The Multiple Approaches for Drug-Drug Interaction Extraction using Machine learning and transformer-based Model

Gurpreet Singh[1], Dr. Kim Yong Il [2], Renuka Ramasamy [3], Sundareswari Thiyagarajan[4]

[1]Endicott College of International Studies, Woosong University, South Korea

[2]Endicott College of International Studies, Woosong University, South Korea

[3]Renuka Ramasamy, University college of Engineering, Anna University, Tiruchirappalli, India

[4]Sundareswari Thiyagarajan, University college of Engineering, Anna University, Tiruchirappalli, India

[*]Corresponding Author: Author's Name. Gurpreet Singh; Email: gurpreetsinghmcse@gmail.com

Abstract: This research paper investigates a machine learning based approach for Drug-Drug Interaction (DDI) extraction for determining the side effects of multi drugs when prescribed simultaneously. In our proposed model we used TAC 2017 Dataset, which has Adverse Drug Reactions (ADRs) data for the classification of drug-drug interaction. TAC 2017 Dataset has various types of information which are related to drugs and their interactions. Our method uses Term Frequency-Inverse Document Frequency (TF-IDF) to transform the textual descriptions of side effects for DDI into numerical feature vectors, followed by a Random Forest Classifier, Gradient Boosting, BioBERT, Support Vector Machine (SVM) Algorithms to predict the potential interactions between drug pairs. One of the key strength of the Random Forest approach is its ability to provide feature importance scores, which allows us to interpret which side effects are most influential in predicting drug interactions. The key advantage of Gradient Boosting is its high predictive performance combined with interpretability. It is able to handle complex, structured data efficiently. Additionally, the model's decisions are more transparent, which is necessary in the biomedical domain. The advantage of SVM is its ability to handle high-dimensional data, capture complex non-linear interactions using kernel functions, and generalize with datasets, making it robust to over-fitting. BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) is advantageous for DDI prediction due to its biomedical domain knowledge, contextual understanding of complex drug-related texts.These method captures the relevance and importance of each side effect of multi drugs and also generate pairs of drugs from the dataset. Our model demonstrates competitive performance in DDI prediction, which highlights the utility of text-based feature extraction combined with an interpretable ensemble learning model.

Keywords: Drug-Drug Interaction, Adverse Drug Reactions, Random Forest Classifier, Gradient Boosting Algorithm, Support vector Machine, BioBERT, Side Effects.

1 Introduction

Drug-Drug Interaction (DDI) refers to effect of consuming two or more drugs simultaneously. There is a study on DDI pairs in vivo and in vitro researches. Laboratory and practical researches can give more accurate results but on the other hand, it was expensive and also has limited resources [12]. Drug-drug interactions (DDI) plays a vital role in the development of new drugs and also for (PhV) pharmacovigilance which contributes 6 - 30% of all Adverse Drug Events [13]. There are many methods proposed for DDI

extraction that are classified into two categories such as rule-based [16] and machine learning-based methods [17], [19]. In comparison between better performance and portability Machine learning models outperform the rule-based models [18]. BERT [43] is a recently proposed pre-trained transformer based model. Due to its multi-layer bidirectional transformer [44] structure, BERT can process the contextual information of sentences into the word vector both from forward and backwards. Classification-based methods formulate DDI prediction as a binary classification task which is represented drug-drug pairs as feature vectors from their constructed pharma cointeraction network [45]. Then they defined the presence or absence of interactions as labels and finally built the logistic regression models for predictions.

In this study, we propose a method using Term Frequency-Inverse Document Frequency (TF-IDF) vectorization combined with Random Forest Classifier, SVM, Gradient Boosting, while BioBERT combined with neural network to predict DDIs. By training the Random Forest model on the feature vectors, we focus to predict whether a given pair of drugs will interact or not. A binary classification label is assigned based on whether the drugs share common side effects or not, indicating a potential interaction. Each model offers distinct advantages,traditional Machine Learning methods are efficient for structured data, while BioBERT leverages deep learning to analyze the unstructured text.

2 Related Work:

From between the drug pairs identifying the drug entities and their interaction is one of the challenging tasks for Drug-Drug Interaction (DDI) Extraction [20]. For the Drug Discovery Computer-Aided Drug-Drug Interaction (DDI) Extraction plays a vital role, because of its process is too expensive and also an time-consuming. So that, in the drug development cycle Machine Learning based approaches can reduce the laborious task. Here the author proposed an approach namely TP-DDI (TP-DDI: Transformer-based pipeline for the extraction of Drug-Drug Interactions) using an Pre-trained weights called BioBERT to improve both drug name entity recognition and DDI extraction task on the DDI Extraction 2013 dataset. In [14], Drug-Drug Interaction (DDI) Extraction is one of the Natural Language Processing task. The author proposed an CNN (Convolutional Neural Network) based model DDI Extraction using DDI 2013 Corpus dataset, their proposed model achieved an F-score of 69.75%.In [21], Preventing adverse effects from drug combinations can be achieve by detecting drug-drug interactions (DDI). The author proposed an Recursive Neural Network to improve the performance of DDI extraction using DDI 2013 dataset. Their proposed model comprises a position feature, a subtree containment feature, and an ensemble method to improve the performance of DDI extraction. They compared with the state-of-the-art models, the DDI detection and type classifiers of their model performed 4.4% and 2.8% better than the existing model. In [22] DDI extraction means extracting information from biomedical text it is a promising approach it gives the better insights to know about the effect of one drug with another drug. However the available of large data makes it more complicated to understand Drug-Drug Interaction and its side effects. Here the author proposed an text mining method for extracting automatically DDI information from text. They combined feature embedding with deep learning methods, and also they achieved improving results in learning the text features by using an LSTM (Long -short term memory) neural network. In [23] Here the author proposes a novel solution to DDI Extraction by using Relation BioBERT (R-BioBERT) to detect and classify Drug-Drug Interaction with the Bidirectional Long Short-Term Memory (BLSTM) for improving the accuracy of predictions for the Drug-Drug Interaction. Their proposed model also can able to identifies the specific type of interaction between the two drug interaction. Their proposed model shows that the use of BLSTM yields higher F-scores compared to their baseline model using the three well-known dataset such as SemEval 2013, TAC 2018, and TAC 2019. In [24], the healthcare system, medical errors is one of the growing concerns because of its adverse events. Drug-Drug Interaction can able to prevent adverse events, by extracting Drug-Drug Interactions from drug labels into a machine-processable form, its an effective step towards an drug safety information in the health care system. They proposed an graph convolutions (GCs) with a novel attention-based gating mechanism (GCA) for predicting task using TAC 2018 DDI Extraction corpus.In [25], In this paper the authors proposed three deep learning models for the classification and identification method for Drug-Drug Interaction such as CNN (Convolutional Neural Network), BiLSTM (Bi-directional Long-Short term Memory), and BiLSTM with attention. They also

introduce three transformer-based models namely BERT, RoBERTa, and ELECTRA. To improve the DDI extraction they proposed an ensemble approach by combining these transformer models using DDI Extraction 2013 dataset. And also their attention-based model gives better yield compared to their existing model.
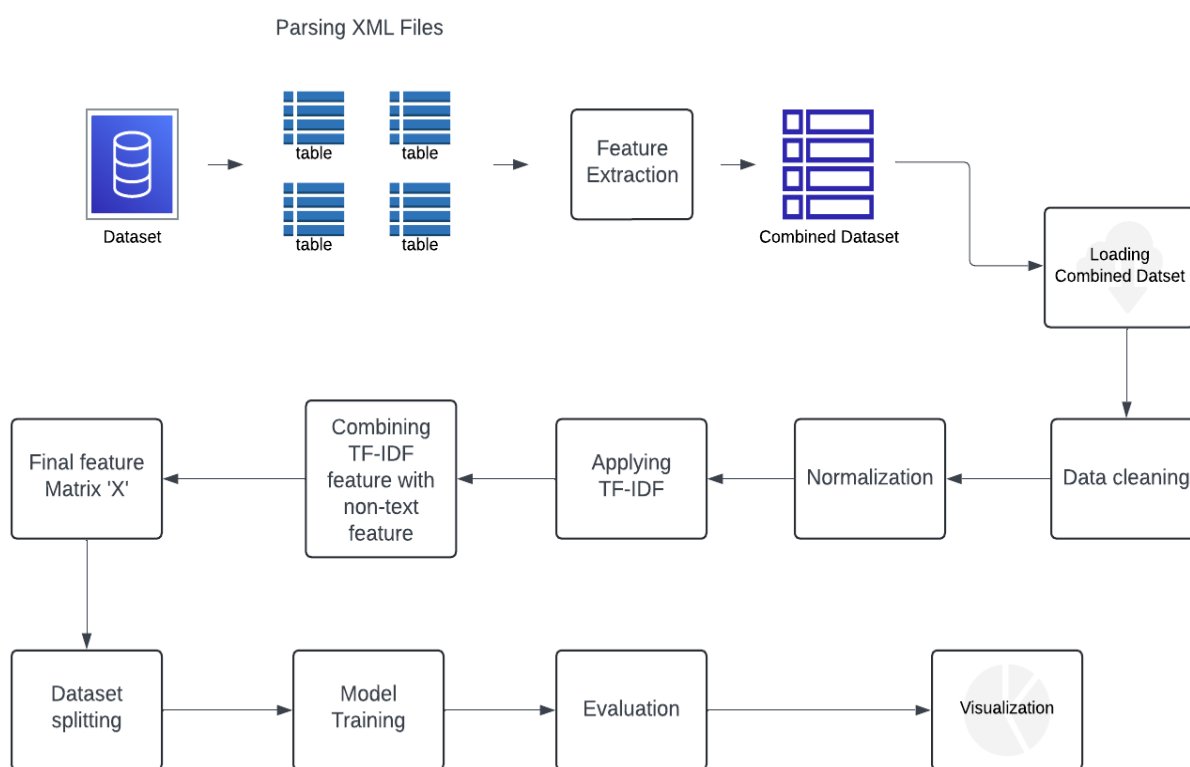
3 Methodology

Figure 1: Architecture Diagram for Proposed Model

3.1 Dataset Description:

This TAC 2017 Adverse Drug Reaction dataset is one of the part of the Text Analysis Conference (TAC) Adverse Drug Reaction Extraction from Drug Labels(Drug prescription) which describe how to use, who should not use, who should use, the reaction from the particular medicine and some specific safety concerns. This dataset is primarily focused on extracting and normalizing adverse drug reactions (ADRs) from structured product labels (SPLs) of the US Food and Drug Administration (FDA) drugs. This dataset includes annotations for ADR(Adverse Drug Reaction), the relation between ADR such as hypothetical, effect, and negated, and their reactions. These Adverse Drug Reactions are then normalized according to the Medical Dictionary in terms of MedDRA terminology, to provide consistency and facilitation for further analysis. These datasets are composed of SPLs (structured product labels), which are one of the official documents, that describe the drug safety, effect, and usage of drugs.

For our proposed model we used TAC 2017 Dataset for the Multi-class classification of drug-drug interaction, TAC 2017 Dataset has various types of information which are related to drugs and their interactions. The dataset consists of Textual Features and Integer Features. Textual Features such as Drug,

Arg1, Arg2, Reaction_Str, Normalization_id, Meddra_PT, Meddra_LLT, which contains the drug information , their relation and their reactions. Integers Features such as Meddra_PT_ID, Meddra_LLT_ID these are numerical codes along with medical terms, it is basically used for categorization and analysis.
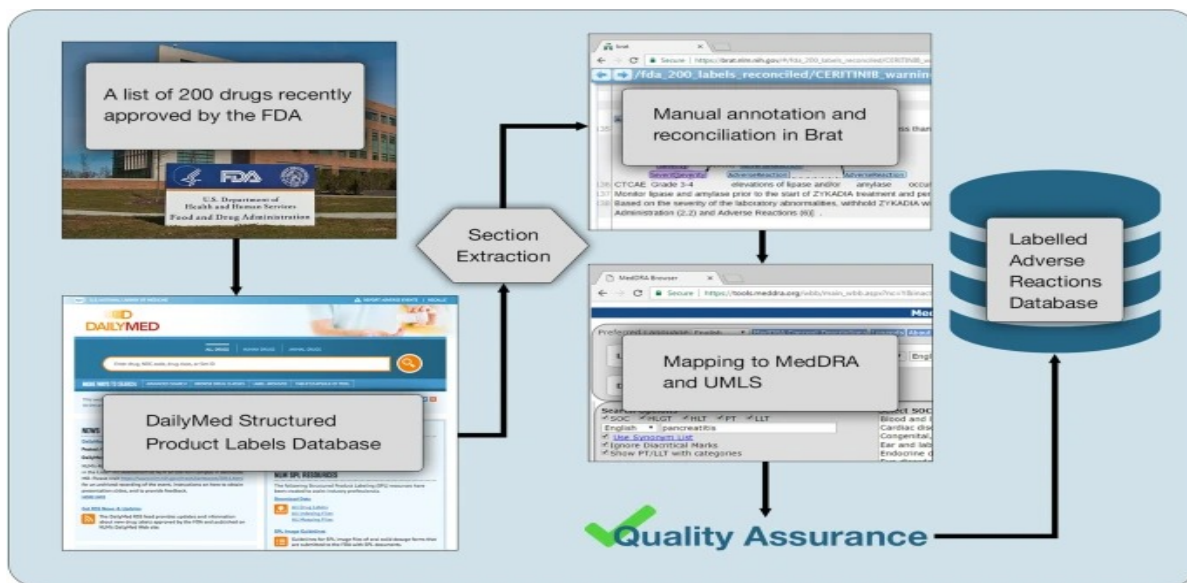

Figure 2: Representation of ADR Database [41]

3.2 Data Preprocessing:

The textual features which was processed by TF-IDF were combined with the integer features into a single feature matrix. By doing this the model can able to learn the data from both textual and numerical data. We combined the dataset which comprises various Drugs and their relation in a separate folder. In each and every folder, it contains details about each drug's properties, such as its name, relation, and reaction or interaction properties, which contains information about the drugs interactions, including whether the interaction is hypothetical, effectual, or negated with annotated fields along with Meddra Codes, The MedDRA (Medical Dictionary for Regulatory Activities) codes which is refers to standardized medical terms for adverse events and the medical conditions, which are used to describe the interactions between the drugs.
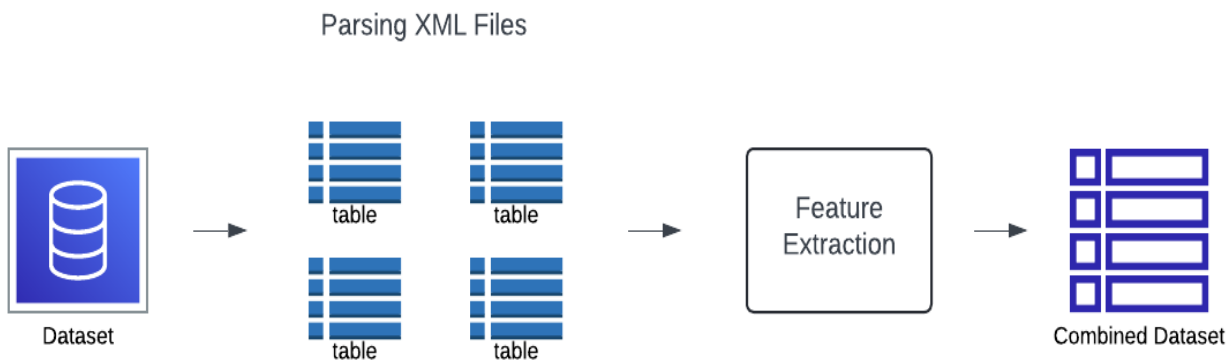

Figure 3: Visual representation of Data preprocessing

3.4 Combination of data:

Textual columns in this dataset were combined into a single column. This combined column comprises the drug and its side effects. Empty strings were placed in place of missing values to ensure valid text representation for each and every data.

3.5 Feature Extraction:

These text data were converted into numerical features applying TF-IDF (Term Frequency-Inverse Document Frequency). By using this technique it converts the text data in the form of a matrix of TF-IDF vectorization, these matrix represent the importance of each word present in the data. To control the complexity of model computational and its dimension we limited the maximum number of features into 2000 and also we removed English stop word to improve the focus on the main or meaningful word.

For our model, we used TF-IDF (Term Frequency-Inverse Document Frequency), which is one of the methods in Natural Language Processing (NLP) used for converting textual data into numerical features. TF means Term Frequency it is used to measure the frequent words in files. IDF means Inverse Document Frequency which is used to measure important words in the entire dataset by decrementing the weight of normal words in files. It will give the result in the matrix form where each row denotes the files about the drug-drug interaction while each column denotes a word, which is weighted by its importance in files. TF-IDF stands for Term Frequency-Inverse Document Frequency. It is vastly used in information retrieval, NLP, text classification, keyword extraction [2-3].

Term Frequency (TF) is used to count how frequently a text comes in a document. Term is represented as 't' and Document as 'd'.

$$TF(t,d) = \frac{Number\ of\ times\ t\ appears\ in\ a\ document\ d}{Total\ number\ of\ terms\ in\ document\ d}$$

IDF measures the importance of a particular term in the entire document. It is used in reducing the common terms ("are", "is","It", etc.) which occurs frequently. |D| is the Total number of documents in the corpus. Number of documents containing term t refers to the count of documents in which the term t appears.

$$IDF(t,D) = log\left(\frac{Total\ number\ of\ document\ |D|}{Number\ of\ Documents\ containing\ term\ t} + 1\right)$$

TF-IDF(t,d,D) = TF(t,d) x IDF(t,D)

TF-IDF Score is calculated by multiplying the TF and IDF. High TF-IDF Score refers to the good keyword for the document and low TF-IDF Score refers to the less informative term.

3.6 Normalization:

After TF-IDF vectorization these features were normalized using L2 normalization. By using normalization it helps in balancing the instances of unique features during the model training.

3.7 Data Splitting:

By using an 80/20 split we split the dataset into training and test sets, training set for model training, while the testing set for evaluating the model performance.

4 Evaluation Metrices:

True positive (TP), False Positive (FP), True Negative (TN), False Negative (FN) are the important factors in Evaluation of Accuracy, Precision, Recall, F1-score. True positive (TP) refers to the model correctly identifies the interaction from the dataset. False Positive (FP) refers to the model incorrectly identifies no interaction from the dataset. True Negative (TN) refers to the model correctly identifies no interaction from the dataset. False Negative (FN) refers to the model incorrectly identifies interaction from the dataset.

4.1 Accuracy:

It shows the overall performance of our model by correctly classifying instances.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$ (1)

4.2 Precision:

Precision is also known as Positive Predictive Value. It is the ratio of correctly predicted positive value to the total predicted positive value. It is effective only the cost of false positive value is high.

$$Precision = \frac{TP}{TP + FP}$$

(2)

4.3 Recall:

Recall is also known as True Positive Rate. Recall refers to the ratio of correctly predicted positive cases to the total number of positive instances

$$Recall = \frac{TP}{TP + FN}$$ (3)

4.4 F1-Score:

It shows the metrics between precision and recall. It is most useful when we are dealing with imbalanced datasets for training the model for multi-class classification.

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$ (4)

4.5 ROC along with AUC:

AUC (Area Under the Curve) of the ROC (Receiver Operating Characteristic) curve, which measures the model's ability to differentiate between the classes across all possible classification thresholds.t shows the difference between the True Positive Rate and False Positive Rate for the threshold values.

$$True\ Positive\ Rate = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

(5)

$$False\ Positive\ Rate = \frac{False\ Positive}{False\ Positive + True\ Negative}$$

(6)

5 Model Training and Evaluation:

5.1 Random Forest Algorithm and its Performance:

For our proposed model we used Random Forest Classifier to perform the classification tasks. Random forest classifier is an ensemble learning method, it combines multiple decision trees to improve performance and reduce overfitting. Using Randomized Search CV we perform tuning for the hyperparameters for model training. Random forest is an ensemble of decision trees that has multiple processes like bagging, random feature selection [5-8]. It is combination of the decision trees for the purpose of increasing the accuracy, reducing the over-fitting. It does not have a structure like single closed-form formula. The final classification or prediction is made based on the majority of the voting among all the decision trees. Most of the decision is based on the bagging and random feature selection, which will helps in reducing the variance and generalization.

5.1.1 Bagging (Bootstrap Sampling):

Bagging is done by creating multiple datasets by randomly sampling the original with replacement. Each tree in Random Forest Algorithm is trained on a different subset of the data.

$$D_b = RandomSample(D, N)$$

Where,
N : Total number of data points
D_b: Bootstraped sample for tree b
D : The original dataset

5.1.2 Random Feature Selection:

For every split in the decision tree, a random subset is considered for performing best split. 'p' represents the total number of features. 'm' represents the number of features selected at each split.

$$m = \sqrt{p} \ \textit{(for classification)}$$

Where,
p - total number of features
m - number of features selected at each split.

5.1.3 Ensemble Voting for Prediction (Majority Voting):

The final prediction is made by aggregation prediction from all the decision trees.

$$\hat{y} = mode(h_1(x), h_2(x), \ldots, h_T(x))$$

Where,
ŷ - Overall Prediction
$h_i(x)$ - Class prediction

Some of the key hyper-parameters tuning include the number of estimators, maximum depth, minimum samples split, and minimum samples leaf. After model training, it visualizes the performance evaluation metrics.
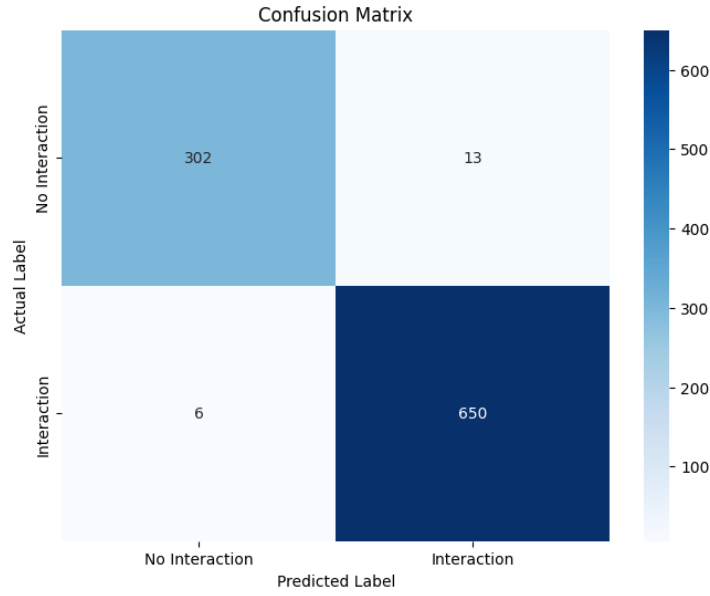
Figure 4: Confusion matrix for Random Forest

True Positive (TP) = 650, False Positive (FP) = 13, True Negative(TN) = 302, False Negative (FN) = 6

From Eq (1) $Accuracy = \frac{650+302}{650+13+302+6} = 0.9804$

From Eq (2) $Precision = \frac{650}{650+13} = 0.9804$

From Eq (3) $Recall = \frac{650}{650+6} = 0.9908$

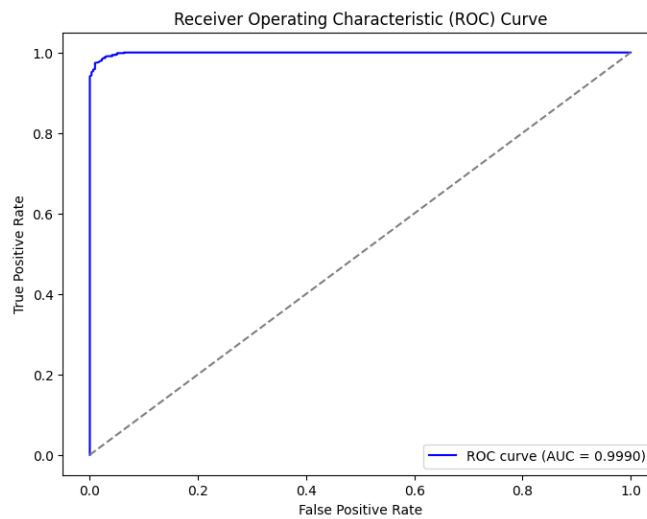From Eq (4) $F1\ score = 2 \times \frac{0.9804 \times 0.9908}{0.9804+0.9908} = 0.9856$

Figure 5: ROC curve for Random Forest Model

Here Fig. 5 shows that ROC curve along with AUC (Area under curve), AUC shows the difference between positive and negative class, if the AUC is pointing to 1 that means the model performance is good, here our model ROC diagram, the AUC is 1.00 that means it clearly shows that our is model performance is good at differentiating positive and negative classes.

Bar Graph: It is a visual comparison of the actual class distribution in the test set and the predicted set, it will help us to diagnose the model's predictions for any biases.
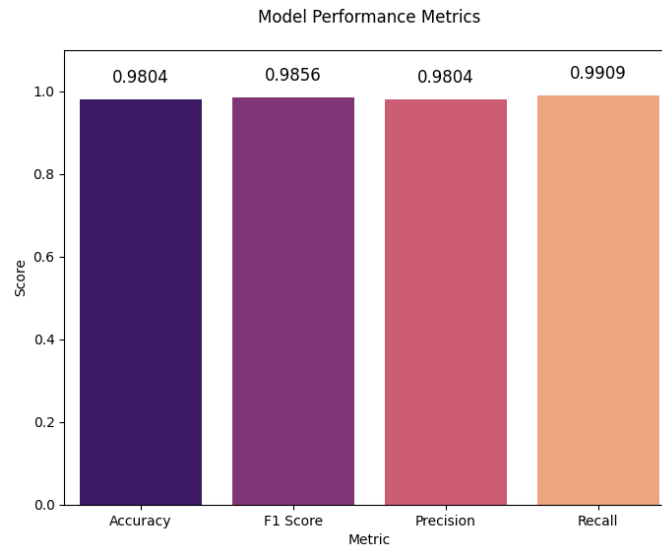

Figure 6: Bar graph for Random forest model

Table 1: Classification Report for Random forest Model

Classification Report			
	precision	recall	f1–score	support
0	0.98	0.96	0.97	315
1	0.98	0.99	0.99	656
accuracy			0.98	971
Macro avg	0.98	0.97	0.98	971
Weighted avg	0.98	0.98	0.98	971

The Fig. 7 shows the important feature of word which was evaluate based on the Random Forest model. Here in this diagram, we visualize the top 20 features from the predicted interaction by visualizing we can able to identify and understand the particular terms which are significant during the prediction.
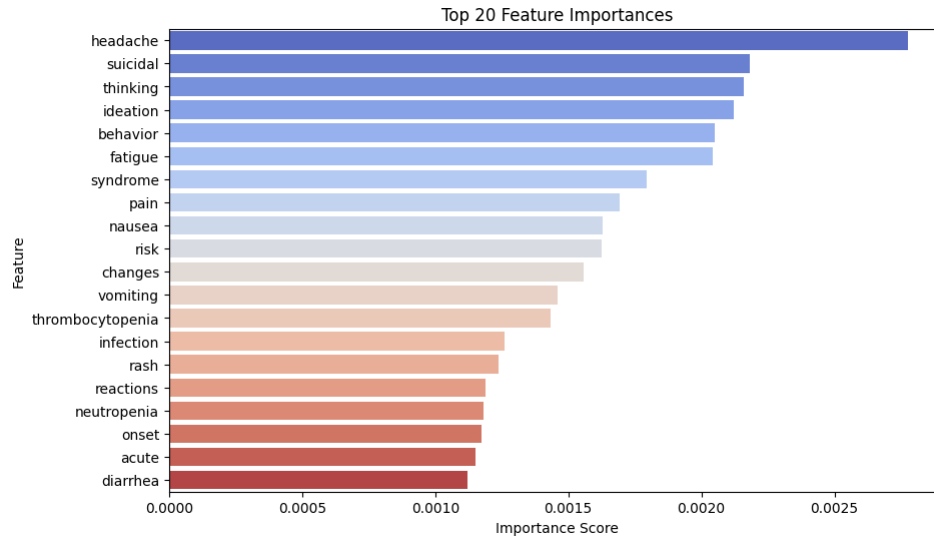
Figure 7 : Important Features in Random Forest Model

5.2 Gradient Boosting Algorithm and its performance:

It is a powerful machine learning technique used mainly for regression and classification tasks [26-30]. It builds a predictive model in a stage wise fashion by sequentially adding models. Boosting models trains models sequentially by new model focusing on residual errors. The initial model makes predictions based on a constant value, such as the mean of the target variable for regression or the most frequent class for classification. Next step is to calculate the residuals based on the difference between the predicted and actual values. These residuals represent the errors. New model is trained based on the residuals of the previous model.The new model is added to the ensemble, and its predictions are combined with previous models.This process is repeated for a set number of iterations until gradually improving the overall model by reducing prediction errors in each step.

$$F_m(x) = F_{m-1}(x) + \lambda.h_m(x)$$

Where,
$F_m(x)$ - the prediction at m
$F_{m-1}(x)$ - previous model's prediction.
$h_m(x)$ - new model trained to predict the residuals.
λ - the learning rate.

Gradient Boosting the disadvantage of overfitting, especially when using complex trees. To prevent this they use regularization techniques like Learning Rate, Tree Depth, Subsampling.
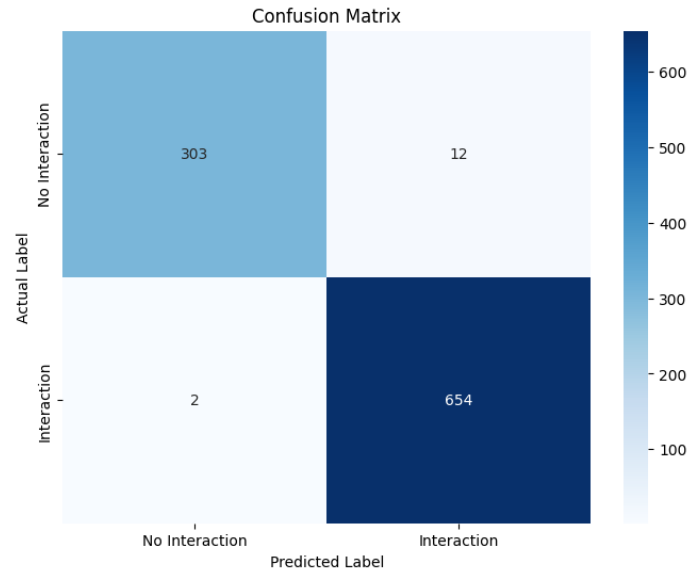
Figure 8 : Confusion matrix for Gradient Boosting

True Positive (TP) = 654, False Positive (FP) = 12, True Negative(TN) = 303, False Negative (FN) = 2

   Based on the implementation using the respective formulas for accuracy , f1-score,precision and recall we get like :
From Eq (1)   Accuracy = 0.9856
From Eq (2)   F1-Score = 0.9874
From Eq (3)   Precision = 0.9820
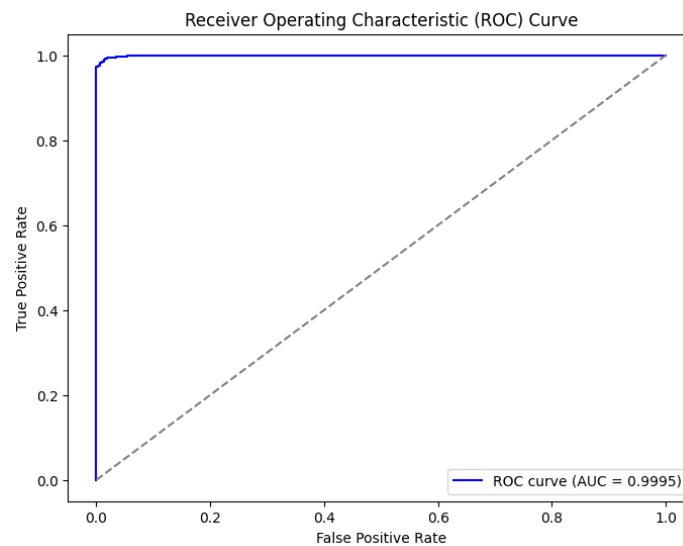From Eq (4)   Recall   = 0.9970

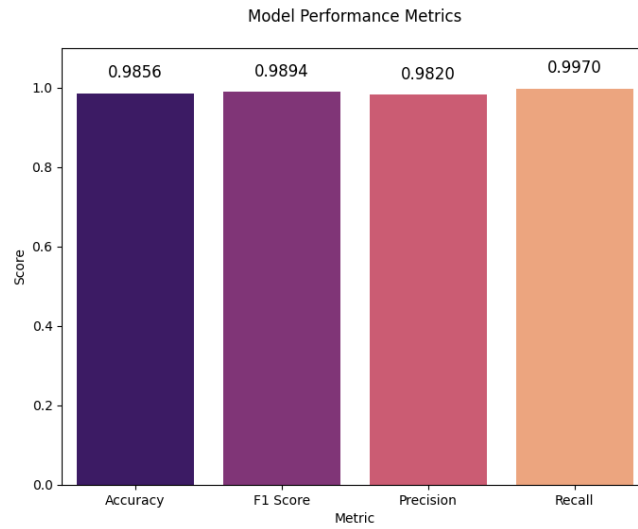**Figure 9 :** ROC curve for Gradient Boosting
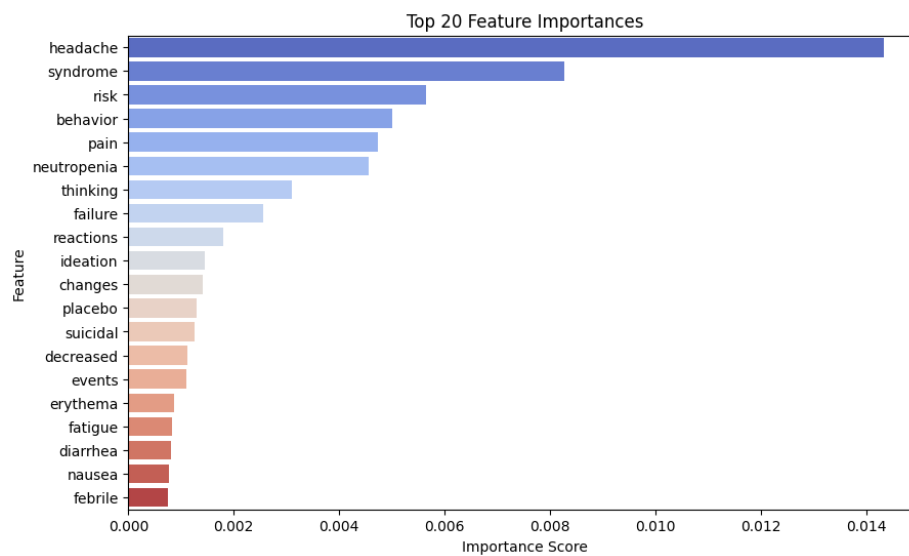
Figure 10 : Bar graph for Gradient Boosting

Figure 11 : Important Features in Gradient Boosting

Table 2 : ClassificationReport for Gradient Boosting

Classification Report			
	precision	recall	f1-score	support
0	0.99	0.96	0.98	315
1	0.98	1.00	0.99	656
accuracy			0.99	971
Macro avg	0.99	0.98	0.98	971
Weighted avg	0.99	0.99	0.99	971

5.3 Support Vector Machine (SVM) and its performance:

It is a powerful and versatile method of supervised machine learning algorithm especially used for classification and regression tasks [31-37]. It works by finding the hyperplane that best divides the dataset into different classes. It has key concepts such as hyperplane, support vectors, Margin, Linear SVM, Non-linear SVM (Kernel Trick), Soft Margin and Regularization. Consider a binary classification problem with two classes (+1 and -1). The SVM algorithm would find a hyperplane that maximizes the margin between the two classes. If the data is not linearly separable,slack variables are introduced to allow margin violations.
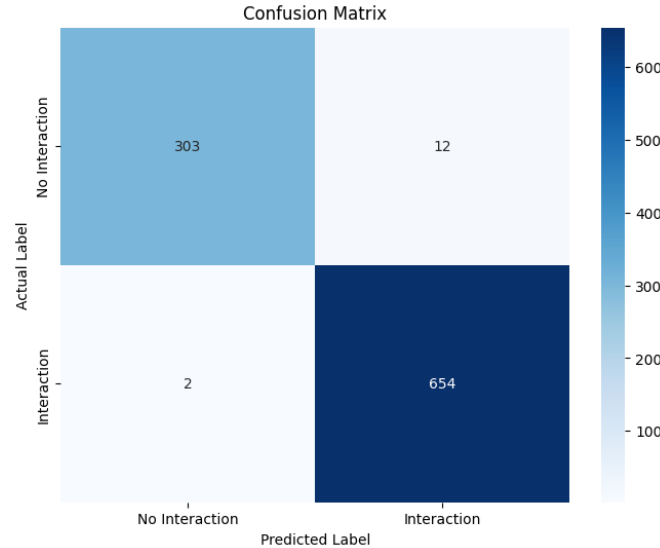

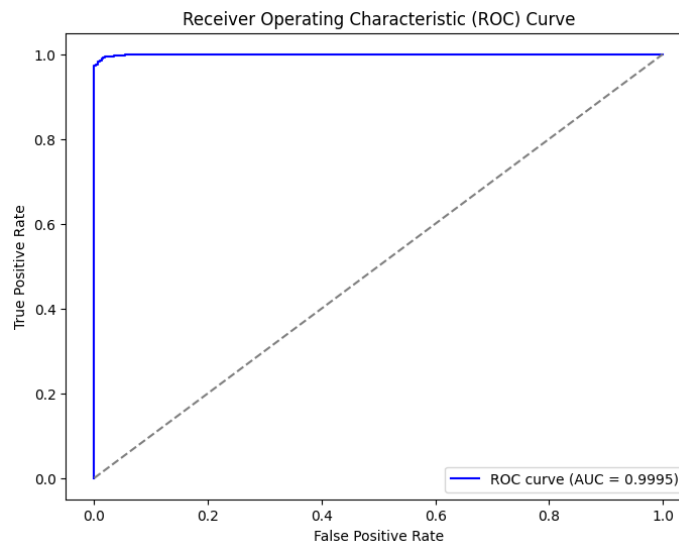**Figure 12 :** Confusion Matrix for SVM

Figure 13 : ROC curve for SVM

True Positive (TP) = 654, False Positive (FP) = 10, True Negative(TN) = 305, False Negative (FN) = 2
Based on the implementation using the respective formulas for accuracy , f1-score,precision and recall we get like :
From Eq (1)   Accuracy = 0.9876
From Eq (2)   F1-Score = 0.9909

From Eq (3)   Precision = 0.9849
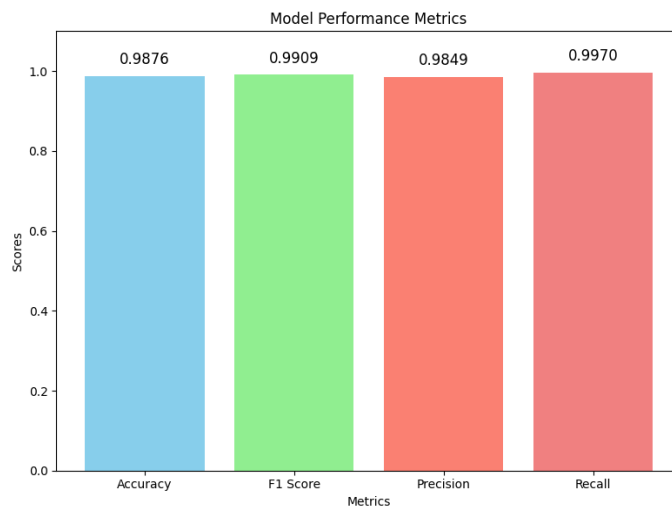From Eq (4)   Recall   = 0.9970

Figure 14 : Bar Graph for SVM

Table 3 : Classification Report for SVM

Classification Report			
	precision	recall	f1–score	support
0	0.99	0.97	0.98	315
1	0.98	1.00	0.99	656
accuracy			0.99	971
Macro avg	0.99	0.98	0.99	971
Weighted avg	0.99	0.99	0.99	971

5.4 BioBERT and its performance:

BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) is a pre-trained language model particularly focused on biomedical domain [38-40]. BERT (Bidirectional Encoder Representations from Transformers), one of the most important models in natural language processing (NLP), because it is specifically designed to handle biomedical and clinical literature. BERT outperform other traditional models by processing words in all direction. BioBERT is designed to handle biomedical text like Named Entity Recognition (NER), Relation Extraction, Questions related to biomedical. BioBERT retains the transformer architecture model from BERT. It consists of layers of self-attention mechanisms that allow it to process each word simultaneously attending every other word. The model's parameters include number of transformer layers, heads, and hidden units are identical to those of BERT, but the difference is pre-training corpus. In our proposed model for BioBERT, the BioBERT embeddings represent the side effects of drug pairs, these drug pairs are then as concatenated embeddings. A simple neural network model with one hidden layer was trained to classify whether the drug pairs have common side effects.
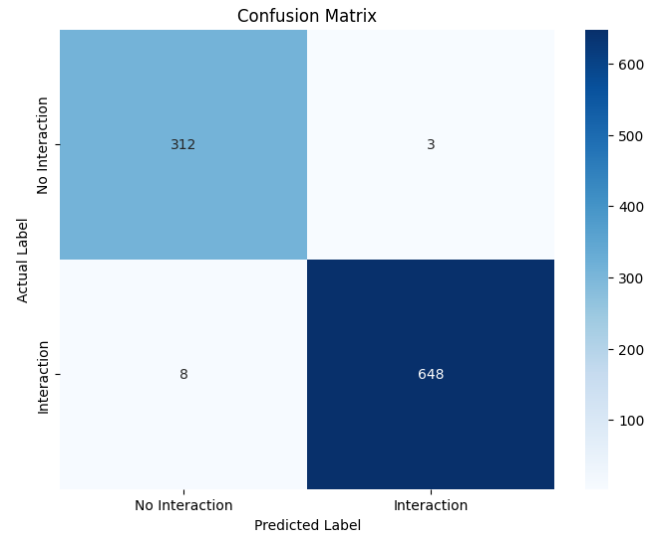
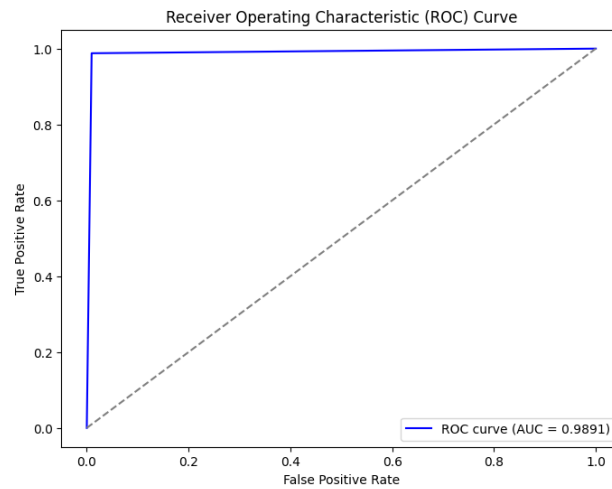**Figure 15 :** Confusion Matrix for BioBERT

Figure 16 : ROC curve for BioBERT

True Positive (TP) = 648, False Positive (FP) = 3, True Negative(TN) = 312, False Negative (FN) = 8
Based on the implementation using the respective formulas for accuracy , f1-score,precision and recall we
get like :

From Eq (1)   Accuracy = 0.9887
From Eq (2)   F1-Score = 0.9916
From Eq (3)   Precision = 0.9954
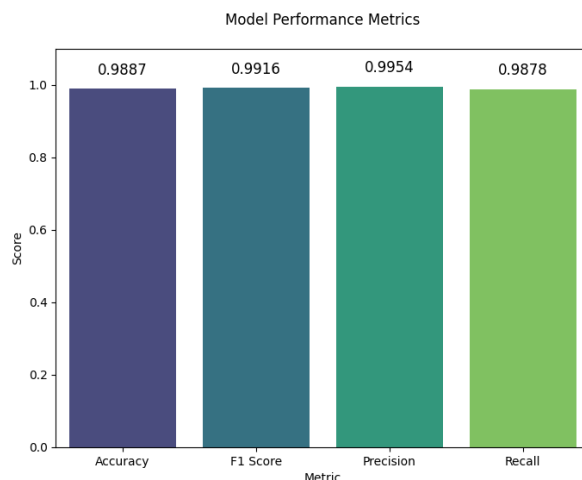From Eq (4)   Recall   = 0.9878

Figure 17 : Bar graph for BioBERT

Table 4 : Classification Report for BioBERT

Classification Report				
	precision	recall	f1-score	support
0.0	0.99	0.97	0.98	315
1.0	0.98	1.00	0.99	656
accuracy			0.99	971
Macro avg	0.99	0.98	0.98	971
Weighted avg	0.99	0.99	0.99	971

6 Final Comparison:

This table show the evaluation metrics we achieved from training the various model for DDI Extraction.

Table 5: Final Comparison Table

Model	Accuracy	F1-Score	Precision	Recall
Random Forest	0.9804	0.9856	0.9804	0.9909
Gradient Boosting	0.9856	0.9894	0.9820	0.9970
Support-Vector Machine	0.9876	0.9909	0.9849	0.9970
BioBERT	0.9887	0.9916	0.9954	0.9878

7 Discussion:

Our model is designed to classify drug-drug interactions using drugs and their Adverse reaction from the TAC 2017 dataset, we used TF-IDF for text processing, After TF-IDF it fed into normalization for balancing the unique features from the data for further processing, we proposed three machine learning models and one transformer model namely Random Forest, Gradient Boosting, Support vector machine and BioBERT for DDI (Drug-Drug Interaction) extraction. However our proposed model is aimed to predict and classify drug interaction accurately, along with increasing training and Visualizing evaluation through

metrics to ensure that our model can able to understand and improve the performance. The final predicted model gives the side effects of combining drug A and drug B. For Example, combining gattex and fycompa may lead to the following side effects : pyrexia, reactions, fatal, oral, hypotension.

8 Limitation

In this study, we done the classification for drug and their adverse reaction using four models, even though they yield promising accuracy for the classification for DDI(Drug-Drug Interaction) extraction but it is not considered as a perfect model. In future, we will try to implement a new approach with new dataset which comprises chemical molecule along with DDI information for giving a new insights for health care.

9 Conclusion:

In this study we proposed Multiple approach for the classification for drug and their adverse reaction by using the TAC 2017 ADR Dataset. We designed three machine learning model with TF-IDF (Term Frequency-Inverse Document Frequency) and one transformer model for classification namely Random Forest, Gradient Boosting, Support Vector Machine and BioBERT(Bidirectional Encoder Representations from Transformers for Biomedical Text). For Random Forest we achieved an Accuracy of 0.9804, an F1-Score of 0.9856, a precision of 0.9804, a recall of 0.9909, for Gradient Boosting we achieved an Accuracy of 0.9856, an F1-Score of 0.9894, a precision of 0.9820, a recall of 0.9970, for Support Vector Machine we achieved an Accuracy of 0.9876, an F1-Score of 0.9909, a precision of 0.9849, a recall of 0.9970, and for BioBERT we achieved an Accuracy of 0.9887, F1-score of 0.9916, a precision of 0.9954, a recall of 0.9878. From this observation BioBERT model outperforms compared to the other Machine learning models with few increasing number of percentages with visualization metrics. The experimental results shows that our proposed model yields better performance for the classification for drug and their adverse reaction.

Author Contributions: The authors confirm their contribution to the paper as follows: Data collection, Model Implementation, Intrepretation and results: Gurpreet Singh, Renuka Ramasamy , Sundareswari Thiyagarajan. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data will be provided on request.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References:

[1]   Park, C., Park, J., & Park, S. (2020). AGCN: Attention-based graph convolutional networks for drug-drug interaction extraction. *Expert Systems with Applications*, *159*, 113538.

[2]   Manning, C. D. (2008). Introduction to information retrieval.

[3]   Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, *24*(5), 513-523.

[4]   Salton, G., & Buckley, C. (1987). *Term weighting approaches in automatic text retrieval*. Cornell University.

[5]   Hastie, T. (2009). The elements of statistical learning: Data mining, inference, and prediction.

[6]   Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. *(No Title)*.

[7]   Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

[8]   James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

[9]   Getman, J., Ellis, J., Song, Z., Tracey, J., & Strassel, S. M. (2017, November). Overview of Linguistic Resources for the TAC KBP 2017 Evaluations: Methodologies and Results. In *TAC*.

[10] Qiu, Y., Zhang, Y., Deng, Y., Liu, S., & Zhang, W. (2021). A comprehensive review of computational methods for drug-drug interaction detection. *IEEE/ACM transactions on computational biology and bioinformatics*, *19*(4), 1968-1985.

[11] Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., ... & Wilson, M. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research*, *46*(D1), D1074-D1082.

[12] Safdari, R., Ferdousi, R., Aziziheris, K., Niakan-Kalhori, S. R., & Omidi, Y. (2016). Computerized techniques pave the way for drug-drug interaction prediction and interpretation. *BioImpacts: BI*, *6*(2), 71.

[13] Heba, I., Amany, A., Ahmed, S. E., & Amr, S. (2014). Novel data-mining methodologies for detecting drug-drug interactions: a review of pharmacovigilance literature. *Adv Environ Sci Dev Chem*, *20*, 301-14.

[14] Liu, S., Tang, B., Chen, Q., & Wang, X. (2016). Drug-drug interaction extraction via convolutional neural networks. *Computational and mathematical methods in medicine*, *2016*(1), 6918381.

[15] Wu, H., Xing, Y., Ge, W., Liu, X., Zou, J., Zhou, C., & Liao, J. (2020). Drug-drug interaction extraction via hybrid neural networks on biomedical literature. *Journal of biomedical informatics*, *106*, 103432.

[16] Segura-Bedmar, I., Martínez, P., & de Pablo-Sánchez, C. (2011, December). A linguistic rule-based approach to extract drug-drug interactions from pharmacological documents. In *BMC bioinformatics* (Vol. 12, pp. 1-11). BioMed Central.

[17] Segura-Bedmar, I., Martinez, P., & de Pablo-Sánchez, C. (2011). Using a shallow linguistic kernel for drug–drug interaction extraction. *Journal of biomedical informatics*, *44*(5), 789-804.

[18] Zhou, D., Zhong, D., & He, Y. (2014). Biomedical relation extraction: from binary to complex. *Computational and mathematical methods in medicine*, *2014*(1), 298473.

[19] Bui, Q. C., Sloot, P. M., Van Mulligen, E. M., & Kors, J. A. (2014). A novel feature-based approach to extract drug–drug interactions from biomedical text. *Bioinformatics*, *30*(23), 3365-3371.

[20] Zaikis, D., & Vlahavas, I. (2021). TP-DDI: Transformer-based pipeline for the extraction of Drug-Drug Interactions. *Artificial Intelligence in Medicine*, *119*, 102153.

[21] Lim, S., Lee, K., & Kang, J. (2018). Drug drug interaction extraction from the literature using a recursive neural network. *PloS one*, *13*(1), e0190926.

[22] Hou, W. J., & Ceesay, B. (2018). Extraction of drug–drug interaction using neural embedding. *Journal of bioinformatics and computational biology*, *16*(06), 1840027.

[23] KafiKang, M., & Hendawi, A. (2023). Drug-drug interaction extraction from biomedical text using relation BioBERT with BLSTM. *Machine Learning and Knowledge Extraction*, *5*(2), 669-683.

[24] Tran, T., Kavuluru, R., & Kilicoglu, H. (2021). Attention-gated graph convolutions for extracting drug interaction information from drug labels. *ACM transactions on computing for healthcare*, *2*(2), 1-19.

[25] Sefidgarhoseini, S., Safari, L., & Mohammady, Z. (2023). Drug-Drug Interaction Extraction Using Transformer-Based Ensemble Model.

[26] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

[27] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

[28] Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, *7*, 21.

[29] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*.

[30] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*, *31*.

[31] Cortes, C. (1995). Support-Vector Networks. *Machine Learning*.

[32] Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*, 273-297.

[33] Schölkopf, B. (2002). Learning with kernels: support vector machines, regularization, optimization, and beyond.

[34] Scholkopf, B., & Smola, A. J. (2018). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

[35] Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, *2*(2), 121-167.

[36] Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.

[37] Vapnik, V. N., & Vapnik, V. (1998). Statistical learning theory.

[38] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234-1240.

[39] Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[40] Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

[41] Demner-Fushman, D., Shooshan, S. E., Rodriguez, L., Aronson, A. R., Lang, F., Rogers, W., ... & Tonning, J. (2018). A dataset of 200 structured product labels annotated for adverse drug reactions. *Scientific data*, *5*(1), 1-8.

[42] Zhu, Y., Li, L., Lu, H., Zhou, A., & Qin, X. (2020). Extracting drug-drug interactions from texts with BioBERT and multiple entity-aware attentions. *Journal of biomedical informatics*, *106*, 103451.

[43] Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[44] Mikolov, T. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[45] Cami, A., Manzi, S., Arnold, A., & Reis, B. Y. (2013). Pharmacointeraction network models predict unknown drug-drug interactions. *PloS one*, *8*(4), e61468.

[46] Zheng, Y., Peng, H., Zhang, X., Zhao, Z., Gao, X., & Li, J. (2019). DDI-PULearn: a positive-unlabeled learning method for large-scale prediction of drug-drug interactions. *BMC bioinformatics*, *20*, 1-12.

[47] Mostafapour, V., & Dikenelli, O. (2019). Attention-wrapped hierarchical blstms for ddi extraction. *arXiv preprint arXiv:1907.13561*.

[48] Boyce, R. D., Gardner, G., & Harkema, H. (2012, June). Using natural language processing to extract drug-drug interaction information from package inserts. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing* (pp. 206-213).

[49] Raja, K., Patrick, M., Elder, J. T., & Tsoi, L. C. (2017). Machine learning workflow to enhance predictions of Adverse Drug Reactions (ADRs) through drug-gene interactions: application to drugs for cutaneous diseases. *Scientific reports*, *7*(1), 3690.

[50] Björne, J., Airola, A., Pahikkala, T., & Salakoski, T. (2011). Drug-drug interaction extraction from biomedical texts with svm and rls classifiers. *Proceedings of DDIExtraction-2011 challenge task*, 35-42.

[51] Jang, H. Y., Song, J., Kim, J. H., Lee, H., Kim, I. W., Moon, B., & Oh, J. M. (2022). Machine learning-based quantitative prediction of drug exposure in drug-drug interactions using drug label information. *NPJ Digital Medicine*, *5*(1), 88.