Shahid Beheshti University
Data Science (2023-2024)

Final Project report

Authors: Mehrdad Baradaran – Katayoun Kobraei
Department of Computer Science
Shahid Beheshti University
meh.baradaran@mail.sbu.ac.ir
k.kobraei@mail.sbu.ac.ir

Abstract

In contemporary times, neural networks and deep learning have become prevalent and widely used in various fields and domains. Their application extends to the medical field, addressing issues in pharmaceuticals. One pertinent concern in the realm of medicine is whether the consumption of a drug in isolation leads to adverse effects. If so, which organ is affected? For instance, in prescription writing, it is crucial for a physician to consider the patient's medical history and potential sensitivities to ensure that prescribed medications do not pose life-threatening risks. Consequently, one task amenable to artificial intelligence is predicting drug pairs interaction. This involves anticipating potential adverse effects based on the specific features of each drug pair.

Several solutions have been proposed and investigated in this regard. One approach is designing a powerful neural network model capable of problem-solving. Another is implementing appropriate preprocessing for the data since datasets related to drugs, especially in the task of side effect prediction, tend to focus more on the similarities between drugs rather than specific drug features. Associating similar drugs based on these features can help attribute similar side effects. Additionally, combining these approaches is a viable strategy, which is the focus of our investigation.

We aim to examine this by employing a method and preprocessing based on drug similarities. The goal is to design a robust neural network model that achieves state-of-the-art performance. We will compare the results with a base model such as k-nearest neighbors (KNN), which is suitable for this task.

1 Introduction

The consumption of a drug in isolation can pose challenges, and if not prescribed correctly, it may lead to adverse effects that not only hinder the individual's improvement but also jeopardize their overall health. Now, consider a scenario where an individual is taking multiple drugs concurrently, presenting various challenges and potential complications. Predicting the possible side effects of such drug combinations becomes crucial. Therefore, addressing this issue and forecasting the potential adverse effects for each drug pair holds significant importance.

In this task, we aim to effectively solve this problem and predict the potential adverse effects for each drug pair using neural networks and a series of analyses and processes. By receiving a drug pair as input and predicting the associated side effects as output, we can expand this task further. By utilizing sigmoid instead of softmax, multiple adverse effects can be predicted simultaneously. We plan to utilize two datasets for this purpose, which will be discussed and introduced in detail later. In summary, one dataset contains drug information, while the other depicts drug pairs interactions.

Solving this task poses numerous challenges that may prove to be problematic. For instance, the datasets available for these tasks are not only incomplete and error-prone but also suffer from a

scarcity of samples. This presents a challenge for neural network models that are data-hungry. Furthermore, existing datasets often exhibit data imbalance issues, strongly favoring one class or label, leading to model bias. These challenges are also prevalent in our dataset. Occasionally, information about a drug is missing, resulting in a highly sparse dataset, where extensive zero and one values exist for each drug (indicating the presence or absence of a specific enzyme or feature).

The dataset we have is highly imbalanced, to the extent that even simple models may achieve high accuracy (e.g., 80%) due to bias towards a specific class. However, in this task, a valuable model or process is one that helps accurately predict the side effects related to classes with very few samples. This is crucial because, in the real world, patients' lives are at stake. As a result, evaluation metrics such as F1 score, recall, and precision carry more importance than loss and accuracy.

The most significant challenge lies in interpreting the drug-related data. The information present in the datasets is intricate and specialized, requiring a certain level of medical knowledge to extract and analyze accurately. In the exploratory data analysis (EDA) phase, we have attempted to analyze these datasets to the best of our knowledge and capabilities.

2 Related work/Background

In the following section, we aim to introduce and analyze some of the approaches, solutions, and methods that have been proposed in recent years to address this task. Notable examples include the GNN-DDI and MDF-SA-DDI models, each presenting distinctive ideas for tackling this problem.

GNN-DDI aims to predict Drug-Drug Interactions (DDIs) by leveraging an attributed heterogeneous graph of drugs and various feature matrices. The approach is divided into two main stages:

First Stage: Constructing Similarity Matrices and Heterogeneous Network

1. **Collect Data and Construct Similarity Matrices:**
    - Collect five adjacency matrices from information sources, representing drug-drug interactions, and various features (enzyme, target proteins, chemical structure, pathway).
    - Utilize the Jaccard similarity function to calculate similarity matrices for each adjacency matrix, capturing the relationships between drugs based on different features.
2. **Constructing a Heterogeneous Network:**
    - Build an attributed heterogeneous network using the drug-drug edge list, indicating interactions between drugs and specifying the type of relationship.
    - Nodes represent drugs, and attributes of nodes correspond to drug features, forming a network with multiple edge types.
3. **Extracting Drug Embedding Vectors:**
    - Employ a Graph Neural Network (GNN)-based model to learn the embeddings of drugs in the attributed heterogeneous network.
    - Generate embeddings for each drug for all interaction types, resulting in four embedding matrices.

Second Stage: DDI Prediction using Fully Connected Deep Learning

4. **Dimensions Reduction of Embedding Matrix:**

- Concatenate (aggregate) drug embedding matrices to reduce dimensions into one-dimensional feature vectors.

- Obtain feature vectors for drug pairs in the DDI list by element-wise product of the feature vectors of the two drugs in the pair.

5. **DDI Prediction by Fully Connected Deep Learning Network:**

    - Employ a fully connected deep learning network with four sub-networks, each using one of the drug feature matrices as input.

    - Use hidden layers, batch normalization, and dropout layers for regularization and generalization.

    - Merge the outputs of sub-networks by calculating the average and employ a softmax layer for final predictions.

    - Utilize cross-entropy loss function, Adam optimizer, and early-stopping for model training.

Overview of the Approach:

- The method starts with data collection and similarity matrix computation.

- It constructs an attributed heterogeneous network to represent drug interactions and features.

- GNN is applied to learn drug embeddings considering both network structure and node attributes.

- The embedding matrices are then used to predict DDIs using a fully connected deep learning network with a concatenation strategy.

- The final predictions are obtained by merging the outputs of sub-networks.

This two-stage approach combines graph-based representation learning with deep learning techniques to predict drug interactions, leveraging diverse information sources and features.

MDF-SA-DDI, focuses on predicting Drug-Drug Interaction (DDI) events by combining information from chemical substructures, targets, and enzymes. The process involves feature extraction, multi-source drug fusion, latent feature fusion, and a multi-head attention mechanism. Here's a breakdown of the method:

1. Feature Extraction:

- **Choice of Features:**

    - The study uses three features - substructures, targets, and enzymes - as they have shown to yield the best performance in previous research.

    - Each feature corresponds to a set of descriptors, and drugs are represented by binary feature vectors.

- **Dimensionality Reduction:**

    - Instead of using bit vectors directly, the Jaccard similarity is calculated from the bit vectors to create a lower-dimensional representation.

    - Jaccard similarity is computed based on the intersection and union of the original bit vectors.

2. Multi-Source Drug Fusion:

1. **Convolutional Neural Networks (CNN):**
    - Each drug is represented as a 1*k-dimensional vector.
    - Drug vectors are combined into a 2*k-dimensional matrix and fed into a CNN.
    - The CNN outputs a row vector as the latent vector of the drug pair.
2. **Autoencoders with Self-Attention Mechanism:**
    - Autoencoders are used to reduce the dimensionality of input features and fuse three features together.
    - A self-attention mechanism is added to focus on internal correlations.
    - Two autoencoders (AE1 and AE2) are employed to obtain latent vectors of the drug pair.
3. **Siamese Network:**
    - The Siamese network, with shared weights, is used to measure the similarity of two inputs.
    - Two autoencoders with self-attention mechanism serve as sub-networks of the Siamese network.
    - The Siamese network extracts information from a single drug, reducing differences between input orders.

3. Latent Feature Fusion:

- **Transformer Encoder Structure:**
    - The encoder structure of the transformer is used to perform latent feature fusion.
    - Four different drug fusion methods and network structures generate four different latent vectors of drug pairs.
    - These vectors are element-wise added and concatenated to form a new feature for the drug pair.

4. Multi-Head Attention Mechanism:

- **Multi-Head Self-Attention Mechanism:**
    - Different latent feature vectors contribute inconsistently to DDI prediction.
    - Latent vectors are concatenated and input into the multi-head attention module.
    - The self-attention mechanism helps recognize important features for prediction.

Additional Techniques:

- **Residual Connections & Layer Normalization:**
    - Residual connections are used to mitigate the problem of gradient disappearance.
    - Layer normalization is applied after self-attention and feed-forward network layers to improve convergence.
- **Data Augmentation (Mixup):**

    - Mixup, a data augmentation algorithm, is employed to improve the generalization ability and robustness of the model.

- **Loss Function:**

    - Focal Loss (FL) is chosen for classification to address imbalances in sample sizes.

    - The model uses cross-entropy loss in the first half of training steps and switches to focal loss in the second half.

    - Mean Squared Error (MSE) loss is used as the auxiliary loss for autoencoders.

MDF-SA-DDI integrates diverse information through various deep learning techniques, including CNNs, autoencoders with self-attention, Siamese networks, and transformer-based latent feature fusion, ultimately employing a multi-head attention mechanism for enhanced DDI prediction. The methodology aims to leverage multiple perspectives on drug features to improve prediction accuracy.

In summary, GNN-DDI and MDF-SA-DDI demonstrate strong performances, with slight variations in different metrics. GNN-DDI excels in precision, while MDF-SA-DDI has a better balance between precision and recall. The choice between the two methods depends on specific use-case requirements, such as the importance of precision-recall balance or the need for high overall accuracy.

Several other models and approaches exist that are both older and demonstrate comparatively weaker results when compared to GNN-DDI and MSTE. These models may not necessitate explicit mention due to their outdated nature and inferior performance in the current landscape.

3 Proposed method

In this section, it is time to discuss our own idea and approach. To facilitate a better understanding of the principal approach, we will delineate it into three stages. In each segment, comprehensive explanations will be provided to elucidate the corresponding details.

In the initial approach, an attempt was made to address the data imbalance issue by designing a powerful model. The conceptualization involves a dual-model neural network, simultaneously incorporating an autoencoder and a classifier in its architecture. The rationale behind this design is twofold: firstly, to prevent bias towards a specific space due to data imbalance, and secondly, to enable the model to create a suitable representation in space for each side effect and drug pair. In this way, the inputs can be clustered in space based on their labels, allowing for high accuracy and minimal error. The final loss of the model is derived from the sum of the classifier loss and the autoencoder loss. This is done to ensure that the model not only retains data for reconstruction but also generates latent representations that align with the aforementioned idea.

Regarding the dataset's structure, it should be noted that the data is highly sparse. If the original form of the dataset were used, it would comprise around 29 thousand features, each taking values of 0 or 1. Each sample and drug has a substantial number of 0 features. Consequently, the reconstruction loss would perform well even in the worst-case scenario. The presence of a classifier aims to force the model not only to represent inputs in space and reduce this sparsity but also to enhance interpretability. The autoencoder is included to provide a representation that allows for task classification. The model should simplify the separation of samples based on their positions in space so that the classification error is minimized. All of these aspects are visualized in **Figure 2**.

Both autoencoder and classifier components employ fully connected layers, with cross-entropy utilized for classification loss and mean squared error (MAE) for the autoencoder section. Training is initiated using the Adam optimizer. To ensure balanced distributions for all labels and validations, a 5-fold cross-validation approach is employed. The final result is reported as the average across all 5 folds. A learning rate scheduler is used to enhance model convergence.

A crucial note is that our input data is the concatenated information of drug A and drug B. Given the inherent order sensitivity, flipping the order of drugs results in different inputs; however, the label remains the same. Consequently, for each drug pair, its inverse is added as new data to the dataset, effectively performing data augmentation. This approach, utilizing the proposed model, is conducted without any specific preprocessing steps, relying solely on the model design outlined with its objectives.

The process of constructing the latent space can be observed in **Figure 1**.

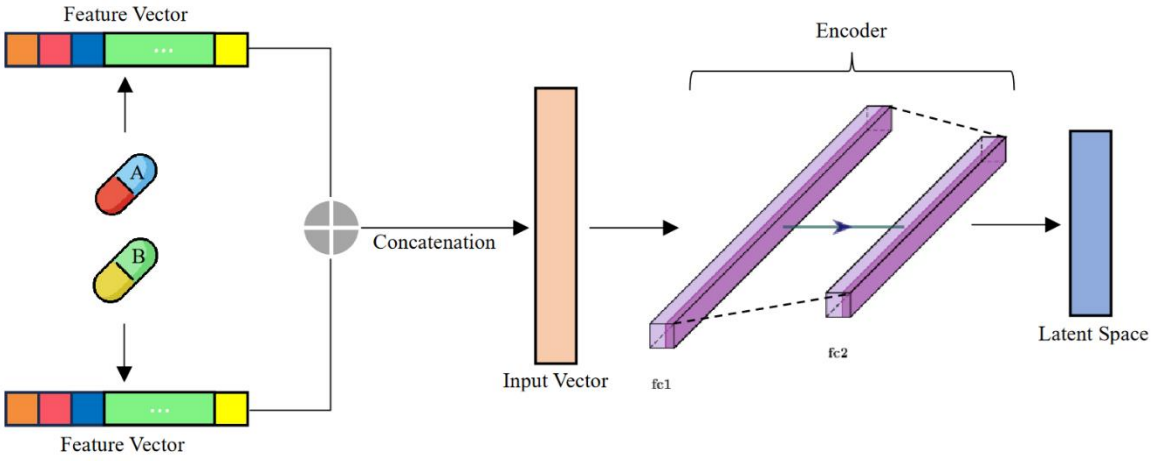

Figure1. Constructing the latent space based on the feature vectors of drug pairs

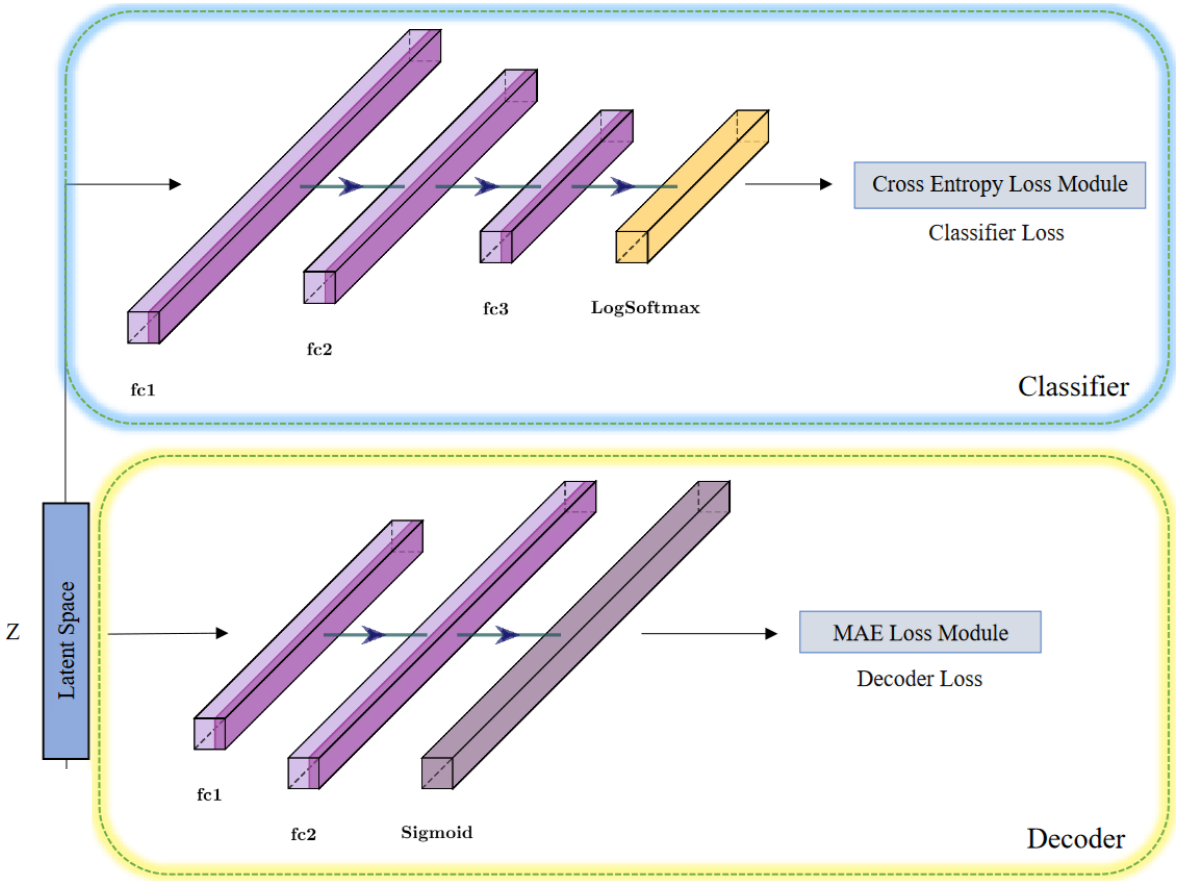

Figure 2. Model Architecture

In conclusion, you can observe the methodology for calculating the total loss for this model based on the individual losses depicted in **Figure 3**.

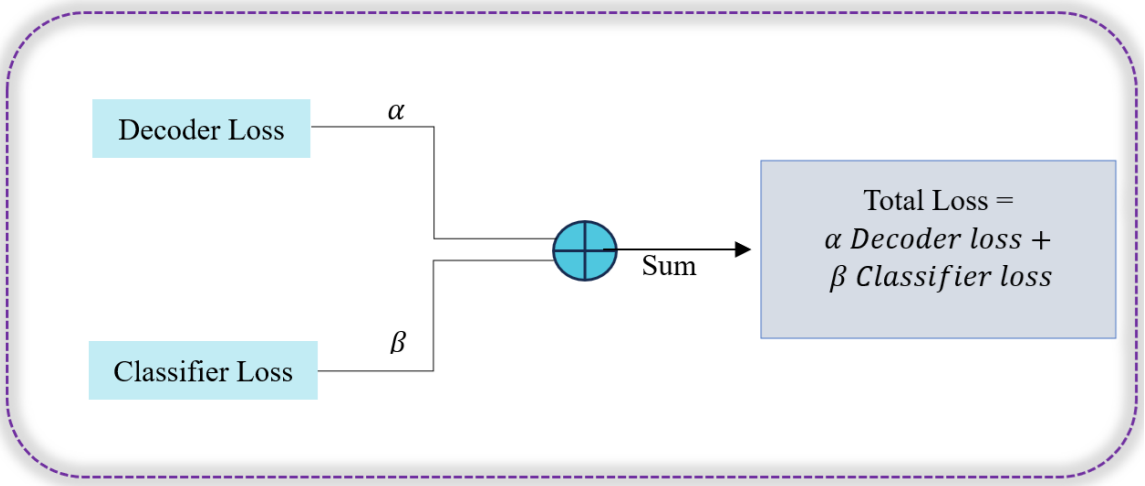

Figure 3. Calculation for Total Loss

In the second approach, we aim to propose a method and algorithm for calculating drug similarity. As mentioned initially, identifying the similarity and entropy of drugs is among the crucial tasks that can be performed as part of preprocessing. This is because drugs with similar content should encompass similar side effects. In this approach, we employ Jaccard similarity to accomplish this task. Subsequently, we provide a step-by-step explanation of this approach.

1. Constructing Drug Features:

- **Feature Extraction Module:**
  - The feature matrix construction involves two modules: Feature extraction and Aggregation.
  - Features include mono side effects, targets, enzymes, chemical substructures, and pathways.
  - Binary matrices are constructed for each feature, representing the presence or absence of a feature for each drug.
  - Jaccard similarity is calculated for each pair of drugs based on these matrices.
- **Aggregation Module:**
  - Similarity matrices for each feature are concatenated, resulting in the main drug feature matrix.
  - The resulting matrix has drug IDs as rows and feature information as columns.
  - The size of the matrix is determined by the number of drugs and features in the dataset.

2. DDI Prediction:

- **Deep Neural Network (DNN) Architecture:**
  - A DNN architecture is employed for DDI prediction, consisting of an input layer, multiple hidden layers, and an output layer.
  - Parameters investigated for optimization include the number of hidden layers, neurons in hidden layers, activation functions, and dropout rate.
  - Training is performed with various configurations, and the best-performing setup is identified.

- **Training Parameters:**
  - Hidden layers: {1, 2, 3}
  - Neurons in hidden layers: {128, 256, 512}
  - Activation functions: {ReLU, Sigmoid, Softmax}
  - Dropout rate: {0.1, 0.2, 0.3, 0.4, 0.5}
  - Training iterations (epochs): 100
  - Batch size: 128
  - Loss function: Binary-cross-entropy
  - Optimization algorithm: Adam optimizer.
- **Training Process:**
  - Drug feature vectors are summed for each drug pair and fed into the DNN.
  - The DNN is trained to predict the probability of each DDI event.
  - The model outputs the probability distribution, and the event with the highest probability is selected as the predicted DDI.

You can refer to **Figure 4** for an illustration of all these explanations.

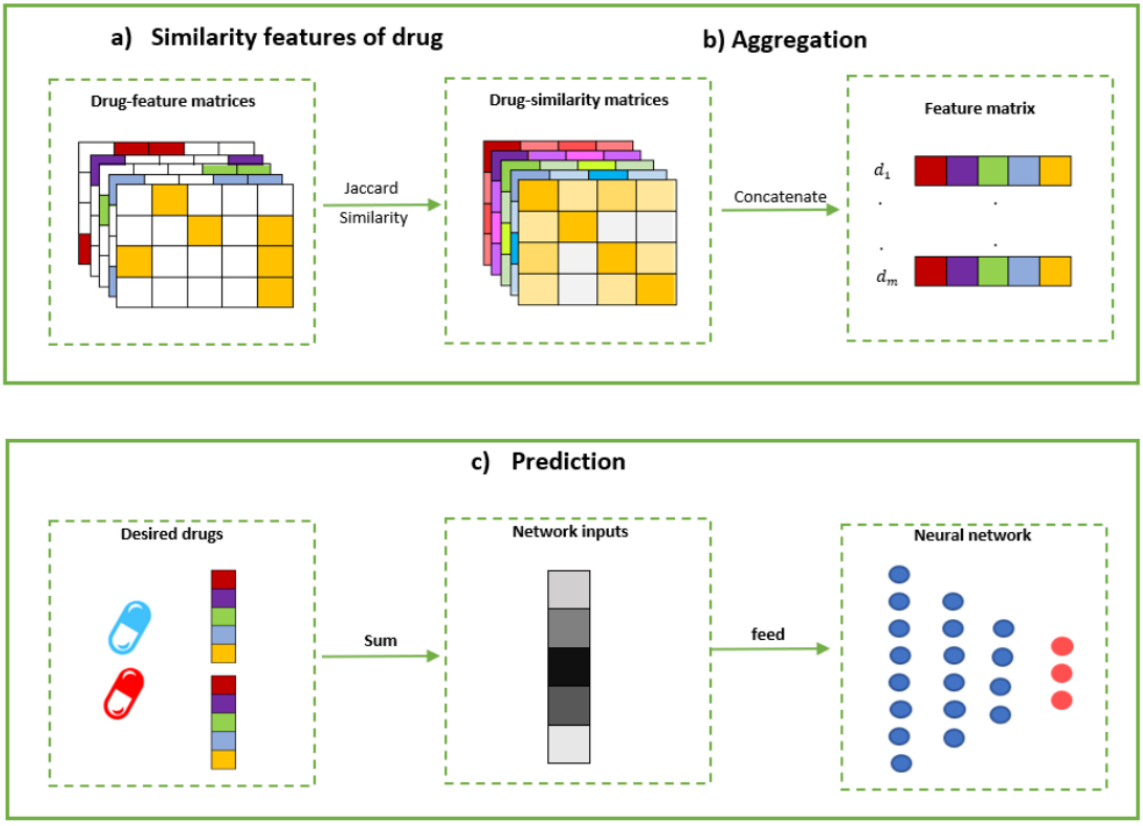

Figure 4. Second Approach Illustration

In the third approach, which constitutes our final model, we achieved notable results by combining the features obtained, incorporating the preprocessing from the second approach, and utilizing the powerful model architecture from the first approach. The obtained results surpass those of the three models and previous state-of-the-art models in this domain and task. We compare the performance of these three models with the earlier state-of-the-art models. Additionally, the data derived from the third approach is not only utilized to shape the final model but is also employed to provide a

baseline model. This baseline model is represented by a k-nearest neighbors (KNN) algorithm, enabling us to evaluate the classification performance of our algorithms and compare it with our developed approach.

4 Results

At first, you should describe the dataset that you have used, and then, you must describe the experiments that you maid, and its results. It is better to test your proposed method in different settings and with various parameters.

The dataset for this task is defined as DDI, comprising two subsets. One dataset includes comprehensive information for each drug, such as SMILES representation, pathway involvement, enzyme details, target information, and side effects. Each of these features lists all relevant entities separated by "|", resulting in approximately 10,000 features. The values for these features are binary (0 or 1), as illustrated in **Table 1**. Another dataset encompasses drug interactions, where each sample involves two drugs and a final label indicating the associated side effect for each drug pair. This can be observed in **Table 2**.

	name	C0854693	C0151694	C3665346	C0152078	C1696849	C0232600	C1504448	C0278835	C2242548	...	C0038273	C0042903	C0349667	C1696575
0	Glucosamine	0	0	0	0	0	0	0	0	0	...	0	0	0	0
1	Azelnidipine	0	0	0	0	0	0	0	0	0	...	1	0	0	0
2	Abiraterone	0	0	0	0	0	0	0	0	0	...	0	0	0	0
3	Flecainide	0	0	0	0	0	0	0	0	0	...	0	0	0	0
4	Caffeine	0	0	0	0	0	0	0	0	0	...	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
567	Ketazolam	0	0	0	0	0	0	0	0	0	...	0	0	0	0
568	Lansoprazole	0	0	0	0	0	0	0	0	0	...	0	0	0	0
569	Bosentan	0	0	0	0	0	0	0	0	0	...	0	0	0	0
570	Ezogabine	0	0	0	0	0	0	0	0	0	...	0	0	0	0
571	Crizotinib	0	0	0	0	0	0	0	0	0	...	0	0	0	0

572 rows × 9992 columns

Table 1. Drug information data frame

	drugA	drugB	side
0	Abemaciclib	Amiodarone	4
1	Abemaciclib	Apalutamide	14
2	Abemaciclib	Aprepitant	15
3	Abemaciclib	Atomoxetine	0
4	Abemaciclib	Bortezomib	0
...	...	...	...
37259	Nefazodone	Netupitant	15
37260	Nefazodone	Nicardipine	0
37261	Neratinib	Netupitant	15
37262	Netupitant	Nicardipine	15
37263	Nicardipine	Naproxen	0

37264 rows × 3 columns

Table 2. Drug pairs interactions

Continuing with the analysis, we examine some crucial Exploratory Data Analysis (EDA) findings to delve into important insights, concepts, and the underlying context of the dataset.

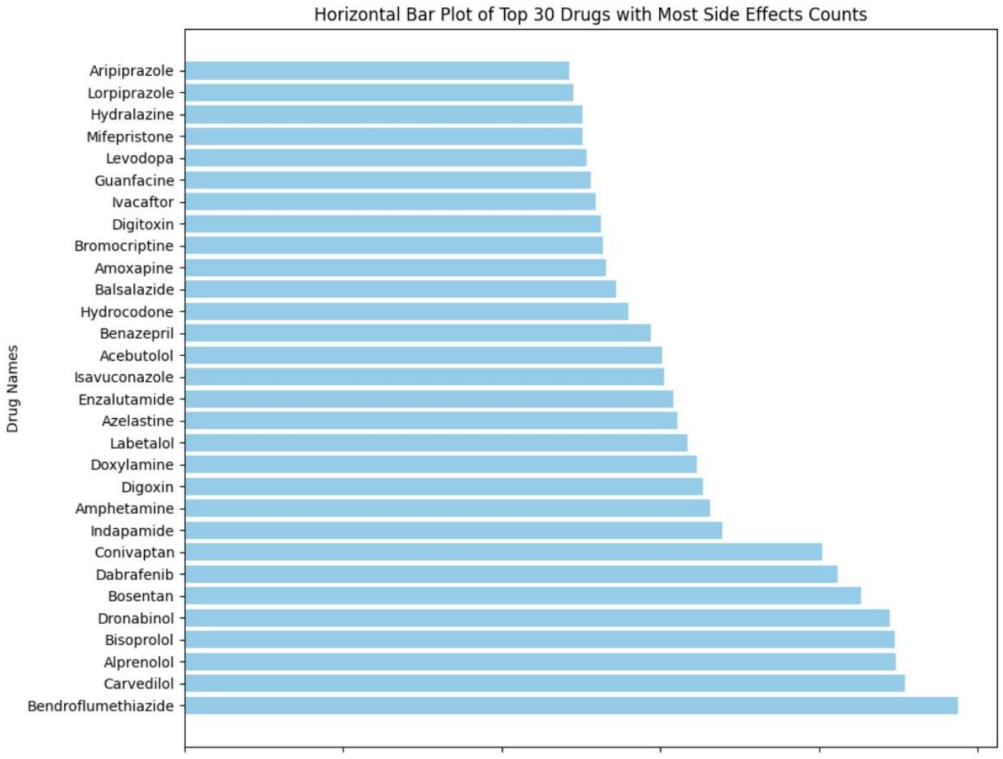

Figure 5. Firstly, we identify drugs with a higher likelihood of side effects compared to others. Subsequently, we select the top 30 from this group. Then, we analyze common features among these drugs to discern patterns. Upon investigation, we find that many high-risk drugs belong to specific categories, such as heart disease medications like Bendroflumethiazide for hypertension and Carvedilol for heart failure and hypertension. Additionally, drugs for specialized conditions like cancer and AIDS, as well as psychiatric medications like Aripiprazole for ADHD or schizophrenia, constitute notable risk-prone groups. Recognizing the drug category is crucial, as certain groups pose higher risks of side effects.

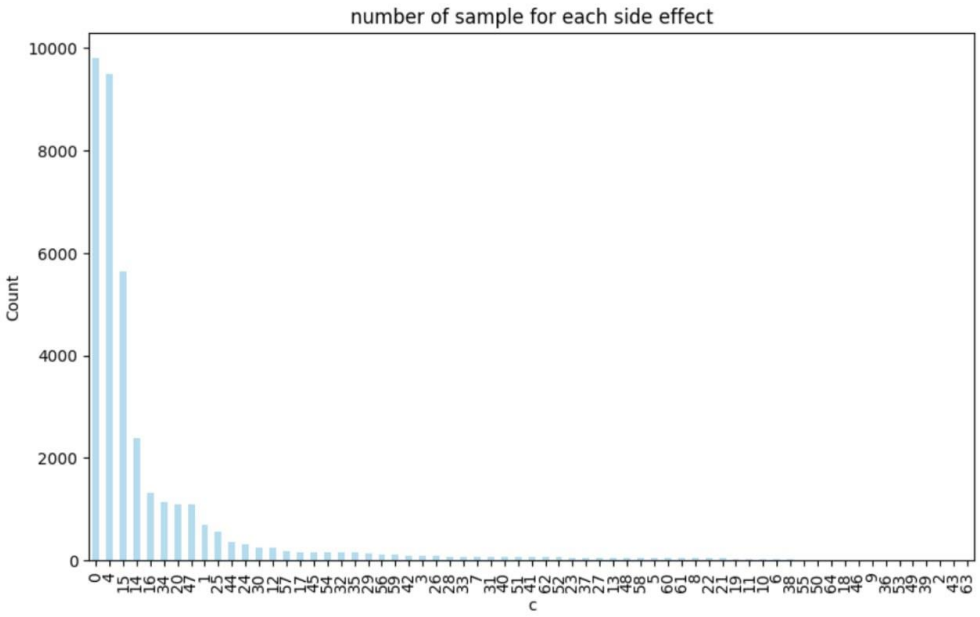

Figure 6. Firstly, we identify drugs with a higher likelihood of side effects compared to others. Subsequently, we select the top 30 from this group. Then, we analyze common features among these drugs to discern patterns. Upon investigation, we find that many high-risk drugs belong to specific categories, such as heart disease medications like Bendroflumethiazide for hypertension and Carvedilol for heart failure and hypertension. Additionally, drugs for specialized conditions like cancer and AIDS, as well as psychiatric medications like Aripiprazole for ADHD or schizophrenia, constitute notable risk-prone groups. Recognizing the drug category is crucial, as certain groups pose higher risks of side effects.

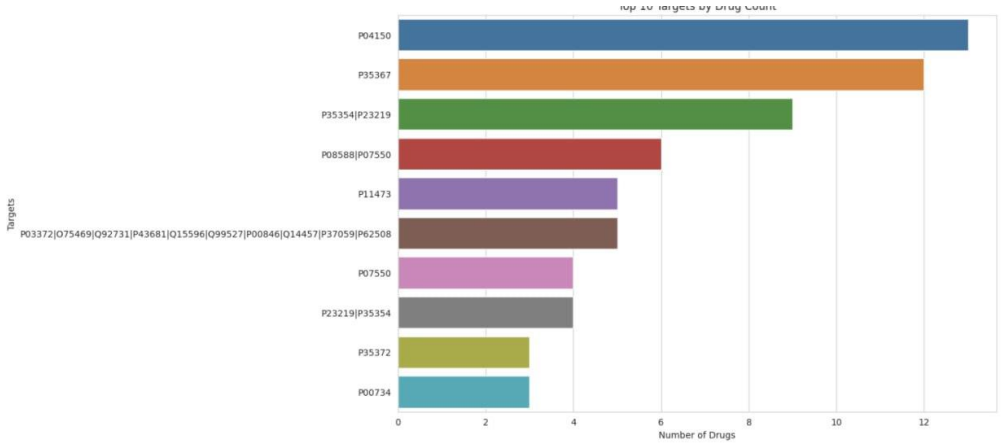

Figure 7. By examining each feature, we identify the most repeated values, such as p04150 in the target feature. Subsequently, we explore whether drugs associated with this value exhibit shared side effects. If not, it suggests that this particular value might not pose a high risk for side effects. This method is systematically applied to all features in the dataset, allowing us to uncover the most repeated values and their associated information, aiding in the identification of potential risks and patterns.

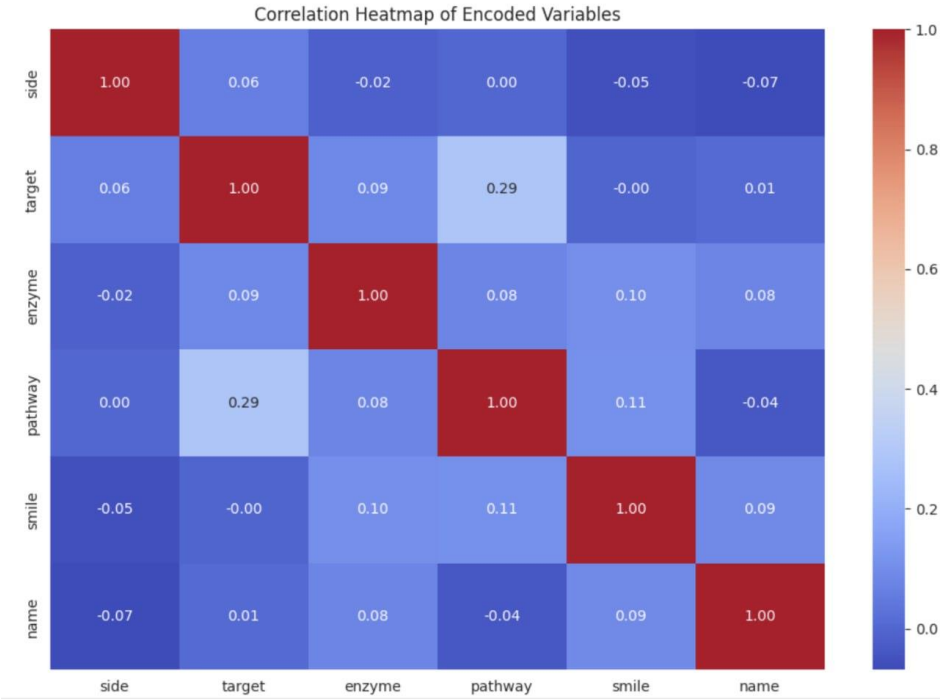

Figure 8. Indeed, exploring relationships between features is crucial. Logically, the target and pathway features are expected to be more closely related. Additionally, observed connections between pathway and smile warrant thorough investigation. Analyzing one feature necessitates an examination of the other to identify potential patterns and establish comprehensive insights into the dataset's inter-feature relationships.

After analyzing the data, it is now time to examine the learning results of the three mentioned approaches, along with the baseline method, which is the k-nearest neighbors (KNN) algorithm.

Initially, we present the results of the previous state-of-the-art models, followed by the results of the three proposed approaches and the KNN method. Subsequently, a comparative analysis is facilitated through the utilization of a radar plot, as depicted in **Figure 9**. Furthermore, the accuracy of the discussed models and KNN can be visualized in **Figure 10**.

GNN-DDI and MDF-SA-DDI:

- GNN-DDI:
    - Accuracy: 0.9180
    - AUPR: 0.9709
    - AUROC: 0.9985
    - F1 Score: 0.8440
    - Precision: 0.9000
    - Recall: 0.8165
- MDF-SA-DDI:
    - Accuracy: 0.9121
    - AUPR: 0.9657
    - AUROC: 0.9989
    - F1 Score: 0.8659
    - Precision: 0.8833
    - Recall: 0.8501

My Approaches:

- **Approach 1:**
  - Accuracy: 0.9336
  - AUPRC (Macro): 0.871626
  - AUROC (Macro): 0.932302
  - F1 (Macro): 0.830795
  - Precision (Macro): 0.859587
  - Recall (Macro): 0.818746
- **Approach 2:**
  - Accuracy: 0.932857
  - ROC AUPR (Micro): 0.976736
  - ROC AUC (Micro): 0.999024
  - F1 (Micro): 0.932857
  - Precision (Micro): 0.932857
  - Recall (Micro): 0.932857
- **Approach 3:**
  - Accuracy: 0.9523
  - AUPRC (Macro): 0.9447
  - AUROC (Macro): 0.9359
  - F1 (Macro): 0.8625
  - Precision (Macro): 0.8937
  - Recall (Macro): 0.8485
- **(KNN):**
  - Accuracy: 0.6901918690460217
  - Precision (Macro): 0.62
  - Recall (Macro): 0.43
  - F1 (Macro): 0.49

Analysis:

- **Approaches 1 and 2:**
  - Both approaches 1 and 2 show good performance across various metrics, with high accuracy, AUPR, AUROC, F1 scores, and precision-recall balance.
- **Approach 3:**
  - Approach 3 performs exceptionally well with high accuracy, AUPR, AUROC, and balanced F1, precision, and recall scores.
- **Approach 4 (KNN):**
  - KNN approach (Approach 4) has lower accuracy and macro-average scores compared to the other approaches, suggesting it may not perform as well on a diverse set of classes.

In summary, approaches 1, 2, and 3 seem to outperform GNN-DDI and MDF-SA-DDI in terms of the provided metrics, while Approach 4 (KNN) performs less favorably.

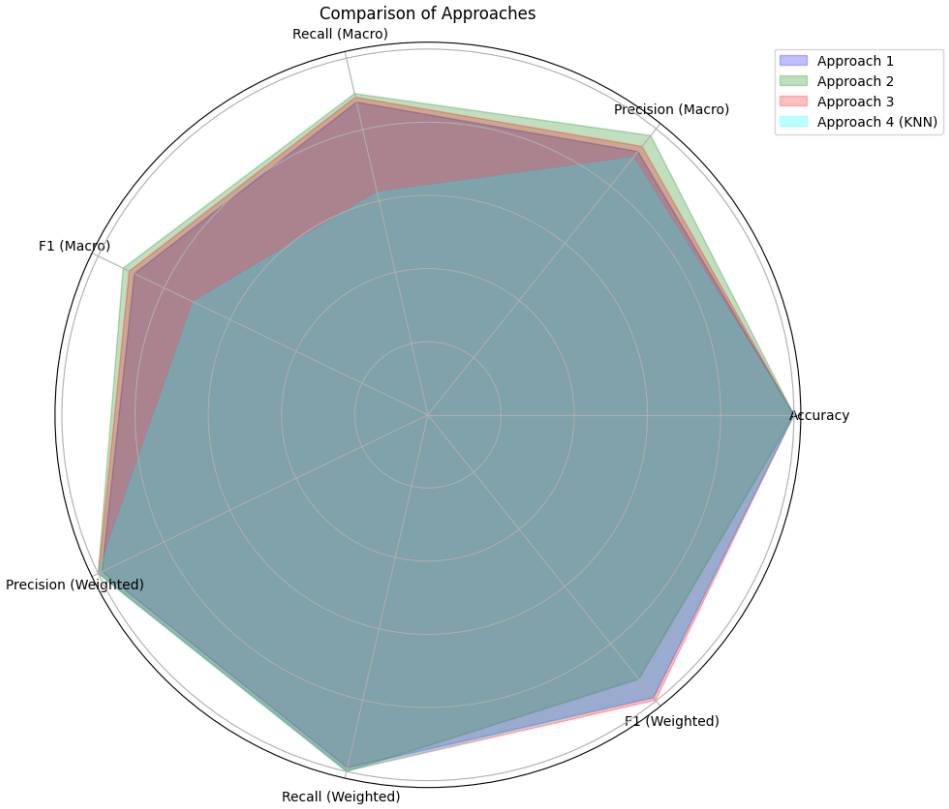

Figure 9. Radar plot for scores of 4 approaches

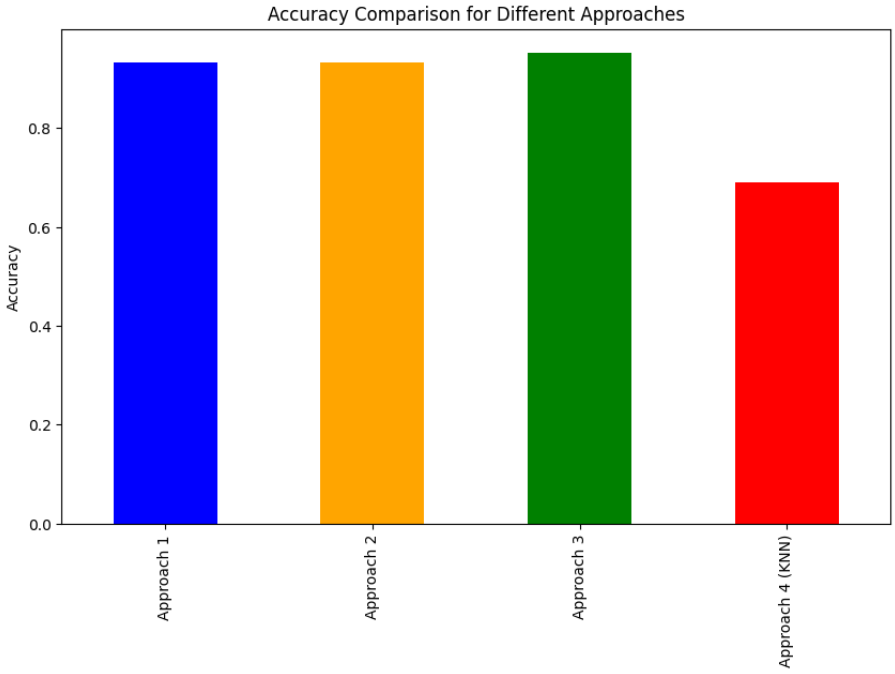

Figure 10. Bar plot of models accuracies

5 Discussion

After scrutinizing the obtained results, it is evident that each of the three approaches individually managed to outperform the state-of-the-art models. As discussed earlier, the first approach successfully addressed data sparsity by combining two neural network models. It not only resolved the sparsity issue but also created representations for inputs in a way that instances with specific labels were positioned closely in the feature space. This property was even verified in the code to ensure the validity of the claim. Additionally, it effectively tackled the imbalance in the data by segregating data in space, generating a latent space that not only yielded very low errors for sparse data but also facilitated the interpretation of drug interactions.

For the second approach, given the high significance of drug similarity in predicting side effects based on similar drugs, the introduced algorithm and method for finding drug similarities and extracting features with enhanced similarity among drugs led to highly satisfactory results.

It is evident that the third approach, leveraging the data generated in the second approach and training it with the model introduced in the first approach, outperforms both of them. The baseline model, KNN, using the data generated in the second approach, does not exhibit acceptable results, as illustrated.

In conclusion, the third approach, which combines the strengths of the initial two approaches and utilizes the generated data from the second approach, demonstrates superior performance in comparison to the state-of-the-art models, establishing its efficacy in addressing the challenges posed by the dataset and task at hand.

References

[1] Masumshah, R. & Eslahchi, C. (2023) "DPSP: a multimodal deep learning framework for polypharmacy side effects prediction." Bioinformatics Advances, 3(1), vbad110. DOI: 10.1093/bioadv/vbad110

[2] Al_Rabeah, M. H. & Lakizadeh, A. (2022) "GNN-DDI: A New Data Integration Framework for Predicting Drug-Drug Interaction Events Based on Graph Neural Networks." Research Square. DOI: 10.21203/rs.3.rs-1803318/v1

[3] Lin, S., Wang, Y., Zhang, L., Chu, Y., Liu, Y., Fang, Y., Jiang, M., Wang, Q., Zhao, B., Xiong, Y., & Wei, D.-Q. (2021) "MDF-SA-DDI: predicting drug–drug interaction events based on multi-source drug fusion, multi-source feature fusion and transformer self-attention mechanism." Briefings in Bioinformatics, 23(1), bbab421. DOI: 10.1093/bib/bbab421

[4] Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., & Frey, B. (Year). "Adversarial Autoencoders." arXiv preprint arXiv:1511.05644 [cs.LG]. DOI: 10.48550/arXiv.1511.05644