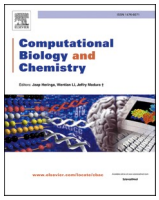
Computational Biology and Chemistry 110 (2024) 108066

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/cbac



K_1K_2NN: A novel multi-label classification approach based on neighbors for predicting COVID-19 drug side effects

Pranab Das, Dilwar Hussain Mazumder *

Department of Computer Science & Engineering, National Institute of Technology Nagaland, Chumukedima, Dimapur, Nagaland 797103, India

ARTICLE INFO

Keywords:
Drug Development
COVID-19
Side Effects
Chemical Properties
Multi-Label
Machine Learning

ABSTRACT

COVID-19, a novel ailment, has received comparatively fewer drugs for its treatment. Side Effects (SE) of a COVID-19 drug could cause long-term health issues. Hence, SE prediction is essential in COVID-19 drug development. Efficient models are also needed to predict COVID-19 drug SE since most existing research has proposed many classifiers to predict SE for diseases other than COVID-19. This work proposes a novel classifier based on neighbors named K_1 K_2 Nearest Neighbors (K_1K_2NN) to predict the SE of the COVID-19 drug from 17 molecules' descriptors and the chemical 1D structure of the drugs. The model is implemented based on the proposition that chemically similar drugs may be assigned similar drug SE, and co-occurring SE may be assigned to chemically similar drugs. The K_1K_2NN model chooses the first K_1 neighbors to the test drug sample by calculating its similarity with the train drug samples. It then assigns the test sample with the SE label having the majority count on the SE labels of these K_1 neighbor drugs obtained through a voting mechanism. The model then calculates the SE-SE similarity using the Jaccard similarity measure from the SE co-occurrence values. Finally, the model chooses the most similar K_2 SE neighbors for those SE determined by the K_1 neighbor drugs and assigns these SE to that test drug sample. The proposed K_1K_2NN model has showcased promising performance with the highest accuracy of 97.53% on chemical 1D drug structure and outperforms the state-of-the-art multi-label classifiers. In addition, we demonstrate the successful application of the proposed model on gene expression signature datasets, which aided in evaluating its performance and confirming its accuracy and robustness.

1. Introduction

In December 2019, a new disease was identified in Wuhan city, China, caused by a novel coronavirus variant, which quickly spread from one person to another (Abdalla and Rabie, 2023; Ahmad et al., 2022). It was then named novel Coronavirus Disease 2019 (COVID-19) (Alici et al., 2022; Atifa et al., 2022) and was quickly declared a world pandemic by the World Health Organization (WHO). As of this study, in early November 2023, more than 77,16,79,618 coronavirus cases have been reported worldwide. It has been realized that most researchers are still trying to develop an effective drug with fewer harmful Side Effects. The essential part of such investigation is to examine the COVID-19 drug side effects before recommending these drugs to treat the patients. Identifying potential side effects of the current COVID-19 drugs is a significant research problem in the drug development procedure because of the high priority of patients' health concerns (Das and Mazumder, 2023d, 2024, 2023c). Therefore, in treating COVID-19, studies regarding the prediction of COVID-19 drug side effects are essential.

The computational model is crucial in drug development, offering insights into unseen drugs and aiding in the reduction of complexity, time, and costs (Anand et al., 2022; Ahmad et al., 2023; Das and Mazumder, 2023a). Therefore, this paper proposes a novel approach based on neighbors, named K_1 K_2 Nearest Neighbors, to predict the side effects of COVID-19 drugs. The proposed model outperformed traditional models, including Decision Tree (DT), Random Forest (RF), Extra Tree Classifier (ETC), K Nearest Neighbor (KNN), Extreme Gradient Boosting (XGBoost), AdaBoost, Multi-layer Perceptron Neural Network (MLPNN), and Deep Neural Network (DNN), due to its capability to identify COVID-19 drug side effects based on both drug-drug similarity and SE-SE similarity.

In their work, Das et al. (2021) predicted drug side effects from drug function by employing various algorithms, including the decision tree, multi-layer perceptron neural network, extra tree classifier, k nearest neighbor, and random forest. The authors found that the extra tree

* Corresponding author.
E-mail address: dilwar2k4@yahoo.co.in (D.H. Mazumder).

https://doi.org/10.1016/j.compbiolchem.2024.108066
Received 5 December 2023; Received in revised form 12 March 2024; Accepted 1 April 2024
Available online 2 April 2024

classifier outperformed other classifier algorithms. In a different study, Das et al. (2022a) integrated drug function with 17 molecular properties and chemical 1D structure. They applied a deep neural network and found that a combination of chemical 1D structure and drug functions performed better than other combinations of drug properties. In another work, Jamal et al. (2017) predicted the side effects of neurological drugs from Biological (B), Phenotypic (P), and Chemical (C) properties of the drugs by applying Support Vector Machine (SVM). The authors used the combinations of two-level (B+C, P+C, B+P) and three-level (B+P+C) drug properties. They showed that combining phenotypic properties with chemical properties provided better neurological side effects prediction results than other combinations. In Wei et al. (2020), the authors predicted the risk level of a drug from side effects collected from patient health records. The data included report ID, report address, drug name, side effects name, gender, and age. They applied AdaBoost, Logistic Regression (LR), random forest, and XGBoost with a class imbalance approach Synthetic Minority Oversampling Technique (SMOTE).

Further, Uner et al. (2019) utilized Gene Ontology (GO), Chemical Structure (CS), Gene Expression (GEX), and META information of drugs to predict drug side effects. They applied deep neural network models such as multi-layer perceptron, multi-task, multi-model, residual variant, and convolutional neural network to each property and their two-level (GO+CS, GEX+CS) and three-level (GEX+CS+META) combination. They found that the multi-model neural network performs well when combining gene expression and chemical structure data. In Zhang et al. (2015), a feature selection-based multi-label KNN model was proposed for adverse drug reaction prediction from drug target and chemical information. In a different work Kanji et al. (2015), proposed a canonical optimization correlation model to predict phenotypic side effects from chemical profiles and drug target profiles. The authors found that their proposed model performed better on chemical profile data than drug target profiles. In Hatmal et al. (2021), the authors predicted the side effects of COVID-19 vaccination by applying machine learning techniques such as MLPNN, XGBoost, RF, and K-star from the demographic data. They found that the random forest machine learning approach performs better than the other algorithms.

The key contributions of this research study are as follows:

- Proposed a novel K_1K_2NN method based on neighbors to predict side effects of COVID-19 drugs from 17 molecules' drug descriptors and 1D chemical drug structure, which have not yet been utilized to predict adverse COVID-19 drug reactions.
- We also compared the performance of the proposed K_1K_2NN model with a multi-label deep neural network model, decision tree, extra tree classifier, k nearest neighbor, random forest, AdaBoost, multi-layer perceptron neural network, and XGBoost.

The framework of this research is as follows: Section 2 demonstrates the materials and methods used in this research work and the functional architecture to predict COVID-19 drug side effects. Further, the results of the experiments have been presented in Section 3. Section 4 discusses the findings of this research. Finally, Section 5 summarizes the conclusions of the research work.

2. Materials and methods

This section explains the materials and methods utilized to predict side effects of the COVID-19 drug from the chemical properties. Furthermore, the problem statement to predict the side effects of the COVID-19 drug and the working architecture for the proposed methodology to solve the stated problem have been presented.

2.1. Drug properties

The process of developing the dataset for the proposed methodology has been shown in Fig. 1. The drugs.com website lists several drugs that can be used to treat COVID-19 illness. Among those drugs, only the chemical properties of 20 COVID-19 drugs are listed in PubChem such as Remdesivir, Azithromycin, Umifenovir, Zyesami, Peginterferon Lambda, PF-07321332, Methylprednisolone, Lopinavir, Ritonavir, MK-4482, Ivermectin, Favipiravir, Hydroxychloroquine, Fluvoxamine, Dexamethasone, Baricitinib, Chloroquine phosphate, Bemcentinib, Hydroxychloroquine sulfate, and Protein kinase inhibitors (Kim et al., 2016). After collecting the drug names from Drugs.com (Link: https://www.drugs.com/condition/COVID-19.html?page_all=1), the chemical 1D structure and 17 molecular descriptors were collected from the PubChem repository. The PubChem CID was used as input in Python programming with the PubChemPy tool to collect chemical properties. Python provides the PubChemPy tool to collect data from PubChem. Furthermore, side effects for each drug were collected from WebMD BLACK (2024). On the WebMD website, 32 side effects are listed for 20 COVID-19 drugs. To prepare the dataset, we combined the COVID-19 drug chemical properties listed on PubChem with the labels of the corresponding COVID-19 drug side effects obtained from WebMD.

2.1.1. Description of the drug properties

The chemical feature of the drug is an essential property in drug development for predicting COVID-19 drug side effects. By predicting drug side effects based on chemical properties, researchers can gain insights into the potential toxicity of drugs and identify safe and effective treatment options. Drug properties play a significant role in predicting drug side effects. These properties can also affect the interaction of drugs with cellular targets and the formation of drug metabolites, which can cause adverse reactions. By understanding the relationship between drug properties and side effects, the current work identifies potential COVID-19 drug side effects.

- **Side Effects**: Side effects are the harmful reactions of drugs that are collected from WebMD BLACK (2024); Das and Mazumder (2023b). The current work consists 32 side effects, namely, Abdominal Cramps, Vomiting, Bad Taste, Weight Decreased, Upset Stomach, Abdominal Pain, Chills, Anaphylaxis, Angioedema, Blood Bilirubin Increased, Blood Pressure Increased, Constipation, Decreased Oxygen in the Tissues or Blood, Decreased Appetite, Diarrhea, Difficulty Sleeping, Drowsiness, Drug Fever, Headache, Heart Rate Irregular, Heartburn, Hypersensitivity, Kidney Function Abnormal, Liver Function Test Abnormal, Nausea, Rash, Seizures, Sore Mouth, Sweating Increased, Tired and Heavy, Trouble Breathing, and Vision Blurred are present in WebMD concerning those 20 drugs.

- **17 Molecules' Descriptors:** The 17 molecules' descriptors are collected from PubChem website Kim et al. (2016); Das and Hussain Mazumder (2021) using the PubChemPy Python package. The 17 molecules' descriptors are the chemical and physical computed properties of a chemical compound, such as Rotatable Bond Count, Undefined Atom Stereocenter Count, Formal Charge, Defined Bond Stereocenter Count, Monoisotopic Mass, Heavy Atom Count, Exact Mass, Hydrogen Bond Donor Count, Covalently-Bonded Unit Count, Undefined Bond Stereocenter Count, Complexity, XLogP3-AA, Isotope Atom Count, Hydrogen Bond Acceptor Count, Defined Atom Stereocenter Count, Topological Polar Surface Area, and Molecular Weight. After mapping the COVID-19 drug side effects with the descriptors of 17 molecules', the dataset consists of 20 samples with 17 input features corresponding to 32 drug side effects.

- **Chemical 1D Structure (SMILES):** The chemical 1D structure of drug molecules is also collected from PubChem Kim et al. (2016). It is a chemical notation of the 1D chemical structure of a drug. The conversion of this chemical 1D structure to a fingerprint is necessary. There are different methods for converting the chemical 1D structure of drugs into fingerprints. For this research work, Morgan circular fingerprints were chosen for 1D structure embedding due to their

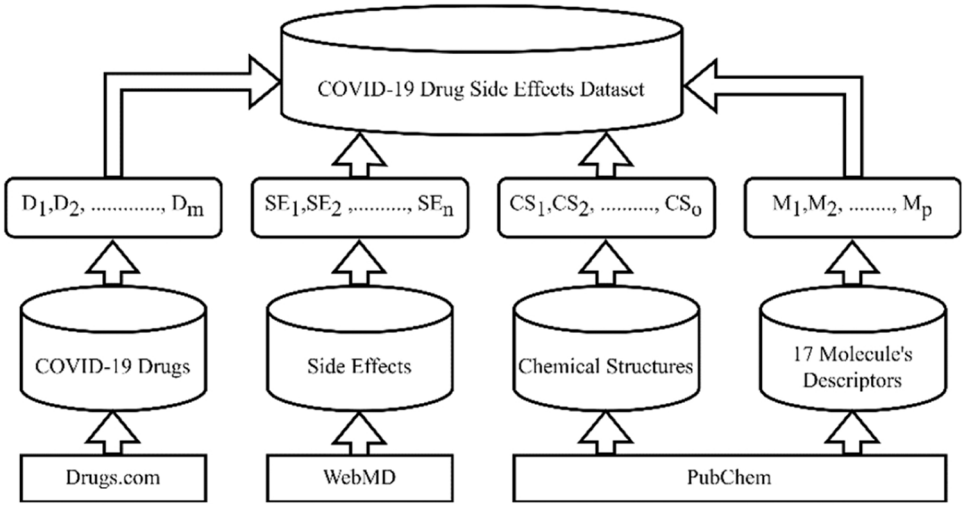


Fig. 1. Preparation process of the COVID-19 drugs side effects dataset.

excellent results reported by other researchers Das et al. (2022b). Each bit of the fingerprint corresponds to a predefined structure, represented by a string of 0's and 1's, where 1 indicates the presence of a specific substructure, and 0 indicates its absence. The length of the Morgan fingerprint is 1024 bits. Python provides an RDKIT tool to generate molecular fingerprints for chemical 1D structures. After mapping the chemical 1D structure with COVID-19 drug side effects, the dataset consists of 20 samples with 1024 chemical substructures corresponding to 32 drug side effects.

2.2. Problem statement

Let Drug = {D_1, D_2, … …, D_i, … …., D_m} be the set of drugs, where m denotes number of the COVID-19 drugs (m=20). Assume that F = {F_1, F_2, …, F_k, …, F_q} be the set of drug input feature, where q is number of drug features (q=1024 for 1D chemical structure and q=17 for 17 molecules' descriptors). Let SE = {SE_1, SE_2, ……., SE_j, ……., SE_n} be the set of side effects of the COVID-19 drug, where n is the number of COVID-19 drug side effects (n=32). A COVID-19 drug (D_i) can have more than one side effect. Therefore, predicting the COVID-19 drug side effects is a multi-label classification problem (Tai et al., 2022; Tan et al., 2022). Fig. 2 demonstrate the multi-label COVID-19 drug side effects prediction task, where zero denotes the absence of a COVID-19 drug side effect, and one denotes the presence of a COVID-19 drug side effect.

2.3. Working architecture

The diagrammatic presentation of the working architecture of the proposed methodology to predict multi-label COVID-19 drug side effects using chemical properties is shown in Fig. 3. Initially, the chemical properties are extracted from PubChem, and side effects are extracted from the WebMD website. After collecting data from the web, it was noticed that pre-processing is needed to handle the missing values in 17 molecules' descriptors properties. Missing values are one of the most common problems in the dataset, and they can significantly hamper our results. Firstly, for handling missing data, we can ignore the tuple that contains a missing value; secondly, fill up the missing values with zero; and lastly, fill up the missing values with the mean. Filling up missing values with zero achieved better results than ignoring the tuple and

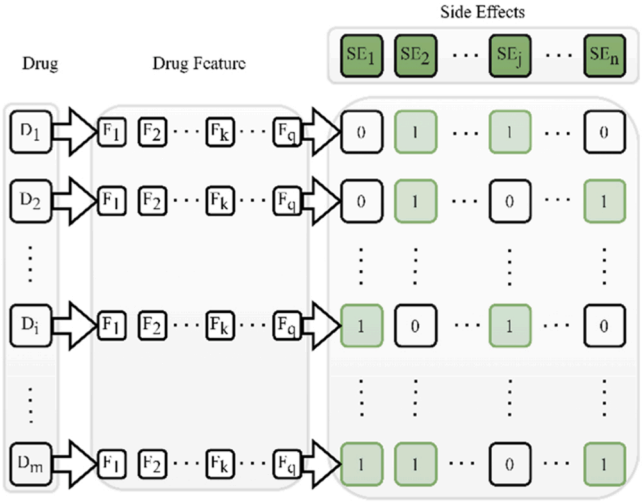


Fig. 2. Representation of multi-label COVID-19 drug side effects.

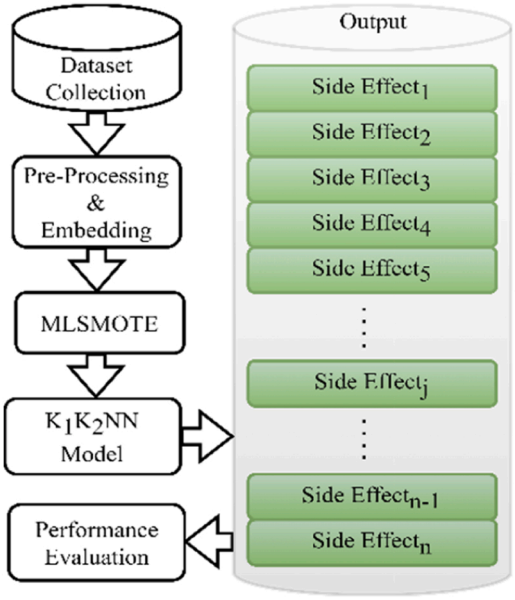


Fig. 3. Proposed model workflow to predict COVID-19 drug side effects.

filling up missing values with the mean. After handling missing data, embedding is required to convert the chemical 1D drug structure to fingerprint. Further, the class imbalance problem, or uneven class distribution, is addressed. When the available dataset is insufficient to describe the underlying phenomena across all classes adequately, the data under-representation problem arises (Liu et al., 2023). This kind of issue could degrade the classification algorithm's performance. It is crucial to address this kind of issue (Lázaro and Figueiras-Vidal, 2023). Therefore, to solve the issue of class imbalance in the datasets and enhance the performance of the classifiers, the Multi-label Synthetic Minority Over-sampling Technique (MLSMOTE) is used. There are other data augmentation techniques, such as SMOTE, Adaptive Synthetic (ADASYN), and SVM-SMOTE, however, these techniques are not executable for multi-label datasets, whereas the MLSMOTE technique has been specifically designed for handling multi-label datasets. Therefore, the MLSMOTE technique has been chosen over other sample strategies for the current investigation. By creating multi-label synthetic data samples based on the reference neighborhood of minority samples, MLSMOTE solves these problems.

For predicting COVID-19 drug side effects, two chemical drug properties - 17 molecules' and chemical 1D drug structure with nine multi-label machine learning approaches, RF, ETC, KNN, DT, Multi-Label Deep Neural Network (MLDNN), Multi-Layer Perceptron Neural Network (MLPNN), AdaBoost, XGBoost Peng et al. (2024, 2023) and K_1K_2NN (proposed model) have been applied. For each classifier, 17 molecules' descriptors and 1D chemical structure are given as input drug features to predict 32 unseen COVID-19 drug side effects. To evaluate the performance of all classifiers, Leave-One Out Cross-Validation (LOOCV) is utilized. To elaborate, LOOCV begins with the dataset containing 'x' samples, where 'x' is the total number of samples. We then train the classification model on the 'x-1′ training samples. This process is repeated for every sample in the dataset, resulting in x iterations of training and testing, leaving out one sample each time. Subsequently, we use the trained model to classify the single left-out sample. This process is repeated for every sample in the dataset, so if we have x samples, it will perform x iterations of training and testing, leaving out one sample each time and calculate the performance matrices in every iterations. Finally, it takes an average of all iteration for every performance matrices to output the final score for the performance metrics. Several methods are used to handle the multi-label machine learning classification problem, namely Classifier Chain (CC) Read et al. (2011), Binary Relevance (BR) Zhang et al. (2018), Multi Output Deep Neural Network (MODNN) Das et al. (2022a), and label powerset Junior et al. (2017). In the current work, BR, CC, and MODNN methods has been employed to solve the task of the multi-label classification problem.

2.4. Proposed model

A similarity-based and co-occurrence-based prediction model is essential in the COVID-19 drug side effects prediction task because it can identify potential side effects of a drug by comparing its chemical properties with known drugs that have similar properties. Additionally, it can identify possible side effects of a drug by analysing the co-occurrence patterns of side effects. With the chemical drug properties, using similarity-based and co-occurrence-based prediction models can help identify the most effective drug for further testing and development while also providing insights into potential side effects that may arise. By leveraging these principles, the K_1K_2NN classifier provides an innovative solution for predicting drug side effects that can aid in drug discovery and development. In the context of COVID-19 drug development, the proposed prediction models can be particularly useful for identifying rare or unexpected side effects that may not have been detected through clinical trials or other traditional approaches. This can help improve patient safety by identifying a drug's potential risks before it is widely used. The proposed model flowchart has been shown in Fig. 4. The K_1K_2NN model chooses the first K_1 neighbours to the test drug sample by calculating its similarity with the train drug samples. It then assigns the test sample with the SE label having the majority count on the SE labels of these K_1 neighbour drugs obtained through a voting mechanism. The model then calculates the SE-SE similarity using the Jaccard similarity measure from the SE co-occurrence values. Finally, the model chooses the most similar K_2 SE neighbours for those SE determined by the K_1 neighbour drugs and assigns these SE to that test drug sample. The algorithm for the proposed model is presented as Algorithm 1. The major steps of the K_1K_2NN nearest neighbours algorithm are described in the following subsections.

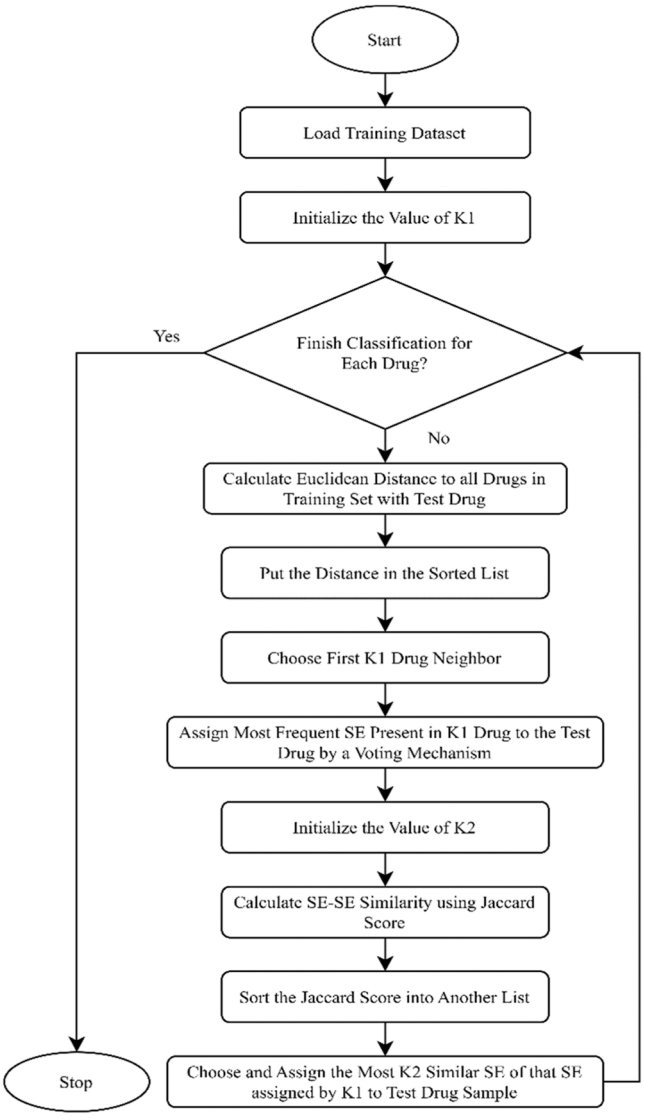


Fig. 4. The flowchart of the proposed model to predict COVID-19 drug side effects.

Algorithm 1. Pseudo code of K_1K_2NN algorithm

Input: COVID-19 drug side effects dataset
OutPut: Predicted COVID-19 drug side effects
1: Begin
2: Train_Drugs ∈ COVID-19 drug side effects dataset
3: for each $Train\_Drug_i$ in Train_Drugs do
4: Calculate ED (Test_Drug, $Train\_Drug_i$) ⊳ ED: Euclidean Distance
5: Store the computed ED in a list
6: end for
7: Organize the store ED in non-decreasing order
8: Assume that K_1 be a positive integer and choose the first K_1 distance from the organized list
9: Choose the first K_1 train drug samples corresponding to those K_1 distances
10: Label_the_Test_Drug based on the most frequent class present in the selected K_1 train drug samples
11: for each $label_i$ in Label_the_Test_Drug assigned by K_1 do
12: Calculate JS ($label_i$ with other SE labels) ⊳ JS: Jaccard Score
13: Store the computed JS in a list
14: end for
15: Organize the store JS in non-decreasing order
16: Assume that K_2 be a positive integer; and choose the first K_2 score from the sorted
17: Assign the labels to the Test_Drug based on the first K_2 similar labels score
18: Return COVID-19 drug side effects
19: End

2.4.1. Calculation of test drug and train drug similarity

Let $TrD = \{TrD_1, TrD_2, ., TrD_t\}$ define the set of t distinct training COVID-19 drugs and $TeD = \{TeD_1, TeD_2, ., TeD_s\}$ define the set of s distinct testing COVID-19 drugs. Let $P = \{p_1, p_2, \ldots\ldots, p_u\}$ represent the set of Train Drug (TrD) features and $Q = \{q_1, q_2, \ldots\ldots, q_u\}$ indicates the set of Test Drug (TeD) features where u is the total number of drug features. The Euclidean Distance (ED) is used to obtain the test drug similarity with train drugs from their chemical features with the help of the following equation.

$$ED(TeD, TrD) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_u - q_u)^2} \quad (1)$$

2.4.2. Choose the First K_1 drug neighbors

After calculating the Euclidean distances between train drug samples and test drug samples, store them in a list, and organize the Euclidean distances in non-decreasing order, and choose the first K_1 train drug samples corresponding to those K_1 distances from the sorted list. The pseudocode for choosing the first K_1 drug neighbors is shown in Algorithm 2.

Algorithm 2. Pseudocode to choose the first K_1 drug neighbours

1: Begin
2: List = []
3: K_1_Distance = []
4: K_1_Drug = []
5: for each TrD_i in TrD do
6: Calculate ED (TeD, TrD_i)
7: List = List.append (ED)
8: end for
9: Sort_List = Sort(List)
10: K_1 _Distance = K_1 _Distance.append (first K_1 distance from Sort_List)
11: K_1 _Drug = K_1_Drug.append (Select K_1 drug neighbors corresponding to those K_1_Distance
12: Return K_1 drug neighbors
13: End

2.4.3. Assign maximum side effects class of K_1 to test drug sample

In this step, the test drug is labeled based on the most frequent class present in the K_1 selected drug samples. The pseudocode for the assigned maximum SE class of K_1 to the test drug sample is shown in Algorithm 3.

Algorithm 3. Pseudo code for assigned maximum side effects class of K_1 to test drug sample

```
1: Begin
2: for each class of K1_Drug do
3:        CPC = Count (Positive Class)          ▷ CPC : Count Positive Class
4:        CPN = Count (Negative Class)          ▷ CNC : Count Negative Class
5:        If CPC > CPN then
6:              Assign the label of TeD as 1    ▷ 1 represents presence of a drug side effects
7:        Else
8:              Assign the label of TeD as 0    ▷ 0 represents absence of a drug side effects
9:        end if
10: end for
11: Return most frequent side effects of K1
12: End
```

2.4.4. Calculation of side effects co-occurrence

Let SE = {SE_1, SE_2, …, SE_n} define the set of *n* distinct side effects of COVID-19 drugs. For any two side effects SE_i and SE_j, the co-occurrence of the COVID-19 drug side effects is calculated using the following equations.

$$\text{Co}-\text{occurrence}\left(SE_i, SE_j\right) = \sum_{k=0}^{t} f(k, TrD) \tag{2}$$

$$f(k, TrD) = \begin{cases} 1, & \text{if} \quad SE_{i\in \text{TrD}_k} \cap SE_{j\in \text{TrD}_k} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

In the above equations, *t* represents the number of drugs in the set of TrD = {TrD_1, TrD_2, ……, TrD_t}.

2.4.5. Calculation of SE-SE similarity

The similarity score between two side effects is calculated according to the Jaccard Similarity (JS) measure score based on the co-occurrence value.

$$\text{Jaccard\_Score}\ (\text{SE}_i, \text{SE}_j) = \frac{\text{V}_i \cap \text{V}_j}{\text{V}_i \cup \text{V}_j} \tag{4}$$

Where V_i and V_j denote the co-occurrence value of side effects SE_i and SE_j, respectively.

2.4.6. Choose the K_2 most similar SE of the K_1 neighbor drugs and assign to test drug sample

Sort the Jaccard similarity score in another list calculated from the co-occurrence of side effects. Choose the first K_2 similar side effects for each of the side effects associated with the K_1 neighbor drugs and assign these K_2 side effects to the test drug sample. The pseudocode for choosing the most K_2 similar side effects assigned by K_1 and labeled to test drug samples is shown in Algorithm 4.

Algorithm 4. Choose the most K_2 similar side effects assigned by K_1 and labeled them to test drug

```
1: Begin
2: Labels_Similarity = []
3: K2_Label_Similarity = []
4: Calculate co-occurrence matrix of training samples labels using equation (2) and (3)
5: for each TeD_Label assigned by K1 do        ▷TeD_Label: Test Drug Labels assigned by K1
6:        Calculate JS (TeD_Label with other Labels)
7:        Labels_Similarity = Labels_Similarity.append(JS)
8: end for
9: Sort_Labels_Similarity = Sort(Labels_Similarity)
10: K2_Label_Similarity=K2_Label_Similarity.append (first K2 similarity score from Sort_Label_Similarity)
11: Assign the class to the Test Drug (TeD) based on the K2 neighbors class corresponding to the first K2
similar label score in K2_Label_Similarity.
12: Return first K2 side effects
13: End
```

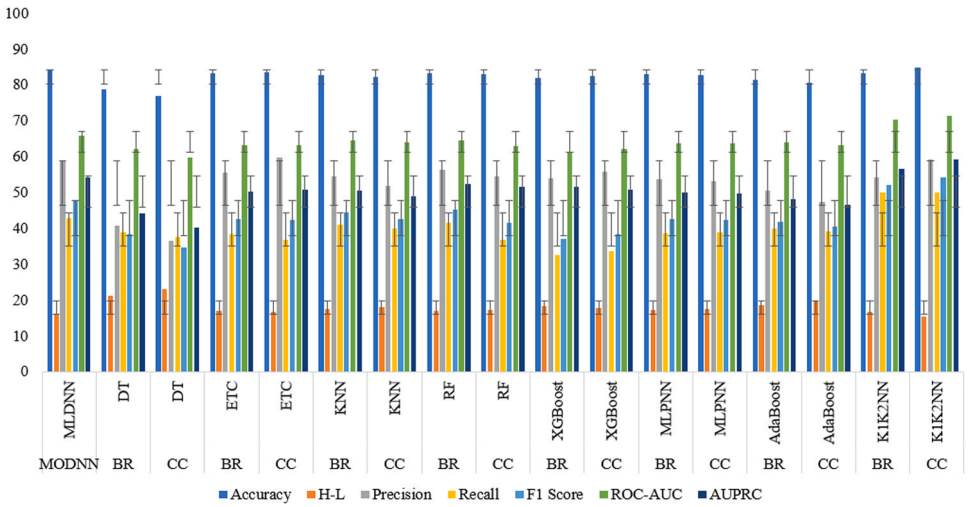


Fig. 5. Results of classifiers applied on the chemical 1D structure to predict the side effects of the COVID-19 drug.

3. Results of the experiments

The results of the experiments have been discussed in this section. The Multi-Label (ML) problem is addressed by employing the multi-output deep neural network, binary relevance, and classifier chain methods with eight multi-label supported classifiers applied to evaluate the research work, and a novel classification approach has been proposed to predict COVID-19 drug side effects. The proposed model performance is better than the state-of-the-art multi-label machine learning classification approaches. The performance of the K_1K_2NN model has been analysed and compared with the decision tree, multi-label deep neural network, k-nearest neighbours, extra tree classifier, and random forest in terms of ROC-AUC, accuracy, precision, hamming-loss, recall, F1-score, and Area Under Precision-Recall Curve (AUPRC). The overall performance of the experiments on chemical 1D structure and 17 molecules' descriptors has been shown in Fig. 5, Fig. 6, Fig. 7, and Fig. 8, respectively.

3.1. Results of classifiers on chemical 1D structure

This section summarizes the results of all classifiers (MLDNN, DT, ETC, KNN, RF, MLPNN, XGBoost, AdaBoost, and K_1K_2NN used to predict side effects of the COVID-19 drug from chemical 1D structure. From Fig. 5, it can be observed that the presented K_1K_2NN model with the CC method performed better than the BR method. The K_1K_2NN model achieved the highest accuracy of 84.68%, a recall value of 50%, an F1 score of 54.20%, a ROC-AUC of 71.18%, and an AUPRC score of 59.12%, and the lowest hamming-loss of 15.31%. For precision, The K_1K_2NN model achieved 59.18%. Compared to the other multi-label supported machine learning classifiers, the proposed model performs well on the chemical 1D structure. On the other hand, the extra tree classifier reported the highest precision value of 59.75% on the CC ML approach.

3.2. Results of classifiers on 17 molecule's descriptors

The following section provides an overview of the results obtained from all classifiers used to predict the side effects of the COVID-19 drug based on 17 molecular descriptors. The performance of the K_1K_2NN model on both the BR and CC ML approaches is the same on the 17 molecules' descriptors to predict COVID-19 drug side effects. The K_1K_2NN model achieved higher accuracy, ROC-AUC, F1, and AUPRC scores of 84.53%, 69.75%, 52.17%, and 57.79%, respectively, as shown in Fig. 6 and the lowest hamming-loss of 15.46% compared to the other classifiers. For precision and recall, the K_1K_2NN model achieved 59.34% and 46.55%, respectively. On the other hand, the KNN classifiers achieved the highest precision of 60.58% and a recall value of 47.69% compared to the other classifiers on both ML approaches. The performance of the other classifiers can be observed from Fig. 6 on the 17 molecular descriptors to predict side effects of the COVID-19 drug.

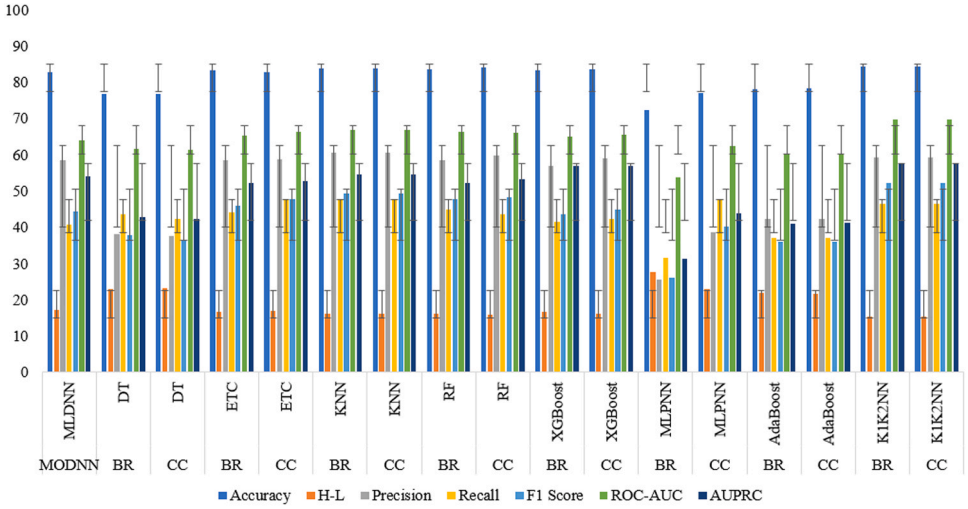


Fig. 6. Results of the classification methods on the 17 molecule's descriptors to predict COVID-19 drug side effects.

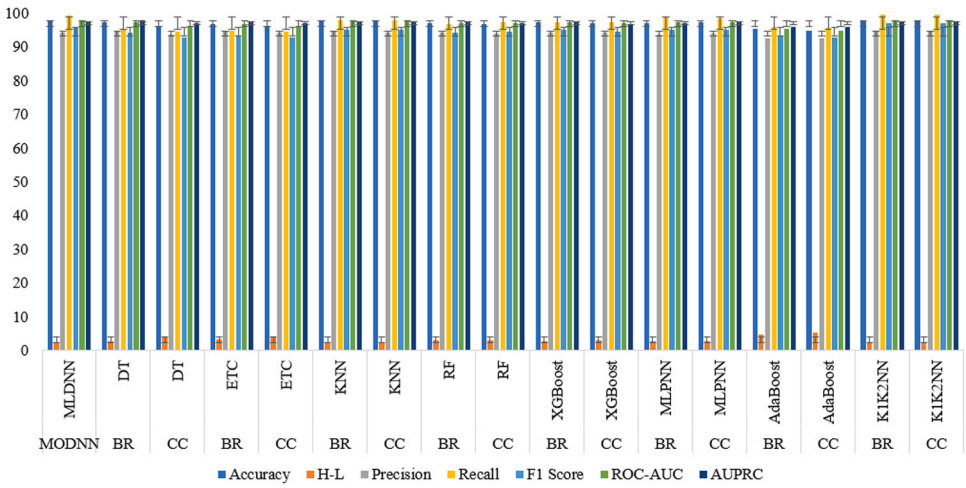


Fig. 7. Results of classifiers with MLSMOTE on the chemical 1D structure to predict COVID-19 drug side effects.

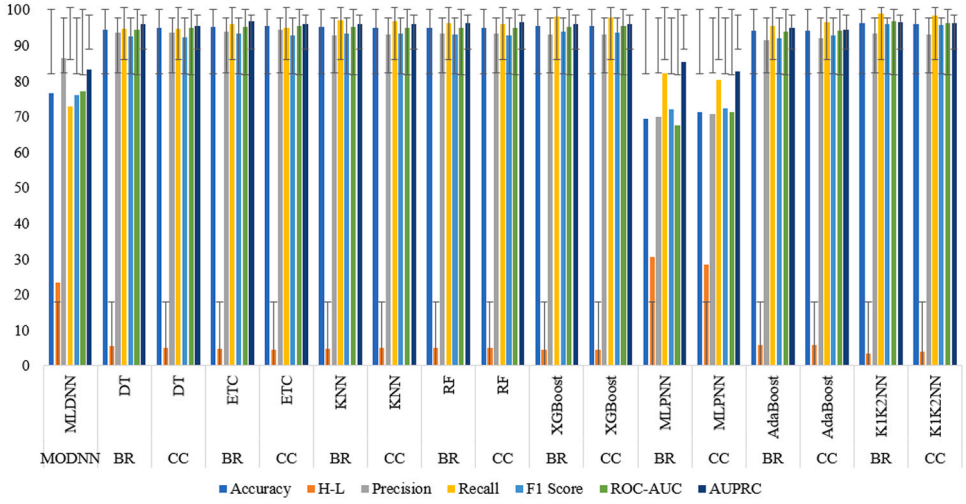


Fig. 8. Results of the classification methods with MLSMOTE on the 17 molecule's descriptors to predict COVID-19 drug side effects.

3.3. Results of classifiers on chemical 1D structure with MLSMOTE

In this section, we will outline the outcomes obtained after implementing MLSMOTE on all classifiers used to predict the side effects of the COVID-19 drug based on chemical 1D structure. The uneven distribution (class imbalance) of COVID-19 drug instances across different side effects poses a challenge, which is magnified due to the joint appearance of minority and majority side effects for COVID-19 drug instances. It can be observed from Fig. 7 and Fig. 8 that after addressing the class imbalance issue with the help of MLSMOTE, the classifiers can learn more appropriately to predict the COVID-19 drug side effects. Additionally, improvements can be noticed in accuracy, precision, F1 score, hamming-loss, recall, ROC-AUC, and AUPRC score. The K_1K_2NN model performs better on BR ML approaches. It achieved the highest accuracy score of 97.53%, precision value of 94.72%, F1 score of 97.01%, recall value of 99.41%, ROC-AUC score of 97.84%, and the lowest hamming-loss of 2.46%. For the AUPRC score, the proposed model reported 97.25%. Additionally, the DT method performed better using the BR approach, achieving the highest AUPRC score of 97.80%.

3.4. Results of classifiers on 17 molecule's descriptors with MLSMOTE

This section will outline the results obtained from applying MLSMOTE to all classifiers used for predicting the side effects of the COVID-19 drug based on 17 molecular descriptors. The performance of the classifiers' methods on 17 molecular descriptors with MLSMOTE can be observed from Fig. 8. The presented K_1K_2NN model with the BR method has performed better than the CC method. It achieved the highest accuracy score of 96.42%, a recall score of 98.96%, an F1 value of 96.03%, an ROC-AUC score of 96.70%, and the lowest hamming loss of 3.57%. For precision, the K_1K_2NN model achieved 93.27%, and the proposed model reported an AUPRC of 96.46%. On the other hand, the ETC classifier reported the highest precision score of 94.35% on the CC method and a 96.70% AUPRC score on the BR method. The effectiveness of the remaining classifiers can be seen in Fig. 8, where 17 molecular descriptors are utilized to predict the side effects of the COVID-19 drug.

3.5. ROC-AUC curves of the proposed models on the chemical properties

It can be observed from Fig. 9*(a)* that the proposed K_1K_2NN model's performance is promising after employing MLSMOTE in terms of ROC-AUC score, achieving the best performance on chemical 1D structure drug properties with the highest ROC-AUC score of 97.84%. On the other hand, the K_1K_2NN model achieved a ROC-AUC score of 96.70% on 17 molecular descriptors after employing MLSMOTE, as shown in Fig. 9 *(b)*.

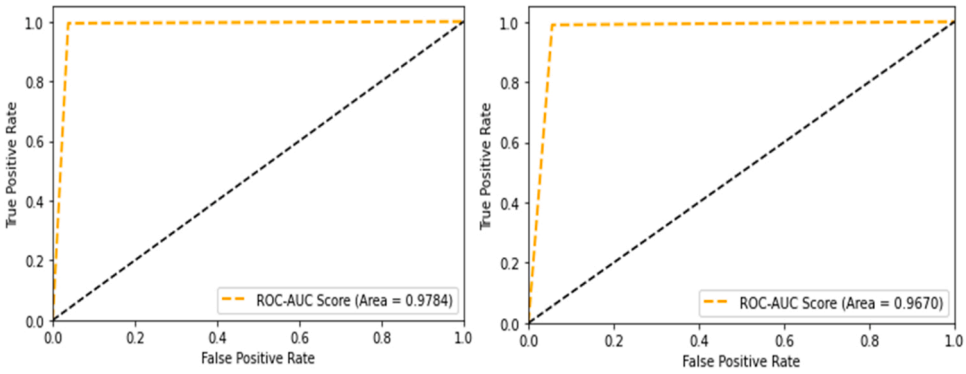


Fig. 9. ROC-AUC curve of K_1K_2NN model on the chemical 1D structure and 17 molecular descriptors.

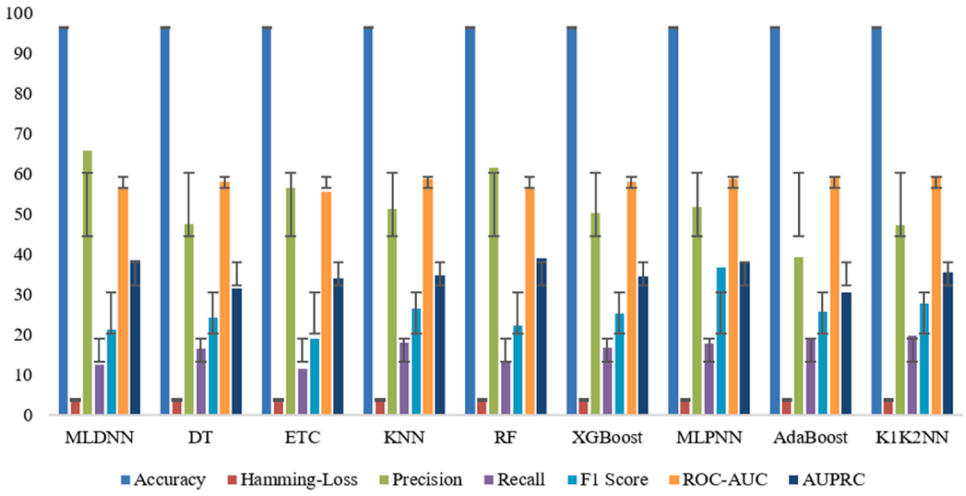


Fig. 10. Results of the classification methods on the gene expression signature dataset to predict drug side effects.

3.6. Results of classifiers on gene expression signature dataset

The data used in this work consist of gene expression profiles measuring changes in the expressions of 978 genes before and after the treatment of individual cells with 20,339 molecules'. Gene expression signatures capture biological information in a computational format and play a vital role in drug discovery for achieving drug targets. These profiles were generated under the Library of Integrated Network-based Cellular Signature (LINCS) project and are publicly available on the Ma'ayanlab website (Link: http://maayanlab.net/SEP-L1000/#download) (Wang et al., 2016). These profiles are used to evaluate the effects of drug-like molecules' on cellular response by analyzing changes in gene expression patterns. On the Ma'ayanlab website, there are 3166 side effects available corresponding to 834 drugs. After mapping the gene expression signature with side effects, the final dataset consists of 791 drugs with 978 genes corresponding to 2881 side effects.

The performance of the proposed model and machine learning classifiers has been analyzed using binary relevance approaches to solve the multi-label task for the gene expression signature dataset. In the gene expression dataset, the proposed model K_1K_2NN and other classification models were trained on a randomly selected 80% of the dataset, and the classification models' performance was evaluated on the remaining 20%. After training and testing the models, the performance of the proposed model and all other models is shown in Fig. 10 based on accuracy, hamming-loss, precision, F1 score, recall, ROC-AUC, and AUPRC score. These metric values indicate that the proposed model K_1K_2NN performed well on the gene expression signature dataset. The model achieved high accuracy in predicting drug efficacy and was able to identify potential side effects of drugs, which could help in designing safer drugs. The proposed model K_1K_2NN achieved the highest accuracy score of 96.54%, recall score of 19.69%, ROC-AUC value of 59.46%, and the lowest hamming-loss of 3.45%. It also achieved a better F1 score of 27.78% compared to the multi-label deep neural network model, decision tree, extra tree classifier, k nearest neighbor, random forest, AdaBoost, and XGBoost classifier models. The proposed model also achieved a better precision score of 47.16% compared to the AdaBoost classifier. The K_1K_2NN model also reported a better AUPRC value of 35.42% compared to the DT, ETC, KNN, XGBoost, and AdaBoost classifiers. On the other hand, the MLDNN model achieved the highest precision of 65.58% on the gene expression dataset compared to the other classifier models, and MLPNN achieved the highest F1 score of 36.64% on the gene expression dataset compared to K_1K_2NN, multi-label deep neural network model, decision tree, extra tree classifier, k nearest neighbor, random forest, AdaBoost, and XGBoost classifier models. In contrast, the RF model exhibited the highest AUPRC of 38.98% on the gene expression dataset, surpassing the AUPRC scores of the other classifier models. In conclusion, the proposed model has been successfully applied to gene expression signature datasets, which helped to evaluate its performance and ensure its accuracy and robustness.

4. Discussions

K_1K_2NN is a novel method that combines similarity-based and co-occurrence-based prediction models. The performance of K_1K_2NN

depends on the values chosen for the hyperparameters K_1 and K_2. The K_1K_2NN model selects the first K_1 neighbours to the test drug sample by calculating their similarity with the train drug samples. It then assigns the test sample with the SE label that has the majority count on the SE labels of these K_1 neighbour drugs, obtained through a voting mechanism. The model then calculates the SE-SE similarity using the Jaccard similarity measure from the SE co-occurrence values. Finally, the model chooses the K_2 most similar SE neighbours for those SEs determined by the K_1 neighbour drugs and assigns these SEs to the test drug sample. In general, K_1K_2NN can work well in situations where the labels have co-occurrence (like the side effects of drugs) and samples are similar (like COVID-19 drugs), as it can help improve prediction performance. The proposed K_1K_2NN based study achieved good results because the authors carefully tuned the hyperparameters K_1 and K_2, or because the dataset had characteristics that made K_1K_2NN a good fit. We also compared the performance of the proposed K_1K_2NN model with a multi-label deep neural network model, decision tree, extra tree classifier, k-nearest neighbour, random forest, AdaBoost, multi-Layer perceptron neural network, and XGBoost. The proposed model (K_1K_2NN) also outperformed state-of-the-art machine learning algorithms performance. Additionally, the proposed model also performed well after employing MLSMOTE to overcome the issue of the uneven distribution of COVID-19 drug instances across different side effects. Overall, the performance of K_1K_2NN depends on the specific problem and the dataset being used.

5. Conclusion

The pharmaceutical drug discovery and development process is challenging, complex, tedious, and costly. Accurate COVID-19 drug side effects predictions are essential for successfully developing and designing COVID-19 drugs. Adverse side effects are one of the main reasons for the failure of COVID-19 drugs, and the appearance of harmful reactions can halt the drug development process. Therefore, a reliable computational model is urgently needed to predict COVID-19 drug side effects, which can reduce design complexity, time, and cost in the drug development process. In this research study, we proposed a neighbour-based model named K_1K_2NN for predicting COVID-19 drug side effects using the chemical properties of drugs. The proposed model utilized 17 molecular descriptors and chemical 1D structures of drugs extracted from PubChem and side effects information from WebMD to develop the model. The performance of the K_1K_2NN model demonstrated that the chemical properties of drugs are sufficient for predicting COVID-19 drug side effects. Moreover, we successfully applied the proposed model to gene expression signature datasets, providing valuable insights into its performance evaluation, accuracy, and robustness. The proposed model's strengths lie in its ability to perform well at predicting class labels (side effects) when the samples (drugs) share similar input features (drug properties), and the labels (side effects) co-occur. However, the model may perform poorly if the dataset has fewer co-occurrence labels. The K_1K_2NN model can benefit researchers and practitioners in various ways. It provides a novel multi-label classification approach to predict drug side effects using nearest neighbours and the co-occurrence of side effects, which can improve the prediction model's performance. Researchers can also investigate the potential applications of the K_1K_2NN model in other areas of drug discovery and development, such as adverse drug reaction detection, drug function prediction, anatomical therapeutic chemical class identification, drug-target prediction, and associated drug-drug interaction, etc. Furthermore, other drug properties such as gene expression profiles, protein sequences (amino acids), drug functions, protein functions, drug targets, chemical 3D structures, and protein targets can be utilized to predict COVID-19 drug side effects. The proposed approach (K_1K_2NN classifier) can be tested for its generalizability to other diseases and drug classes beyond COVID-19. The proposed model can also effectively investigate COVID-19 drug side effects in drug discovery, design, and development.

CRediT authorship contribution statement

Pranab Das: Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Dr. Dilwar Hussain Mazumder:** Writing – review & editing, Validation.

Declaration of Competing Interest

We know that no conflicts of interest associated with this publication, and there has been no significant financial support for this work that could have influenced its outcome.

Acknowledgements

None.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.compbiolchem.2024.108066.

References

Abdalla, M., Rabie, A.M., 2023. Dual computational and biological assessment of some promising nucleoside analogs against the covid-19-omicron variant. Comput. Biol. Chem. 104, 107768.

Ahmad, I., Khan, H., Serdaroğlu, G., 2023. Physicochemical properties, drug likeness, admet, dft studies, and in vitro antioxidant activity of oxindole derivatives. Comput. Biol. Chem. 104, 107861.

Ahmad, Z., El-Kafrawy, S.A., Alandijany, T.A., Giannino, F., Mirza, A.A., El-Daly, M.M., Faizo, A.A., Bajrai, L.H., Kamal, M.A., Azhar, E.I., 2022. A global report on the dynamics of covid-19 with quarantine and hospitalization: a fractional order model with non-local kernel. Comput. Biol. Chem. 98, 107645.

Alici, H., Tahtaci, H., Demir, K., 2022. Design and various in silico studies of the novel curcumin derivatives as potential candidates against covid-19-associated main enzymes. Comput. Biol. Chem. 98, 107657.

Anand, P.K., Kumar, A., Saini, A., Kaur, J., 2022. Mutation in eth a protein of mycobacterium tuberculosis conferred drug tolerance against enthinoamide in mycobacterium smegmatis mc2155. Comput. Biol. Chem. 98, 107677.

Atifa, A., Khan, M.A., Iskakova, K., Al-Duais, F.S., Ahmad, I., 2022. Mathematical modeling and analysis of the sars-cov-2 disease with reinfection. Comput. Biol. Chem. 98, 107678.

BLACK, S., 2024. Uses, side effects, interactions and warnings–webmd.

Das, P., Hussain Mazumder,D.,2021. Predicting anatomical therapeutic chemical drug classes from 17molecules'properties of drugs by multi-label binary relevance approach with mlsmote, in: 2021 5th International conference on computational biology and bioinformatics, pp. 1–7.

Das, P., Mazumder, D.H., 2023c. Identify unfavorable covid medicine reactions from the three-dimensional structure by employing convolutional neural network. In: Mathematical Modeling and Intelligent Control for Combating Pandemics. Springer, pp. 155–167.

Das, P., Mazumder, D.H., 2023b. An extensive survey on the use of supervised machine learning techniques in the past two decades for prediction of drug side effects. Artif. Intell. Rev. 1–28.

Das, P., Mazumder, D.H., 2023a. Advances in predicting drug functions: a decade-long survey in drug discovery research. IEEE Transactions on Molecular, Biological and Multi-Scale Communications.

Das, P., Mazumder, D.H., 2023d. Mlcnn-cov: a multilabel convolutional neural network-based framework to identify negative covid medicine responses from the chemical three-dimensional conformer. ETRI Journal.

Das, P., Mazumder, D.H., 2024. Inceptionv3-lstm-cov: a multi-label framework for identifying adverse reactions to covid medicine from chemical conformers based on inceptionv3 and long short-term memory. ETRI Journal.

Das, P., Pal, V., et al., 2022a. Integrative analysis of chemical properties and functions of drugs for adverse drug reaction prediction based on multi-label deep neural network. J. Integr. Bioinforma.

Das, P., Sangma, J.W., Pal, V., et al., 2021. Predicting adverse drug reactions from drug functions by binary relevance multi-label classification and mlsmote. in: International Conference on Practical Applications of Computational Biology & Bioinformatics. Springer, pp. 165–173.

Das, P., Thakran, Y., Anal, S.N., Pal, V., Yadav, A., 2022b. Brmcf: binary relevance and mlsmote based computational framework to predict drug functions from chemical and biological properties of drugs. IEEE/ACM Trans. Comput. Biol. Bioinforma.

Hatmal, M.M., Al-Hatamleh, M.A., Olaimat, A.N., Hatmal, M., Alhaj-Qasem, D.M., Olaimat, T.M., Mohamud, R., 2021. Side effects and perceptions following covid-19 vaccination in jordan: a randomized, cross-sectional study implementing machine learning for predicting severity of side effects. Vaccines 9, 556.

Jamal, S., Goyal, S., Shanker, A., Grover, A., 2017. Predicting neurological adverse drug reactions based on biological, chemical and phenotypic properties of drugs using machine learning models. Sci. Rep. 7, 1–12.
Junior, J.C., de Faria Paiva, E., de Andrade Silva, J., Cerri, R., 2017. Label powerset for multi-label data streams classification with concept drift. Proc. 5th Symp. Knowl. Discov., Min. Learn 97–104.
Kanji, R., Sharma, A., Bagler, G., 2015. Phenotypic side effects prediction by optimizing correlation with chemical and target profiles of drugs. Mol. Biosyst. 11, 2900–2906.
Kim, S., Thiessen, P.A., Bolton, E.E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B.A., et al., 2016. Pubchem substance and compound databases. Nucleic Acids Res. 44, D1202–D1213.
Lázaro, M., Figueiras-Vidal, A.R., 2023. Neural network for ordinal classification of imbalanced data by minimizing a bayesian cost. Pattern Recognit., 109303
Liu, Y., Liu, Y., Bruce, X., Zhong, S., Hu, Z., 2023. Noise-robust oversampling for imbalanced data classification. Pattern Recognit. 133, 109008.
Peng, L., Xiong, W., Han, C., Li, Z., Chen, X., 2023. Celldialog: a computational framework for ligand-receptor-mediated cell-cell communication analysis iii. IEEE Journal of Biomedical and Health Informatics.
Peng, L., Huang, L., Su, Q., Tian, G., Chen, M., Han, G., 2024. Lda-vghb: identifying potential lncrna–disease associations with singular value decomposition, variational graph auto-encoder and heterogeneous newton boosting machine. Brief. Bioinforma. 25, bbad466.
Read, J., Pfahringer, B., Holmes, G., Frank, E., 2011. Classifier chains for multi-label classification. Mach. Learn. 85, 333–359.
Tai, M., Kudo, M., Tanaka, A., Imai, H., Kimura, K., 2022. Kernelized supervised laplacian eigenmap for visualization and classification of multi-label data. Pattern Recognit. 123, 108399.
Tan, A., Liang, J., Wu, W.Z., Zhang, J., 2022. Semi-supervised partial multi-label classification via consistency learning. Pattern Recognit. 131, 108839.
Uner, O.C., Cinbis, R.G., Tastan, O., Cicek, A.E., 2019. Deepside: a deep learning framework for drug side effect prediction. Biorxiv, 843029.
Wang, Z., Clark, N.R., Ma'ayan, A., 2016. Drug-induced adverse events prediction with the lincs l1000 data. Bioinformatics 32, 2338–2345.
Wei, J., Lu, Z., Qiu, K., Li, P., Sun, H., 2020. Predicting drug risk level from adverse drug reactions using smote and machine learning approaches. IEEE Access 8, 185761–185775.
Zhang, M.L., Li, Y.K., Liu, X.Y., Geng, X., 2018. Binary relevance for multi-label learning: an overview. Front. Comput. Sci. 12, 191–202.
Zhang, W., Liu, F., Luo, L., Zhang, J., 2015. Predicting drug side effects by multi-label learning and ensemble learning. BMC Bioinforma. 16, 1–11.

Pranab Das is currently pursuing a Ph.D. degree from the National Institute of Technology Nagaland, India, in Computer Science & Engineering with a major area of research as bioinformatics. His research interests include Computational Biology and Bioinformatics, Drug Discovery and Design, Machine Learning, Data Mining, and Deep Learning. He has published several research papers in reputed journals and prestigious conferences, found in

Dr. Dilwar Hussain Mazumder received the B.E. and M.Tech. degree in Computer Science and Engineering from Jorhat Engineering College, Assam, India, and Rajiv Gandhi University, Arunachal Pradesh, India, in 2008 and 2012, respectively. He pursued the Ph.D. degree in Computer Science and Engineering from National Institute of Technology Nagaland, India, in 2019. His research interests include computational methods for gene selection in cancer prediction, such as biogeography-based optimizers, particle swarm optimizers, hybrid approaches, genetic algorithms, neural networks, data analytics, and bioinformatics with emphasis to drug discovery. He has published several research papers in reputed journals and prestigious conferences. He is having more than 13 years of experience in Teaching and Research. Currently, he is working as an Assistant Professor at the Department of Computer Science and Engineering, National Institute of Technology Nagaland, India.