown on the x-axis) into the contributions

of each feature. Negative values indicate that the feature pushes the prediction towards a lower output, while positive values push it towards a higher output. Feature 1 has the most significant negative impact on this particular prediction.

Figure 5. The SHAP decision plot shows the cumulative impact of features on the model's output across all samples. Each line represents a feature, and the x-axis shows the model's output value. Features that contribute positively to the prediction are in red, while those that contribute negatively are in blue. Feature 1 again stands out as having a major influence, significantly shifting the model's output value.

6. Conclusion

The research presents a full method for drug-drug interaction (DDI) prediction through advances in machine learning and AI explainable systems. The DrugBank dataset received substantial enhancement through our data preprocessing and augmentation pipeline which included GAN-based synthetic data generation. The performance of machine learning models reached high levels because the dataset became more diverse and robust through the augmentation process.

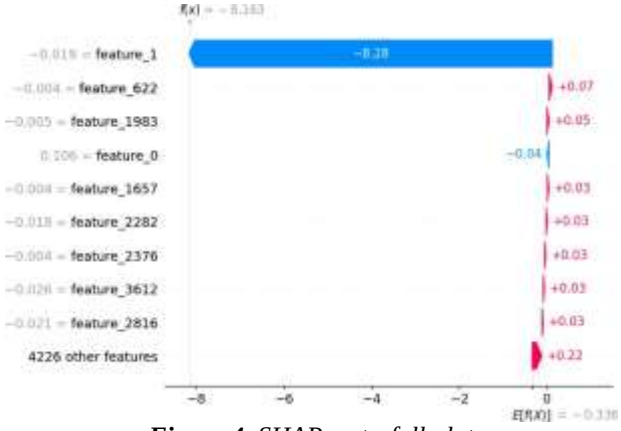


Figure 4. *SHAP waterfall plot*

Figure 5. *The SHAP decision plot*

Table 2. *Comparison of Different Methods for Drug-Drug Interaction Prediction*

Method	Feature Type	Performance Metric	Reference
Support Vector Machine (SVM)	Kernel Methods	F1-Score: 0.5965	[12, 13, 6, 14, 15, 16, 17]
All-paths Graph (APG) Kernel	Graph-based Features	F1-Score: 0.606	[18, 19]
Logistic Regression(LR)	Clinical Safety	Not specified	[23]
Drug-Entity-Topic (DET) Model	Text-mining Features	Improved discrimination and calibration	[24]
XGBoost	Ensemble Method	Accuracy: 99%, Precision: 98.73%, Recall: 99.03%, F1-Score: 0.9888	[27, 26]
Hybrid Meta-Heuristic Algorithms	Ensemble Method	F-score: 70.4%	[34]
Multi-modal Deep Auto-Encoders (DDI- MDAE)	Deep Learning	Suitable for large-scale, noisy, and sparse data	[35]
Sparse Feature Learning Ensemble Method (SFLLN)	Linear Neighborhood Regularization	High accuracy, surpasses benchmarks	[36]
Deep Learning (DL)	Neural Networks	Not specified	[37, 38, 39]
Artificial Neural Networks (ANN)	Neural Networks	Not specified	[40, 41, 42]
Recurrent Neural Networks (RNN) with LSTM	Sequential Data Processing	Used for relation extraction	[46, 47, 48]
Attention Mechanism in RNN	Text Classification	Enhances information retention in long sequences	[57, 58, 59]

Testing on the validation dataset demonstrated superior prediction aptitude of the XGBoost model which operated using the augmented dataset through its exceptional metric results including accuracy together with precision and recall and F1 score values. Detailed classification reports demonstrated the predictive capability of the model to accurately identify synergistic as well as antagonistic drug-drug interactions in critical application areas.

The research integrated SHapley Additive exPlanations (SHAP) for clear and complete interpretation of how the model makes its decisions. SHAP analysis provided essential information about which features are most important in model predictions as well as the way each feature impacts model predictions. Medical professionals should understand the reasoning mechanism behind predictions since this direct approach enhances trust and subsequent adoption of the system.

XAI stands out as a critical factor which helps explain complex machine learning models to human operators in clinical settings according to the study findings. This research made the model architecture accessible to users thereby clearing the path toward practical applications in clinical operations so healthcare providers could depend on these predictive systems.

Advanced machine learning combinations with explainable AI approaches create new possibilities for tackling intricate problems like drug interaction foretelling systems. The generated results support both effectiveness and practical implementation of the method for better medical care quality and treatment success in healthcare environments. Future research should concentrate on developing better prediction models together with examining new interpretability methods to advance the transparency and trustworthiness of healthcare AI predictions.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] Askari, M., et al. (2013). Frequency and nature of drug-drug interactions in the intensive care unit. *Pharmacoepidemiology and Drug Safety*. 22(4);430-437. https://doi.org/10.1002/pds.3415

[2] Raschetti, R., et al. (1999). Suspected adverse drug events requiring emergency department visits or hospital admissions. *European Journal of Clinical Pharmacology*. 54(12);959-963. https://doi.org/10.1007/s002280050582

[3] Budnitz, D.S., et al. (2006). National surveillance of emergency department visits for outpatient adverse drug events. *JAMA*. 296(15);1858-1866. https://doi.org/10.1001/jama.296.15.1858

[4] Reis, A.M., & Cassiani, S.H. (2010). Evaluation of three brands of drug interaction software for use in intensive care units. *Pharmacy World & Science*. 32(6);822-828. https://doi.org/10.1007/s11096-010-9445-2

[5] Vonbach, P., et al. (2008). Evaluation of frequently used drug interaction screening programs. *Pharmacy World & Science*. 30(4);367-374. https://doi.org/10.1007/s11096-008-9191-x

[6] Cheng, F., & Zhao, Z. (2014). Machine learning-based prediction of drug–drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. *Journal of the American Medical Informatics Association*. 21(e2);e278-e286. https://doi.org/10.1136/amiajnl-2013-002512

[7] Ryu, J.Y., Kim, H.U., & Lee, S.Y. (2018). Deep learning improves prediction of drug–drug and drug–food interactions. *Proceedings of the National Academy of Sciences*. 115(18);E4304. https://doi.org/10.1073/pnas.1803294115

[8] Vilar, S., et al. (2014). Similarity-based modeling in large-scale prediction of drug-drug interactions. *Nature Protocols*. 9(9);2147-2163. https://doi.org/10.1038/nprot.2014.151

[9] Vilar, S., Uriarte, E., Santana, L., Tatonetti, N.P., & Friedman, C. (2013). Detection of drug-drug interactions by modeling interaction profile fingerprints. *PLoS ONE*. 8(3);e58321. https://doi.org/10.1371/journal.pone.0058321

[10] Gunning, D., et al. (2019). XAI—explainable artificial intelligence. *Science Robotics*. 4(37);eaay7120. https://doi.org/10.1126/scirobotics.aay7120

[11] Hailu, N., Hunter, L., & Cohen, K.B. (2013). Ucolorado_som: extraction of drug-drug interactions from biomedical text using knowledge-rich and knowledge-poor features. *Second Joint Conference on Lexical and Computational*

Semantics (SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). https://aclanthology.org/S13-2112/

[12] Hunta, S., Aunsri, N., & Yooyativong, T. (2017). Integrated action crossing method for drug-drug interactions prediction in noncommunicable diseases based on neural networks. *2017 International Conference on Digital Arts, Media and Technology (ICDAMT)*. https://doi.org/10.1109/icdamt.2017.7904973

[13] Song, D., et al. (2019). Similarity-based machine learning support vector machine predictor of drug-drug interactions with improved accuracies. *Journal of Clinical Pharmacy and Therapeutics*. 44(2);268-275. https://doi.org/10.1111/jcpt.12786

[14] Wang, H., et al. (2021). Gognn: graph of graphs neural network for predicting structured entity interactions. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. Article 183. https://doi.org/10.24963/ijcai.2020/183

[15] Minard, A.-L., et al. (2011). Feature selection for drug-drug interaction detection using machine-learning based approaches. *Challenge Task on Drug-Drug Interaction Extraction (DDI) SEPLN*. 43–50. https://hal.science/hal-02289969v1

[16] Boyce, R., Gardner, G., & Harkema, H. (2012). Using natural language processing to identify pharmacokinetic drug-drug interactions described in drug package inserts. *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*. 206-213.

[17] Thomas, P., et al. (2011). Relation extraction for drug-drug interactions using ensemble learning. *1st Challenge task on Drug-Drug Interaction Extraction (DDIExtraction 2011)*. 11-18. https://ceur-ws.org/Vol-761/paper1.pdf

[18] Bokharaeian, B., Diaz, A., & Chitsaz, H. (2016). Enhancing extraction of drug-drug interaction from literature using neutral candidates, negation, and clause dependency. *PLoS ONE*. 11(10);e0163480. https://doi.org/10.1371/journal.pone.0163480

[19] Dhami, D.S., et al. (2018). Drug-drug interaction discovery: kernel learning from heterogeneous similarities. *Smart Health*. 9;88-100. https://doi.org/10.1016/j.smhl.2018.07.007

[20] Zhang, Y., et al. (2012). A single kernel-based approach to extract drug-drug interactions from biomedical literature. *PLoS ONE*. 7(11);e48901. https://doi.org/10.1371/journal.pone.0048901

[21] Zhang, P., Wang, F., Hu, J., & Sorrentino, R. (2015). Label propagation prediction of drug-drug interactions based on clinical side effects. *Scientific Reports*. 5;12339. https://doi.org/10.1038/srep12339

[22] Xie, W., et al. (2021). Integrated random negative sampling and uncertainty sampling in active learning improve clinical drug safety drug-drug interaction information retrieval. *Frontiers in Pharmacology*. 11;2225. https://doi.org/10.3389/fphar.2020.582470

[23] Yan, S., Jiang, X., & Chen, Y. (2013). Text mining driven drug-drug interaction detection. *Proceedings. IEEE International Conference on Bioinformatics and Biomedicine*. 349-355. https://doi.org/10.1109/bibm.2013.6732517

[24] Zhan, C., et al. (2020). Detecting high-quality signals of adverse drug-drug interactions from spontaneous reporting data. *Journal of Biomedical Informatics*. 112;103603. https://doi.org/10.1016/j.jbi.2020.103603

[25] Qian, S., Liang, S., & Yu, H. (2019). Leveraging genetic interactions for adverse drug-drug interaction prediction. *PLoS Computational Biology*. 15(5);e1007068. https://doi.org/10.1371/journal.pcbi.1007068

[26] Dang, L.H., et al. (2021). Machine learning-based prediction of drug-drug interactions for histamine antagonist using hybrid chemical features. *Cells*. 10(11);3092. https://doi.org/10.3390/cells10113092

[27] Park, C., Park, J., & Park, S. (2020). Agcn: Attention-based graph convolutional networks for drug-drug interaction extraction. *Expert Systems with Applications*. 159;113538. https://doi.org/10.1016/j.eswa.2020.113538

[28] Zhang, Y., et al. (2017). Drug–drug interaction extraction via hierarchical rnns on sequence and shortest dependency paths. *Bioinformatics*. 34(5);828-835. https://doi.org/10.1093/bioinformatics/btx659

[29] Mahadevan, A.A., et al. (2019). A predictive model for drug-drug interaction using a similarity measure. *2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. https://doi.org/10.1109/cibcb.2019.8791458

[30] Patrick, M.T., et al. (2021). Advancement in predicting interactions between drugs used to treat psoriasis and its comorbidities by integrating molecular and clinical resources. *Journal of the American Medical Informatics Association*. 28(6);1159-1167. https://doi.org/10.1093/jamia/ocaa335

[31] Zhang, Y., & Lu, Z. (2019). Exploring semi-supervised variational autoencoders for biomedical relation extraction. *Methods*. 166. https://doi.org/10.1016/j.ymeth.2019.02.021

[32] Hung, T.N.K., et al. (2022). An ai-based prediction model for drug-drug interactions in osteoporosis and paget's diseases from smiles. *Molecular Informatics*. 2100264. https://doi.org/10.1002/minf.202100264

[33] Bobic, T., Fluck, J., & Hofmann, M. (2013). Scai: Extracting drug-drug interactions using a rich feature vector. *Second Joint Conference on Lexical and Computational Semantics (SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013).* https://aclanthology.org/S13-2111/

[34] Zhang, Y., et al. (2020). Predicting drug-drug interactions using multi-modal deep auto-encoders based network embedding and positive-unlabeled learning. *Methods*. 179;37-46. https://doi.org/10.1016/j.ymeth.2020.05.007

[35] Zhang, W., et al. (2019). Sflln: A sparse feature learning ensemble method with linear neighborhood regularization for predicting drug–drug interactions.

Information Sciences. 497;189-201. https://doi.org/10.1016/j.ins.2019.05.017

[36] Li, D., et al. (2016). A topic-modeling based framework for drug-drug interaction classification from biomedical text. *AMIA Annual Symposium Proceedings*. 789-798. https://pmc.ncbi.nlm.nih.gov/articles/PMC5333320/

[37] Shukla, P.K., et al. (2020). Efficient prediction of drug-drug interaction using deep learning models. *IET Systems Biology*. 14(4);211-216. https://doi.org/10.1049/iet-syb.2019.0116

[38] Sejnowski, T.J. (2020). The unreasonable effectiveness of deep learning in artificial intelligence. *Proceedings of the National Academy of Sciences*. 117(48);30033-30038. https://doi.org/10.1073/pnas.1907373117

[39] Hou, W.J., & Ceesay, B. (2018). Extraction of drug-drug interaction using neural embedding. *Journal of Bioinformatics and Computational Biology*. 16(6);1840027. https://doi.org/10.1142/s0219720018400279

[40] Shtar, G., Rokach, L., & Shapira, B. (2019). Detecting drug-drug interactions using artificial neural networks and classic graph similarity measures. *PLoS ONE*. 14(8);e0219796. https://doi.org/10.1371/journal.pone.0219796

[41] Rohani, N., & Eslahchi, C. (2019). Drug-drug interaction predicting by neural network using integrated similarity. *Scientific Reports*. 9(1);13645. https://doi.org/10.1038/s41598-019-50121-3

[42] Masumshah, R., Aghdam, R., & Eslahchi, C. (2021). A neural network-based method for polypharmacy side effects prediction. *BMC Bioinformatics*. 22(1);385. https://doi.org/10.1186/s12859-021-04298-y

[43] Collobert, R., et al. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*. 12;2493-2537. https://doi.org/10.48550/arXiv.1103.0398

[44] Sutskever, I., Vinyals, O., & Le, Q.V. (2014). Sequence to sequence learning with neural networks. *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. 3104-3112. https://doi.org/10.48550/arXiv.1409.3215

[45] Zhang, S., et al. (2015). Bidirectional long short-term memory networks for relation classification. *PACLIC*. https://aclanthology.org/Y15-1009/

[46] Sahu, S.K., & Anand, A. (2018). Drug-drug interaction extraction from biomedical texts using long short-term memory network. *Journal of Biomedical Informatics*. 86;15-24. https://doi.org/10.1016/j.jbi.2018.08.005

[47] Wang, W., et al. (2017). Dependency-based long short term memory network for drug-drug interaction extraction. *BMC Bioinformatics*. 18(16);578. https://doi.org/10.1186/s12859-017-1962-8

[48] Kim, Y., et al. (2016). Character-aware neural language models. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. 2741-2749. https://doi.org/10.1609/aaai.v30i1.10362

[49] Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*. 649-657. https://doi.org/10.48550/arXiv.1509.01626

[50] Kavuluru, R., Rios, A., & Tran, T. (2017). Extracting drug-drug interactions with word and character-level recurrent neural networks. *IEEE International Conference on Healthcare Informatics*. https://doi.org/10.1109/ichi.2017.15

[51] Luo, Q., et al. (2021). Novel deep learning-based transcriptome data analysis for drug-drug interaction prediction with an application in diabetes. *BMC Bioinformatics*. 22(1);318. https://doi.org/10.1186/s12859-021-04241-1

[52] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*. 9(8);1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[53] Cho, K., et al. (2014). On the properties of neural machine translation: Encoder–decoder approaches. *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. https://doi.org/10.3115/v1/w14-4012

[54] Zhou, D., Miao, L., & He, Y. (2018). Position-aware deep multi-task learning for drug-drug interaction extraction. *Artificial Intelligence in Medicine*. 87;1-8. https://doi.org/10.1016/j.artmed.2018.03.001

[55] Jiang, Z., Gu, L., & Jiang, Q. (2017). Drug drug interaction extraction from literature using a skeleton long short term memory neural network. *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. https://doi.org/10.1109/bibm.2017.8217708

[56] Zaikis, D., & Vlahavas, I. (2020). Drug-drug interaction classification using attention based neural networks. *11th Hellenic Conference on Artificial Intelligence*. 34-40. https://doi.org/10.1145/3411408.3411461

[57] Yi, Z., et al. (2017). Drug-drug interaction extraction via recurrent neural network with multiple attention layers. *International Conference on Advanced Data Mining and Applications*. 448-462. https://doi.org/10.1007/978-3-319-69179-4_39

[58] Zheng, W., et al. (2017). An attention-based effective neural model for drug-drug interactions extraction. *BMC Bioinformatics*. 18(1);445. https://doi.org/10.1186/s12859-017-1855-x

[59] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*. 2169-3536. https://doi.org/10.1109/access.2018.2870052

[60] Guidotti, R., et al. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*. 51(5), 93. https://doi.org/10.1145/3236009

[61] Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*. abs/1312.6034.

[62] Liu, S., et al. (2016). Dependency-based convolutional neural network for drug-drug interaction extraction. *2016 IEEE International*

Conference on Bioinformatics and Biomedicine (BIBM). https://doi.org/10.1109/bibm.2016.7822671

[63] Sun, X., et al. (2018). Deep convolution neural networks for drug-drug interaction extraction. *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. https://doi.org/10.1109/bibm.2018.8621405

[64] Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. 3145-3153. https://doi.org/10.1109/tnnls.2022.3228102

[65] Zeng, T., et al. (2015). Deep convolutional neural networks for annotating gene expression patterns in the mouse brain. *BMC Bioinformatics*. 16(1);147. https://doi.org/10.1186/s12859-015-0553-9

[66] Springenberg, J.T., et al. (2015). Striving for simplicity: The all convolutional net. *CoRR*. abs/1412.6806.

[67] Lei, T., Barzilay, R., & Jaakkola, T. (2016). Rationalizing neural predictions. *2016 Conference on Empirical Methods in Natural Language Processing*. 2143-2152. https://doi.org/10.18653/v1/d16-1011

[68] Quan, C., et al. (2016). Multichannel convolutional neural network for biological relation extraction. *BioMed Research International*. 2016;1850404. https://doi.org/10.1155/2016/1850404

[69] Xiong, W., et al. (2019). Extracting drug-drug interactions with a dependency-based graph convolution neural network. *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. https://doi.org/10.1109/bibm47256.2019.8983150

[70] Bolukbasi, T., et al. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in Neural Information Processing Systems*. 29. https://doi.org/10.48550/arXiv.1607.06520

[71] Zügner, D., Akbarnejad, A., & Günnemann, S. (2018). Adversarial attacks on neural networks for graph data. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*. 2847-2856. https://doi.org/10.1145/3219819.3220078

[72] Sun, M., Wang, F., Elemento, O., & Zhou, J. (2020). Structure-based drug-drug interaction detection via expressive graph convolutional networks and deep sets. *Proceedings of the AAAI Conference on Artificial Intelligence*. https://doi.org/10.1609/aaai.v34i10.7236

[73] Lin, Z.Q.X., et al. (2020). Knowledge graph neural network for drug-drug interaction prediction. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. https://doi.org/10.24963/ijcai.2020/380

[74] Miyato, T., Dai, A.M., & Goodfellow, I. (2016). Adversarial training methods for semi-supervised text classification. arXiv preprint arXiv:1605.07725.

[75] Feng, S., et al. (2018). Pathologies of neural models make interpretations difficult. *arXiv preprint*. https://doi.org/10.48550/arXiv.1605.07725

[76] Schwarz, K., et al. (2021). Attentionddi: Siamese attention-based deep learning method for drug–drug interaction predictions. *BMC Bioinformatics*. 22(1);412. https://doi.org/10.1186/s12859-021-04325-y