information

MDPI

Article

Chem2Side: A Deep Learning Model with Ensemble Augmentation (Conventional + Pix2Pix) for COVID-19 Drug Side-Effects Prediction from Chemical Images

Muhammad Asad Arshed [1,*], Muhammad Ibrahim [1], Shahzad Mumtaz [2], Muhammad Tanveer [3] and Saeed Ahmed [4]

[1] Department of Computer Science, The Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan; muhammad.ibrahim@iub.edu.pk

[2] Department of Data Science, The Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan; shahzad.mumtaz@iub.edu.pk

[3] School of Systems and Technology, University of Management and Technology, Lahore 54770, Pakistan; muhammad_tanveer@umt.edu.pk

[4] Department of Experimental Medical Science, Biomedical Center (BMC), Lund University, 22184 Lund, Sweden; saeed.ahmed@med.lu.se

* Correspondence: muhammadasadarshed0900@gmail.com; Tel.: +92-3060771040

Abstract: Drug side effects (DSEs) or adverse drug reactions (ADRs) are a major concern in the healthcare industry, accounting for a significant number of annual deaths in Europe alone. Identifying and predicting DSEs early in the drug development process is crucial to mitigate their impact on public health and reduce the time and costs associated with drug development. **Objective:** In this study, our primary objective is to predict multiple drug side effects using 2D chemical structures, especially for COVID-19, departing from the conventional approach of relying on 1D chemical structures. We aim to develop a novel model for DSE prediction that leverages the CNN-based transfer learning architecture of ResNet152V2. **Motivation:** The motivation behind this research stems from the need to enhance the efficiency and accuracy of DSE prediction, enabling the pharmaceutical industry to identify potential drug candidates with fewer adverse effects. By utilizing 2D chemical structures and employing data augmentation techniques, we seek to revolutionize the field of drug side-effect prediction. **Novelty:** This study introduces several novel aspects. The proposed study is the first of its kind to use 2D chemical structures for predicting drug side effects, departing from the conventional 1D approaches. Secondly, we employ data augmentation with both conventional and diffusion-based models (Pix2Pix), a unique strategy in the field. These innovations set the stage for a more advanced and accurate approach to DSE prediction. **Results:** Our proposed model, named CHEM2SIDE, achieved an impressive average training accuracy of 0.78. Moreover, the average validation and test accuracy, precision, and recall were all at 0.73. When evaluated for COVID-19 drugs, our model exhibited an accuracy of 0.72, a precision of 0.79, a recall of 0.72, and an *F*1 score of 0.73. Comparative assessments against established transfer learning and machine learning models (VGG16, MobileNetV2, DenseNet121, and KNN) showcased the exceptional performance of CHEM2SIDE, marking a significant advancement in drug side-effect prediction. **Conclusions:** Our study introduces a groundbreaking approach to predicting drug side effects by using 2D chemical structures and incorporating data augmentation. The CHEM2SIDE model demonstrates remarkable accuracy and outperforms existing models, offering a promising solution to the challenges posed by DSEs in drug development. This research holds great potential for improving drug safety and reducing the associated time and costs.

Keywords: COVID-19; medicine; drug side effects; image classification; chemical structure images; stable diffusion; Pix2Pix; machine learning; deep learning; transfer learning

check for updates

Citation: Arshed, M.A.; Ibrahim, M.; Mumtaz, S.; Tanveer, M.; Ahmed, S. Chem2Side: A Deep Learning Model with Ensemble Augmentation (Conventional + Pix2Pix) for COVID-19 Drug Side-Effects Prediction from Chemical Images. *Information* **2023**, *14*, 663. https://doi.org/10.3390/info14120663

Academic Editor: Jose Santamaria

Received: 9 November 2023
Revised: 15 December 2023
Accepted: 15 December 2023
Published: 16 December 2023

1. Introduction

Drugs are widely used to treat various medical conditions and diseases, and while they can be very effective, they also carry the risk of side effects. Side effects are unintended and usually undesired effects that can occur when a medication is used. These effects can range from mild to severe, impacting the patient's quality of life and, in some cases, leading to serious health problems. Understanding drug side effects is essential for healthcare providers and patients to make informed decisions about medication use. The side-effect profile of a drug is an important consideration when prescribing or administering medication.

Side effects can occur for various reasons, such as an overdose, drug interactions, individual susceptibility, genetic variations, and adverse drug reactions (ADRs) to the drug's components. Drug side effects can be categorized into different types based on their severity, duration, and mechanism of action, e.g., typical side effects, serious side effects, long-term side effects, and interactions with other drugs.

DSEs significantly impact public health and drug discovery costs [1,2] and can lead to morbidity and mortality. Drug side effects are reported as one of the leading causes of death in the United States [3].

Pharmaceutical companies must conduct clinical trials to evaluate the safety and efficacy of their drugs before they are approved for use by regulatory agencies such as the U.S. Food and Drug Administration (FDA) [4]. During these trials, pharmaceutical companies monitor participants for adverse reactions or side effects that may appear in humans due to the drug. Once a drug is approved, pharmaceutical companies continue to monitor its safety through post-marketing surveillance. The traditional methods to predict DSEs require several clinical trials and monitoring after drug release in the market [5]. This involves tracking adverse event reports from healthcare providers, patients, and other sources and investigating potential safety concerns. The computational approach can mitigate the burden on pharmaceutical companies and public health and reduce drug development costs with an early prediction of DSEs.

Several computational methods have been proposed to predict DSEs at different stages of drug development [6,7]. In recent decades, proposed computational models have evolved for the prediction of DSEs from similarity-based methods [8] to machine learning models, e.g., support vector machines (SVMs) [9], clustering [10], more complex predictors that are based on random forests (RFs) [10], and deep learning (DL) [11].

The existing models are based on the assumption that similar drugs have the same properties in terms of their biological and chemical features. For the prediction of DSEs, Pauwels et al. [12] proposed a chemical structure-based model using sparse-canonical correlation analysis (SCCA), and Yamanishi et al. proposed a method based on the target protein and chemical structure [13]. The biological activities of drugs with similar chemical structures are frequently comparable [14]. Common drug targets that produce relative therapeutic effects will also have comparable signaling cascades and, as a result, comparable side effects.

After focusing on biological and chemical properties, DSE prediction research was extended to phenotypic traits [15]. Zheng et al. [16] used therapeutic data, drug substitutes, targets, and chemical structures. Their study was based on the idea that drugs may have similar side effects due to similar therapeutic effects.

Scheiber et al. [17] predict DSEs in their study through the association of DSEs with chemical structures. Their analysis considered the PharmaPendium [18] database, which consists of 1842 drugs and 4210 side effects. To associate the DSEs with chemical features, the Laplacian-based Naïve Bayes (NB) classifier was proposed in their study. The logistic regression (LR) model was used in the study [17] to predict the side effects events after considering the Lexicom website [19] for the extraction of drug side effect association. In their research, experiments were based on 809 drugs and 852 side effects events, and as a result, an AUROC score of 87% was achieved with the LR model.

Huang et al. [20] identified drug cardiotoxicity side effects with LR and an SVM. To obtain drug targets, protein–protein interactions (PPIs), and side effects, they considered

DrugBank, human protein–protein interactions (HAPPIs), and SIDER, respectively, in their study. Fisher's exact test (FET) and the Wilcoxon rank-sum test (WRST) were used to reduce the data dimension and, in their study, the class imbalance problems were handled with the sample-balancing method. They achieved an accuracy score of 67.50% with the SVM.

In the proposed study of Jiang and Zheng [21], potential side effects were predicted with supervised ML algorithms, including an SVM, maximum entropy (ME), and NB from Twitter posts. Their experiments comprised 6829 tweets for five drugs and achieved an *F*1 score of 84.8% with the ME classifier, whereas Ginn et al. [22] studied experiments based on 10,822 tweets for 76 drugs, and the SVM outperformed the NB with an accuracy of 76.6%. Zhang et al. [23] proposed an effective model named Feature Selection-Based Multi-Label KNN (FS-MLKNN) based on drug chemical information and drug targets for the prediction of DSEs. The 2260 side effects and 1080 drugs were considered for experiments and achieved an area under the precision–recall (AUPR) score of 48.02%, 40.04% and 42.86% for three benchmark datasets ([24–26], respectively) in their proposed study.

Different drug information was integrated into the study of Zhang et al. [8] to predict DSEs with the proposed Linear Neighborhood Similarity Method. Further, the LNSM extended and proposed two approaches (Cost Minimization Integration and Similarity Interaction) and their proposed model outperformed [24–26], with an AUC score of 90.91%. The study of Jamal et al. [27] predicted neurological side effects from the combination of chemical, biological, and phenotypic information. In their research, feature extraction and synthetic minority oversampling techniques (SMOTEs) were used for dimensionality reduction and class imbalance problems, and they achieved an accuracy of 94.18%.

The study by De et al. [28] explored the importance of machine learning and deep learning to identify the relationship between side effects and chemical substructures via distinct fingerprint extraction from the compound with DL methodology. Their experiment dataset consisted of 6123 DSEs and 1420 drugs, and they were able to achieve an accuracy of 97.70% for skin striae DSEs. Wang et al. [29] proposed a DL-based model to analyze the DSEs using drug descriptors. In their study, textual information was collected from MEDLINE, the biological properties (e.g., target and enzymes) were collected from DrugBank, and 17 drug properties were collected from PubChem. Furthermore, to integrate drug properties, a multi-layer perception (MLP) model with an extension of two hidden layers was considered in their research to achieve an effective AUC of 84.40%. Lee et al. [30] proposed a DL model that can predict drug adverse events (DAEs) using preclinical data. Their study used a dataset of over 5000 compounds with preclinical data, including the chemical structure and in vitro assay results, to train and test the model. The deep learning model was based on a multi-task neural network architecture that was designed to predict both drug efficacy and AEs. The model achieved an accuracy of 0.857, a precision of 0.891, a recall of 0.702, and an *F*1 score of 0.785, demonstrating its effectiveness in predicting drug AEs.

Zhou et al. [31] proposed a boosted random forest approach to predict DSEs from chemical structures, protein targets, transporters, treatments, pathways, and enzymes. The number of side effects and drugs in their study was 4251 and 1426, respectively, and they were able to achieve an effective precision score of 78% with the proposed model compared to KNN and MLP.

Mohsen et al. [32] proposed a DL-based approach to predict adverse drug reactions (ADRs) using two large-scale databases: Open TG-GATEs and FAERS. They developed a deep neural network that takes the input of drug molecular descriptors and gene expression data and outputs the probability of adverse reactions (ADRs) for each drug. The model was trained on a large dataset of drug–gene–ADR associations and evaluated on an independent test set. Their results showed that the proposed approach achieved high performance in predicting ADRs, outperforming several baseline methods. The authors also conducted extensive experiments to investigate the importance of different input features and showed that the gene expression data significantly improved the performance of the model.

Jiang et al. [33] proposed a method based on the assumption that drugs with similar structures are likely to have similar side effects. They constructed a network representation of drug structures and side effects and use a path-based algorithm to identify significant associations between drugs and side effects.

Liang et al. [34] proposed an approach that utilizes two sources of data: drug–drug similarities and drug–side effect similarities. The drug–drug similarities were calculated based on the similarities of their chemical structures, while the drug–side effect similarities were calculated based on their occurrence in clinical trials. They applied a transductive matrix co-completion algorithm to the constructed matrix to predict the missing entries, i.e., potential side effects. The algorithm leverages the similarities between drugs and side effects to predict the missing entries in the matrix.

Computer vision is a field of study in computer science and artificial intelligence that focuses on enabling machines to interpret and understand visual information from the world around them. It involves developing algorithms and techniques that can analyze and interpret images and videos in a way that is similar to how humans perceive the world. Computer vision has many practical applications, such as object recognition, facial recognition, motion detection, and image segmentation. It is used in a variety of fields, including healthcare, the automotive sector, entertainment, security, and robotics.

The development of deep learning techniques, such as convolutional neural networks, has greatly advanced the field of computer vision in recent years. These techniques enable computers to recognize and classify images with high accuracy, and they have led to significant advances in areas such as medical imaging, where computer vision is used to diagnose diseases from medical images. Owing to recent progress in computational methodologies, our research has successfully pinpointed various side effects associated with drugs using 2D chemical structure images. To the best of our knowledge, this study stands as a pioneering effort in this domain. The major contributions of this study are listed below:

- We compiled a comprehensive dataset by gathering information from reputable sources such as SIDER and PubChem;
- We converted our problem from a multi-label to a multi-class format, employing effective techniques for enhanced clarity and precision;
- We opted for the reliable diffusion method and traditional augmentation techniques to produce synthetic data;
- Our study introduces a novel model designed for predicting multiple drug side effects (DSEs) by leveraging 2D chemical structures of drugs;
- In an effort to streamline the training process, we incorporated a transfer learning approach, thereby minimizing the required training time;
- Our proposed model streamlines the intricate transformation process, in contrast to the NLP domain's approach, which involves converting smiles into fingerprints and extracting features.

The remainder of this paper is divided as follows: Section 2 outlines the methodology of the proposed model, while the experimental results and comparison are presented in Section 3, and Section 4 provides the conclusion of the proposed study.

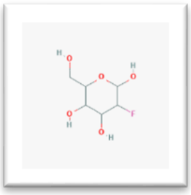
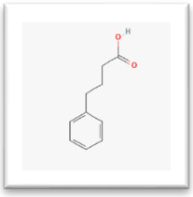
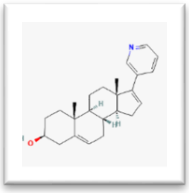
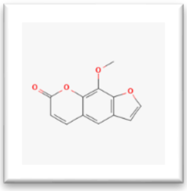
2. Proposed Methodology

This section outlines and presents the methodologies utilized and proposed to accurately predict multiple side effects of drugs using 2D drug chemical structures. These methods are carefully designed to enhance the precision and effectiveness of side-effect detection.

2.1. Dataset

We have retrieved the information of drugs and their associated side effects from the publicly available SIDER database (Version 4.1) [35], maintained by the European Molecular Biology Laboratory (EMBL). SIDER provides details about marketed medicines and their corresponding side effects. Our dataset includes information on 1430 drugs (excluding any data related to the drug named 'x', resulting in a consideration of 1429 drugs). Additionally, this database is linked with PubChem [36]. In this study, we considered fever and vomiting as multiple side effects of drugs. Furthermore, we obtained COVID-19 drug information, i.e., drug names from DrugBank [37], specifically those indicating fever and vomiting as side effects. This step was crucial to assess the generalization and robustness of our proposed model. In the initial phase, we collected and categorized the data in a multi-label binary form, assigning labels of 0 and 1. Here, 0 signifies the absence of the specific side effect, while 1 indicates the presence of the particular side effect; see Table 1.

Table 1. Prepared dataset labeling as multi-label side effects ([35,36]).

Drug Name	Drug Structure	Fever	Vomiting
18F-FDG		0	0
4-PBA		0	1
abiraterone		1	0
8-MOP		1	0

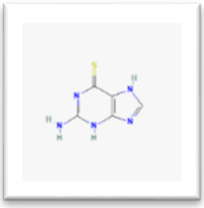
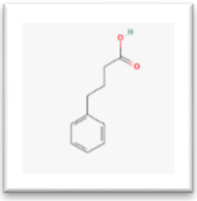
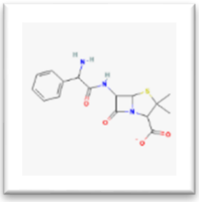
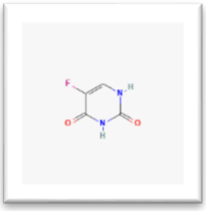
2.2. *Transformation of Multi-Label Problem to Multi-Class*

Our research dives into the tricky challenges of dealing with multi-label classification. This is a bit more complicated than sorting things into just two or a few categories. Some of the big issues we tackle include labels depending on each other, having lots of possible labels, some labels showing up much more than others, and the need for careful ways to figure out how well our models are performing. To make things more manageable, we took a smart approach, transforming our multi-label problem into something simpler called multi-class classification. We used an equation (Equation (1)) that looks at whether fever and vomit are present or not, similar to checking boxes with "yes" or "no" to make things easier. Our goal is to make classifying things less complicated, reduce how much computing power we need, and make our models easier to understand. We know there are still challenges in designing the computer programs for this, so we picked our methods carefully to make sure we dealt well with the unique problems that come with multi-label classification.

$$T(f,v) = f \times (2^1) + v \times (2^0) \quad : f \in \{0,1\} \;\&\; v \in \{0,1\} \tag{1}$$

Table 2 presents the classes along with their associated labels, derived from the transformation process.

Table 2. Problem transformation to multi-class from multi-label.

Drug	Fever	Vomit	Class
	0	0	0
	0	1	1
	1	0	2
	1	1	3

In the initial phase of data processing, we identified a subset of images that solely featured the names of elements without relevant chemical structure information. Consequently, we deemed these images unsuitable for analysis and opted to exclude them from the dataset. Refer to Figure 1 for dataset counts post initial preprocessing.

2.3. Data Augmentation with Diffusion and Conventional Augmentation Techniques

In this study, we harnessed the potent capabilities of the Pix2Pix model, an advanced framework introduced by Brooks et al. [38]. Designed to emulate stable diffusion techniques, this model serves as a powerful tool to artificially expand our dataset, mitigating the common challenge of data scarcity in image-related tasks. By generating synthetic images, Pix2Pix significantly broadens the scope and diversity of our dataset, bolstering the comprehensiveness of our research samples. Operating as a conditional Generative Adversarial Network (GAN) [39], Pix2Pix learns to produce images based on specified conditions or input data. This conditioning mechanism ensures that the generated images align precisely with the attributes outlined in the stable diffusion technique, promising to enhance the robustness and generalizability of our research outcomes. The augmentation process involves resizing the original Pix2Pix-based generated images of 1024 × 1024 pixels, followed by saving the generated images in 300 × 300 pixels. For transparency, Table 3 presents the empirically determined hyperparameters crucial for the effectiveness of the Pix2Pix model in our artificial image creation process.

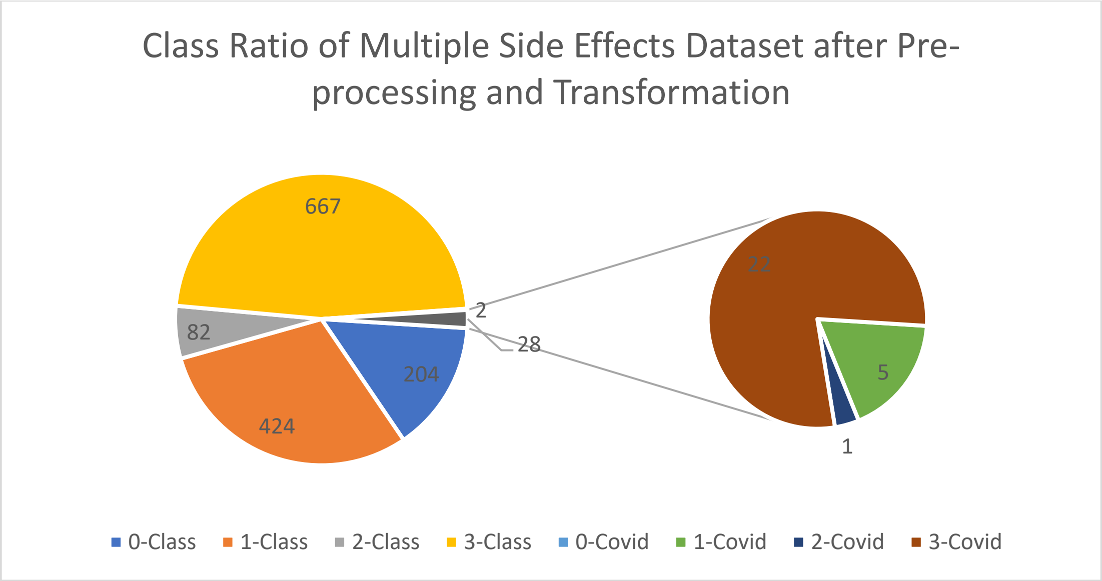


Figure 1. DSE's Samples Counts after Preprocessing and Transformation.

Table 3. Hyperparameter configurations for Pix2Pix model.

Hyperparameter	Value
Image Guidance Scale	2.0
Number of Inference Steps	20
Mode	RGB
Input Image Size	300 × 300
Generated Image Size	1024 × 1024
Save Size	300 × 300

Furthermore, to enhance the model's ability to generalize effectively, we integrated traditional augmentation techniques on the images produced by the Pix2Pix model. For these conventional augmentation processes, we carefully selected and applied specific parameters, as outlined in Table 4.

Table 4. Hyperparameters of conventional augmentation method.

Hyperparameter	Value
Rotation Range	10–30
Shear Range	0.1–0.2
Zoom Range	0.1–0.2
Brightness Range	0.8–1.2
Horizontal Flip	True
Fill Mode	Nearest

As far as we are aware, this marks the first implementation of augmentation methods to artificially broaden the dataset through two distinct techniques, as depicted in Figure 2. After this augmentation process, we obtained a balanced dataset of 1000 samples of each class excluding the COVID-19 test samples.

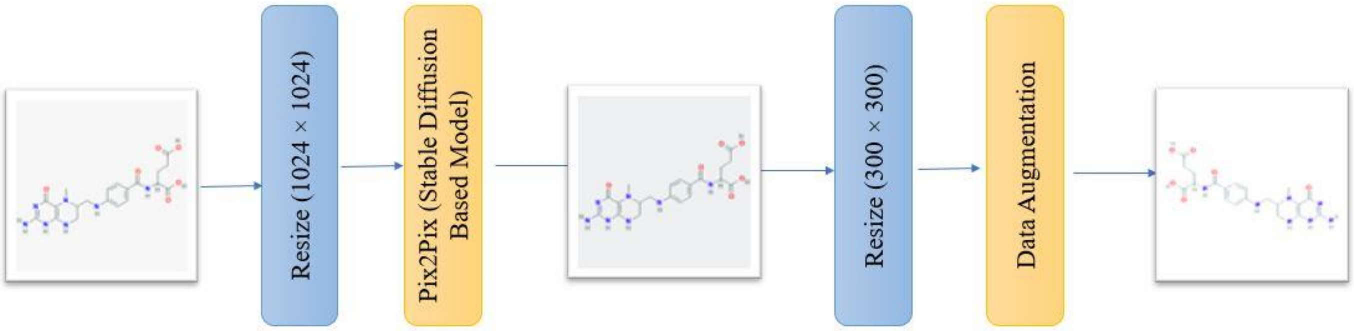


Figure 2. Data augmentation with the combination of Pix2Pix and conventional augmentation techniques.

2.4. Proposed Model Based on Transfer Learning and Fine-Tuning

The proposed model, CHEM2SIDE, is based on the ResNet152V2 architecture [40] that is the most recent progression within the ResNet series, encompassing a grand total of 152 layers that incorporate improved skip connections and batch normalization. These enhancements contribute significantly to the facilitation of gradient flow and the preservation of key features. The process of fine-tuning this model is a meticulous endeavor. It involves freezing all layers of the ResNet152V2 architecture to harness the knowledge contained in the pre-existing ImageNet weights. Specifically, the uppermost layers are deliberately excluded from this approach, leading to the creation of a tailored architecture. This personalized structure begins with the introduction of a flattening layer, followed by the deliberate addition of six subsequent layers. These six layers include a dense layer with 512 units, dropout with a rate of 0.2, a 256-unit dense layer, another dropout layer with a rate of 0.2, a 128-unit dense layer, and ultimately, a fully connected layer comprising four units. These four units align seamlessly with the four distinct classes relevant to our classification task, as depicted in Figure 3 of the proposed framework architecture.

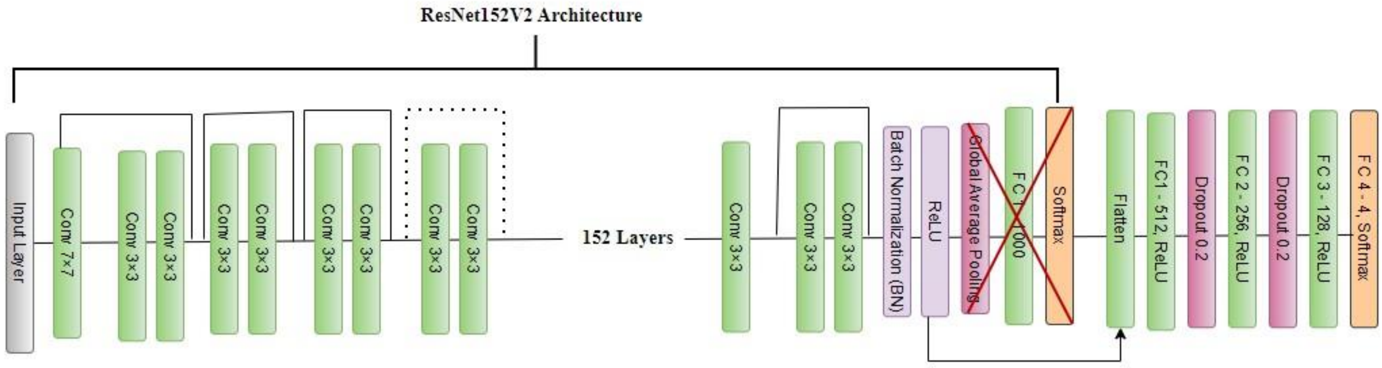


Figure 3. Proposed architecture (CHEM2SIDE) framework.

Activation functions play a crucial role in determining the performance of a model. ReLU is used in hidden layers to introduce non-linearity, and SoftMax is used in the output layer for class probability generation in multi-class prediction. The model is optimized for multi-class prediction using categorical cross-entropy as the loss function and the Adam optimizer (lr = 0.001). Strategies like learning rate reduction and early stopping are used to ensure robustness and prevent overfitting; see Table 5 for the hyperparameters and their optimized values.

Table 5. Hyperparameters and values for proposed model training.

Parameters	Values
Batch Size	32
Epochs	100
Learning Rate	0.0001
Reduced Learning Rate	Yes
Patience for Reduced Learning Rate	3
Early-Stopping Patience	5
Stratified K-Fold K Value	3
Optimizer	Adam
FC-Layer Activation Function	ReLU
FC-Layer Neurons	512, 256, 128
Output-Layer Neurons	4
Output Activation Function	SoftMax
Dropout between FC Layers	0.2, 0.2
Compile Loss	Categorical Cross-Entropy

3. Experiment Results and Discussion

In this section, we present a comprehensive discussion of the evaluation measures, experimental details, and the results obtained through the proposed methodology. We delve into the assessment criteria used to gauge the performance of our approach, provide insights into the experimental setup and configurations, and present the outcomes achieved during our evaluation process.

3.1. Evaluation Metrics

In the realm of machine learning and deep learning, evaluation metrics play a vital role in gauging model performance. These measures are fundamental to statistical research and are essential in assessing the effectiveness of our proposed model. In this study, we emphasized the following key assessment measures to evaluate the efficacy of our approach.

- **Accuracy:** Accuracy is a metric that assesses the overall correctness of a model's predictions. It calculates the proportion of correctly classified samples out of the total samples. While accuracy is a crucial evaluation measure, it may not be sufficient in certain scenarios, such as imbalanced datasets or cases where different types of errors have varying consequences. In such situations, additional evaluation metrics may be necessary to provide a more comprehensive understanding of the model's performance and capabilities. In Equations (2)–(4), *TP*, *TN*, *FP*, and *FN* represent true positive, true negative, false positive, and false negative, respectively.

$$Accuracy = {}^{(TP + TN)}\Big/_{(TP + FP + TN + FN)} \tag{2}$$

- **Precision:** Precision is a metric that evaluates a model's capability to correctly identify positive samples among the predicted positive samples. It calculates the proportion of true-positive predictions to the total number of positive predictions (which includes both true positives and false positives). Precision provides valuable insights into how accurately the model detects and classifies positive instances, making it an essential measure in many classification tasks.

$$P = {}^{TP}\Big/_{(TP + FP)} \tag{3}$$

- **Recall:** Recall, also known as sensitivity or the true-positive rate, measures the model's capacity to correctly identify positive samples among all actual positive samples. It calculates the ratio of true positives to the sum of true positives and false negatives. Recall reflects the model's ability to be comprehensive in capturing positive instances, making it a critical evaluation metric in classification tasks.

$$R = {}^{TP}\Big/_{(TP + FN)} \tag{4}$$

- **Score:** The $F1$ score is computed as the harmonic mean of precision and recall, providing a single statistic that balances the two metrics. This makes it particularly useful when dealing with imbalanced class distributions or scenarios where equal emphasis is placed on both types of errors. The $F1$ score ranges from 0 to 1, with 1 representing the best possible performance of the model. By incorporating both precision and recall, the $F1$ score offers a comprehensive evaluation of the model's overall effectiveness in classification tasks.

$$F1 = {}^{(2 \times P \times R)}\Big/_{(P + R)} \tag{5}$$

3.2. Stratified K-Fold (Train, Validation, Test)

In the context of this particular study, we turned from the conventional method of Stratified K-Fold by introducing a more granular division in the validation set. Specifically, we subdivided the validation set further, allocating 70% of its samples for traditional validation purposes and keeping the remaining 30% for testing. This departure from the standard practice is motivated by the need for a more effective evaluation, allowing us to assess model performance on a separate subset within the validation data; see Algorithm 1.

Algorithm 1: Stratified K-Fold Approach for the Proposed Study

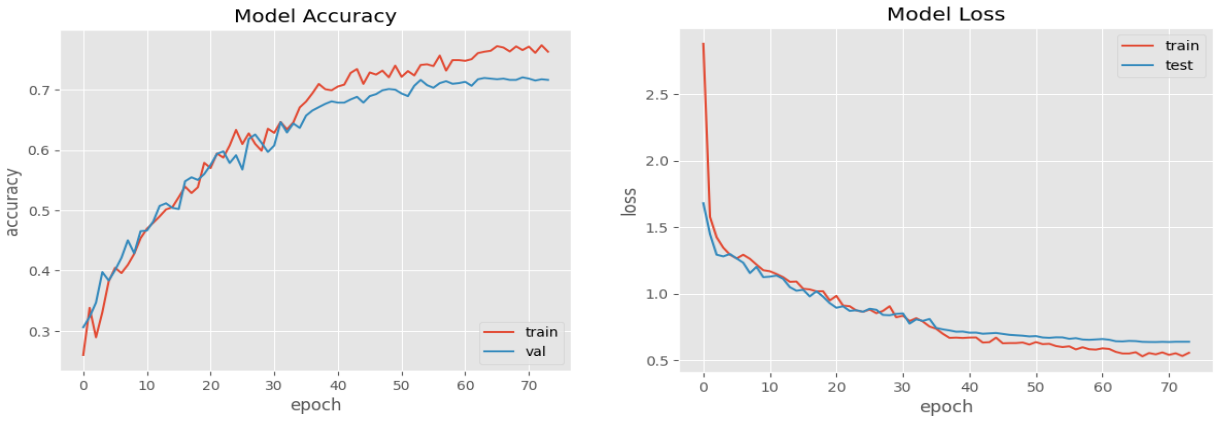
```
Input: Drug 2D Chemical Structures Dataset
Step 1: Split the dataset into 3 folds.
Repeat
For fold i = 1 to 3 do
Step 2: Select fold i as the validation and the remaining folds as the training set.
Step 3: Divide the validation set with ratio 70:30, 30% used for testing.
Step 4: Fit the model on training set.
Step 5: Evaluate for validation set during training.
Step 6: At the end of the fold i, evaluate the model for test set.
Step 7: Store the evaluation scores in list S.
End-for
Step 8: Find the average performance with S.
Output: Average Performance of the Model
```

The foundation behind this approach is to enhance the robustness of the model evaluation process. By having a dedicated testing subset within the validation set, we aim to obtain a more comprehensive understanding of how the model generalizes to unseen data. For model training, we utilized the following software: Jupyter Notebook (6.4.12) and Python (3.11.5) with updated TensorFlow and Keras. As for hardware, we employed an RTX-3080 GPU with 20 GB of memory.

3.3. Results of Proposed CHEM2SIDE Model and Discussion

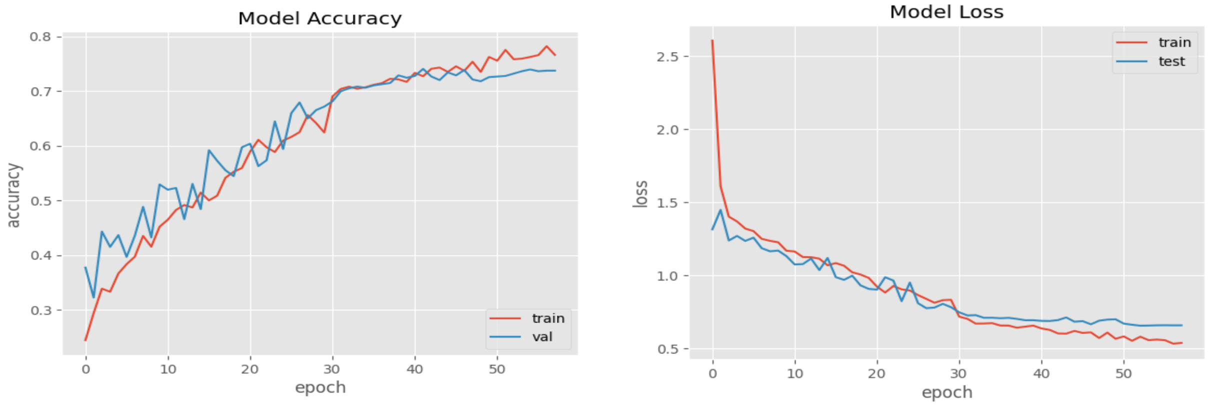
For assessing the model's performance, we employed the proposed CHEM2SIDE based on a fine-tuned ResNet152V2 architecture for 300 × 300-sized 2D structure images. We selected the full size of the images so as to not remove any necessary information. Employing a Stratified K-Fold cross-validation technique with a fold value of three, we ensured rigorous validation. This approach minimizes bias by segmenting the dataset into homogeneous subsets for both training and evaluation.

The proposed model underwent training for a total of 100 epochs. To mitigate overfitting, we implemented early stopping with a patience level of five, coupled with the reduced learning rate technique featuring a patience level of three. The outcomes presented in Figures 4–6 unmistakably indicate that the model training concluded prior to reaching the 100-epoch mark. This conclusion arose as the early-stopping criteria were met consistently across all folds—namely, Fold 1, Fold 2, and Fold 3. This strategic approach ensures that the model's training halts when optimal performance is achieved to prevent overfitting and enhance generalization across the dataset.



(**a**) (**b**)

Figure 4. Model accuracies and losses graph for Fold 1.



(**a**) (**b**)

Figure 5. Model accuracies and losses graph for Fold 2.

In the proposed study, we achieved average maximum training accuracy, maximum average training loss, average minimum training accuracy, and average minimum training loss scores of 0.7832, 2.7820, 0.2466, and 0.5284, respectively. Whereas, the average maximum validation accuracy, maximum validation training loss, average mini-mum validation accuracy, and average minimum validation loss scores achieved were 0.7377, 1.5051, 0.3139, and 0.6391, respectively, with the proposed model.

Furthermore, we designated 30% of the data for each fold as the test set. Specifically, three distinct test sets correspond to Fold 1, Fold 2, and Fold 3; the classification report of 30% test data of each fold can be seen in Table 6. Approximately in each fold, the training set consists of 2666 samples, validated with 933–934 and 400–401 test samples. The total test samples for each fold are mentioned in the support column of Table 6.

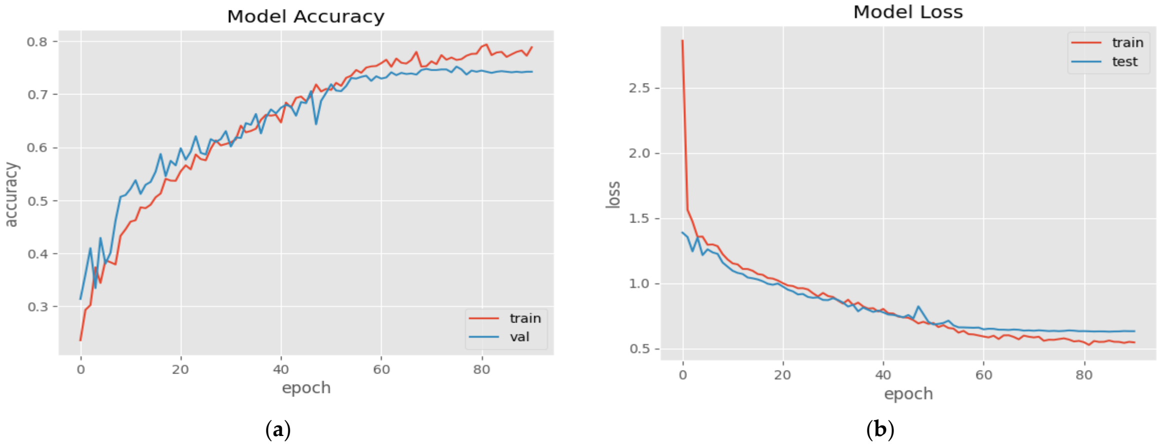


Figure 6. Model accuracies and losses graph for Fold 3.

Table 6. Test accuracies of each fold's test data of Stratified K-Fold.

Test Set Fold	Per-Class Samples	Accuracy	Weighted Precision	Weighted Recall	Weighted *F*1	Support
Fold 1	Class 0: 97, Class 1: 90, Class 2: 106, Class 3: 108	0.72	0.73	0.72	0.72	401
Fold 2	Class 0: 95, Class 1: 90, Class 2: 106, Class 3: 109	0.73	0.73	0.73	0.73	400
Fold 3	Class 0: 95, Class 1: 91, Class 2: 106, Class 3: 108,	0.74	0.74	0.74	0.74	400
Average		0.73	0.73	0.73	0.73	400

In scenarios involving imbalanced classes or substantial discrepancies in the misclassification between different classes, the utilization of a confusion matrix becomes pivotal for evaluating the effectiveness of a classification model. The confusion matrix of each fold's test data can be seen in Figures 7–9.

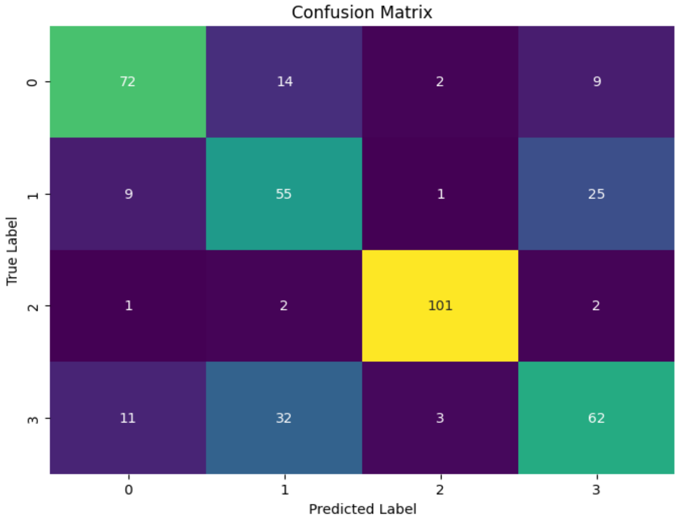


Figure 7. Confusion matrix for Fold 1 test data.

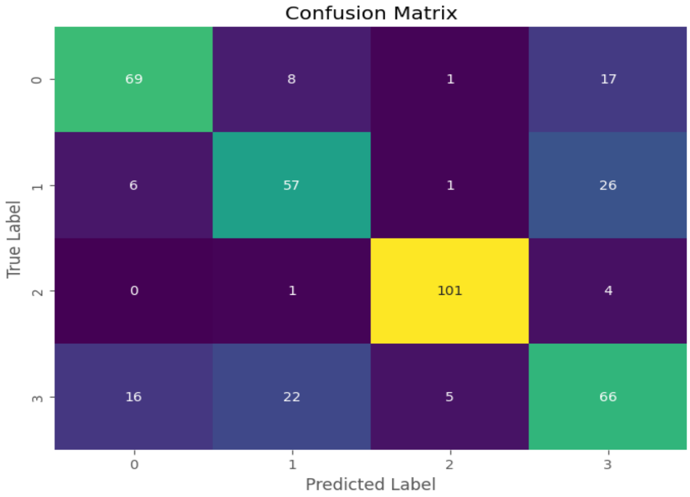


Figure 8. Confusion matrix for Fold 2 test data.

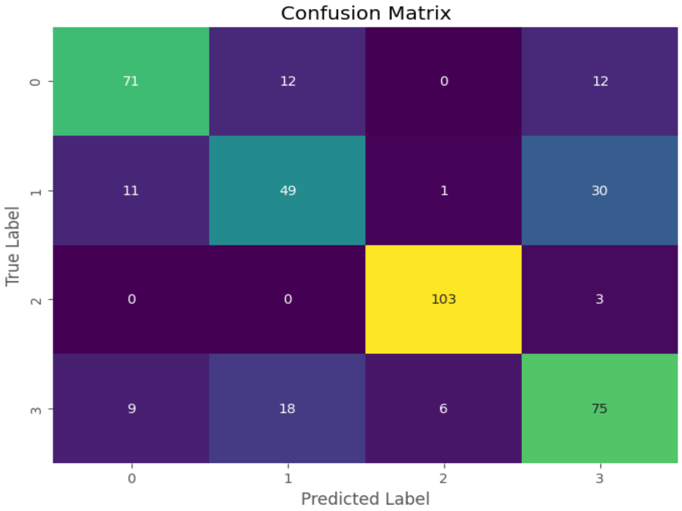


Figure 9. Confusion matrix for Fold 3 test data.

3.4. Comparison with the Literature Contributions

In our proposed study, there is a notable absence of prior studies focusing on the prediction of multiple drug side effects solely relying on 2D chemical structure images. While certain research has explored predictions based on canonical SMILES, primarily within the domain of natural language processing (NLP), our study stands out as pioneering. To the best of our knowledge, our research marks the first instance of employing a deep learning approach for this purpose. This is why a direct comparison may not be feasible due to the pioneering nature of our study. However, we conducted a comparative analysis by benchmarking our proposed model against established state-of-the-art models, i.e., MobileNetV2 and KNN, to assess its performance and efficacy in comparison to well-established frameworks; see Table 7.

Table 7. Proposed CHEM2SIDE model comparison with state-of-the-art models.

Model	**Average Scores (Stratified K-Fold = 3, Validation → 70% Validation and 30% Testing)**					
	Train Accuracy	**Validation Accuracy**	**Test Accuracy**	**Weighted Precision**	**Weighted Recall**	**Weighted *F*1**
Proposed CHEM2SIDE	0.78	0.73	0.73	0.73	0.73	0.73
MobileNetV2	0.68	0.66	0.66	0.67	0.66	0.65

Table 7. *Cont.*

Model	Average Scores (Stratified K-Fold = 3, Validation → 70% Validation and 30% Testing)					
	Train Accuracy	Validation Accuracy	Test Accuracy	Weighted Precision	Weighted Recall	Weighted *F*1
VGG16	0.51	0.51	0.51	0.51	0.51	0.50
DenseNet121	0.62	0.62	0.61	0.62	0.61	0.61
KNN	0.38	0.33	0.32	0.32	0.33	0.25

3.5. Robustness of Proposed CHEM2SIDE

COVID-19, or coronavirus disease 2019, is a global viral respiratory illness caused by the SARS-CoV-2 virus. Drug discovery has been vital in combating this pandemic. Medications like remdesivir, dexamethasone, and monoclonal antibodies have been repurposed or developed to treat COVID-19. These drugs target different aspects of the virus or the body's response. However, they can have side effects, such as fever with ritonavir and increased blood sugar, weight gain, and mood changes with dexamethasone. Careful consideration of benefits versus side effects is crucial, and ongoing research seeks new treatments and strategies to minimize the impact on public health. Our study tested a model's performance using COVID-19 drug 2D chemical structures, extracted from DrugBank [37], to evaluate their potential side effect of fever; see Table 8.

Table 8. COVID-19 Drugs.

COVID-19 Drugs		
bromhexine	ivermectin	budesonide
chloroquine	losartan	celecoxib
colchicine	montelukast	chlorpromazine
dipyridamole	nitazoxanide	darunavir
methylprednisolone	quetiapine	dexamethasone
rivaroxaban	ribavirin	famotidine
tranexamic acid	ritonavir	fondaparinux
argatroban	ruxolitinib	heparin
azithromycin	simvastatin	hydroxychloroquine
bicalutamide	sofosbuvir	ibuprofen

Table 9 reveals inconsistencies in the number of samples for the 30 COVID-19 drugs across different classes; insufficient samples are particularly evident for classes 0 and 2.

Table 9. COVID-19 drugs' class-wise ratio.

Class Label	Class Samples
0	2
1	5
2	1
3	22

We achieved an accuracy of 0.57, a precision of 0.61, a recall of 0.57, and an *F*1 score of 0.58 with these insufficient data of COVID-19 with the proposed model. We addressed the issue of insufficient samples by utilizing a diffusion-based model, Pix2Pix, in conjunction with conventional augmentation techniques discussed in Section 3.3; see Table 10.

Table 10. COVID-19 drugs' class-wise ratio after augmentation process.

Class Label	Class Samples
0	22
1	22
2	22
3	22

We have evaluated our proposed model with the augmented COVID-19 drugs and achieved an accuracy of 0.72, a precision of 0.79, a recall of 0.72, and an *F*1 score of 0.73 with these sufficient data of COVID-19 drugs. The class-wise scores can be seen in Table 11 as a classification report. For a better visualization, the confusion matrix can be seen in Figure 10. From the confusion matrix, it is observable that our proposed model is generalized as effectively predicting all the classes rather than exhibiting bias.

Table 11. Classification report of COVID-19 drugs (original + augmented).

	Precision	Recall	*F*1	Support
0	1.00	0.68	0.81	22
1	0.55	0.82	0.65	22
2	1.00	0.68	0.81	22
3	0.60	0.68	0.64	22
Accuracy			0.72	88
Macro Average	0.79	0.72	0.73	88
Weighted Average	0.79	0.72	0.73	88

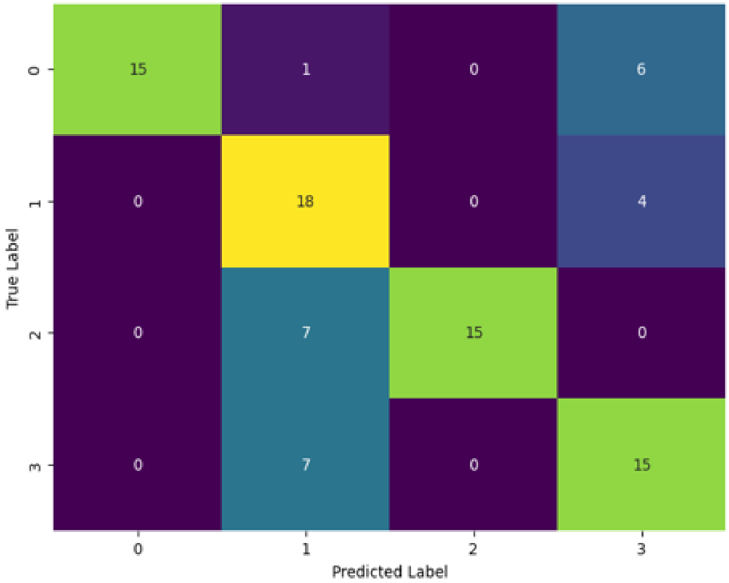


Figure 10. Confusion matrix for COVID-19 drugs (Original + Augmented Samples).

This study employs a critical assessment of the proposed framework that includes the receiver operating characteristics curve (ROC) and a confusion matrix. The ROC curve serves as a valuable visual representation of the balance between true-positive and false-positive rates. Subsequently, the model was assessed by testing it with samples of COVID-19 drugs that were not part of the initial training dataset; see Figure 11.

To the best of our knowledge, this study represents the first of its kind in predicting multiple drug side effects using 2D chemical structure images. Prior to this, side effects were predicted using 1D chemical structures, which are based on the NLP domain. Consequently, a direct comparison of our proposed model with existing studies is not feasible. However, we have also conducted comparisons of our models for COVID-19 drugs to demonstrate the robustness of our proposed model; see Table 12.

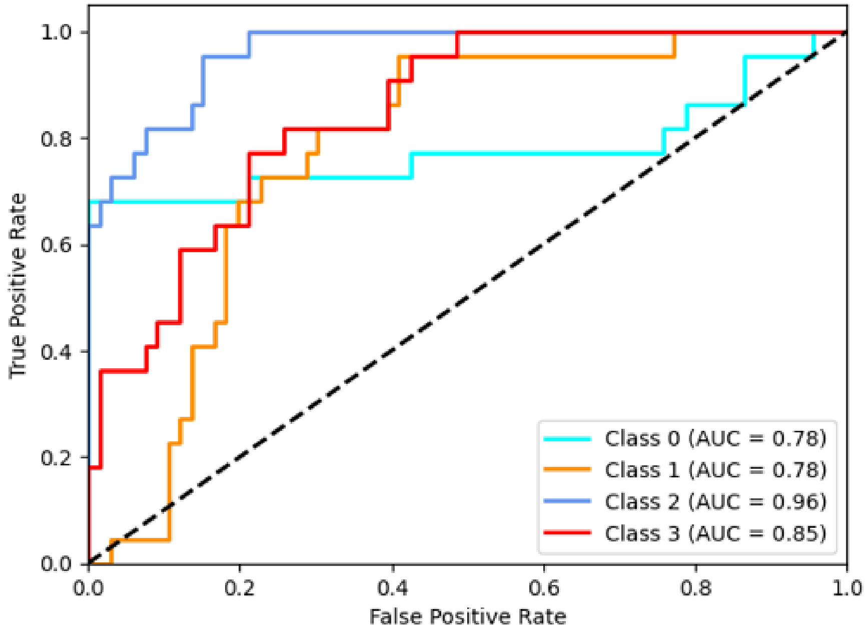


Figure 11. ROC curve for COVID-19 drugs' side-effect prediction.

Table 12. Comparison of proposed model with state-of-the-art models for COVID-19 drugs (original + augmented).

	COVID-19 Accuracy	COVID-19 Weighted Precision	COVID-19 Weighted Recall	COVID-19 Weighted *F*1
Proposed CHEM2SIDE	0.72	0.79	0.72	0.73
MobileNetV2	0.45	0.47	0.45	0.39
VGG16	0.43	0.51	0.43	0.33
DenseNet121	0.34	0.16	0.34	0.22
KNN	0.35	0.17	0.35	0.23

4. Conclusions

To overcome the hefty process of 1D chemical structure transformation, we have introduced a model to predict drug side effects directly from 2D chemical structure images due to the advancement in the image-processing models, i.e., transfer learning. The proposed CHEM2SIDE model is based on the ResNet152V2 architecture and transfer learning approach. To create our dataset, we utilized ground-truth labels from SIDER and 2D structures from PubChem. Although the 2D chemical images are sparse, the ensemble augmentation and transfer learning approach were considered in this study to mitigate this problem. Our model exhibited promising results, with an average training accuracy of 0.78, alongside commendable average validation and test accuracy, precision, and recall, all consistently at 0.73 for multiple drug side-effect prediction, i.e., fever and vomit.

Moreover, when assessing our model's performance in predicting side effects for COVID-19 drugs, we achieved an accuracy of 0.72, a precision of 0.79, a recall of 0.72, and an *F*1 score of 0.73. In comparative evaluations against established transfer learning and ML models such as VGG16, MobileNetV2, DenseNet121, and KNN, our CHEM2SIDE model outperformed them all, underlining a substantial advancement in the field of drug side-effect prediction. This research not only streamlines the prediction process but also underscores the potential for a transformative impact in drug development and safety. Future work may involve expanding the dataset for the drug side-effect prediction model to enhance its real-world effectiveness. Consequently, exploring multi-modal data integration for improved accuracy, along with collaboration with industry and regulatory agencies, is undertaken for practical adoption.

Author Contributions: Conceptualization, M.A.A.; methodology, M.A.A.; validation, M.A.A., S.M., S.A. and M.T.; supervision, M.I.; investigation, M.A.A.; data curation, M.A.A.; writing—original draft preparation, M.A.A.; writing—review and editing, M.A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The prepared dataset and codes will be provided on request.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

TP	True Positive
FP	False Positive
TN	True Negative
TP	True Positive
FN	False Negative
LR	Learning Rate
GAN	Generative Adversarial Networks
ML	Machine Learning
DL	Deep Learning
CNN	Convolutional Neural Network
AI	Artificial Intelligence
1D	One-Dimensional
2D	Two-Dimensional
ROC	Receiver Operating Characteristics
FDA	Food and Drug Administration
DSEs	Drug Side Effects
KNN	K-Nearest Neighbors

References

1. Khalil, H.; Huang, C. Adverse drug reactions in primary care: A scoping review. *BMC Health Serv. Res.* **2020**, *20*, 5. [CrossRef] [PubMed]
2. Billingsley, M.L. Druggable Targets and Targeted Drugs: Enhancing the Development of New Therapeutics. *Pharmacology* **2008**, *82*, 239–244. [CrossRef] [PubMed]
3. Giacomini, K.M.; Krauss, R.M.; Roden, D.M.; Eichelbaum, M.; Hayden, M.R.; Nakamura, Y. When good drugs go bad. *Nature* **2007**, *446*, 975–977. [CrossRef] [PubMed]
4. Drugs | FDA. Available online: https://www.fda.gov/drugs (accessed on 10 April 2023).
5. Yao, B.; Zhu, L.; Jiang, Q.; Xia, H.A. Safety Monitoring in Clinical Trials. *Pharmaceutics* **2013**, *5*, 94–106. [CrossRef] [PubMed]
6. Ho, T.-B.; Le, L.; Thai, D.T.; Taewijit, S. Data-driven Approach to Detect and Predict Adverse Drug Reactions. *Curr. Pharm. Des.* **2016**, *22*, 3498–3526. [CrossRef] [PubMed]
7. Boland, M.R.; Jacunski, A.; Lorberbaum, T.; Romano, J.D.; Moskovitch, R.; Tatonetti, N.P. Systems biology approaches for identifying adverse drug reactions and elucidating their underlying biological mechanisms. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **2015**, *8*, 104–122. [CrossRef] [PubMed]
8. Zhang, W.; Chen, Y.; Tu, S.; Liu, F.; Qu, Q. Drug side effect prediction through linear neighborhoods and multiple data source integration. In Proceedings of the 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Shenzhen, China, 15–18 December 2016; pp. 427–434.
9. Shaked, I.; Oberhardt, M.A.; Atias, N.; Sharan, R.; Ruppin, E. Metabolic Network Prediction of Drug Side Effects. *Cell Syst.* **2016**, *2*, 209–213. [CrossRef] [PubMed]
10. Cakir, A.; Tuncer, M.; Taymaz-Nikerel, H.; Ulucan, O. Side effect prediction based on drug-induced gene expression profiles and random forest with iterative feature selection. *Pharmacogenomics J.* **2021**, *21*, 673–681. [CrossRef]
11. Uner, O.C.; Cinbis, R.G.; Tastan, O.; Cicek, A.E. DeepSide: A Deep Learning Framework for Drug Side Effect Prediction. *bioRxiv* **2019**, 843029. [CrossRef]

12. Pauwels, E.; Stoven, V.; Yamanishi, Y. Predicting drug side-effect profiles: A chemical fragment-based approach. *BMC Bioinform.* **2011**, *12*, 169. [CrossRef]
13. Yamanishi, Y.; Pauwels, E.; Kotera, M. Drug side-effect prediction based on the integration of chemical and biological spaces. *J. Chem. Inf. Model.* **2012**, *52*, 3284–3292. [CrossRef] [PubMed]
14. Martin, Y.C.; Kofron, J.L.; Traphagen, L.M. Do Structurally Similar Molecules Have Similar Biological Activity? *J. Med. Chem.* **2002**, *45*, 4350–4358. [CrossRef] [PubMed]
15. Duffy, Á.; Verbanck, M.; Dobbyn, A.; Won, H.H.; Rein, J.L.; Forrest, I.S.; Nadkarni, G.; Rocheleau, G.; Do, R. Tissue-specific genetic features inform prediction of drug side effects in clinical trials. *Sci. Adv.* **2020**, *6*, 6242. [CrossRef] [PubMed]
16. Zheng, Y.; Peng, H.; Ghosh, S.; Lan, C.; Li, J. Inverse similarity and reliable negative samples for drug side-effect prediction. *BMC Bioinform.* **2019**, *19*, 91–104. [CrossRef] [PubMed]
17. Cami, A.; Arnold, A.; Manzi, S.; Reis, B. Predicting Adverse Drug Events Using Pharmacological Network Models. *Sci. Transl. Med.* **2011**, *3*, 114ra127. [CrossRef] [PubMed]
18. Rees, K.E.; Chyou, T.-Y.; Nishtala, P.S. A Disproportionality Analysis of the Adverse Drug Events Associated with Lurasidone in Paediatric Patients Using the US FDA Adverse Event Reporting System (FAERS). Available online: https://link.springer.com/article/10.1007/s40264-020-00928-1 (accessed on 13 April 2023).
19. Drug Decision Support from Wolters Kluwer | Wolters Kluwer. Available online: https://www.wolterskluwer.com/en/know/drug-decision-support-solutions (accessed on 13 April 2023).
20. Huang, L.-C.; Wu, X.; Chen, J.Y. Predicting adverse side effects of drugs. *BMC Genom.* **2011**, *12*, S11. [CrossRef] [PubMed]
21. Jiang, K.; Zheng, Y. Mining Twitter data for potential drug effects. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: New York, NY, USA, 2013; Volume 8346, pp. 434–443. [CrossRef]
22. Ginn, R.; Pimpalkhute, P.; Nikfarjam, A.; Patki, A.; O'Connor, K.; Sarker, A.; Smith, K.; Gonzalez, G. Mining Twitter for Adverse Drug Reaction Mentions: A corpus and Classification Benchmark. Available online: https://www.researchgate.net/profile/Abeed-Sarker/publication/280301158_Mining_Twitter_for_adverse_drug_reaction_mentions_a_corpus_and_classification_benchmark/links/56d205b608ae85c8234ae39d/Mining-Twitter-for-adverse-drug-reaction-mentions-a-corpus-and-classification-benchmark.pdf (accessed on 13 April 2023).
23. Zhang, W.; Liu, F.; Luo, L.; Zhang, J. Predicting drug side effects by multi-label learning and ensemble learning. *BMC Bioinform.* **2015**, *16*, 1–11. [CrossRef]
24. Elaziz, M.A.; Yousri, D. Automatic selection of heavy-tailed distributions-based synergy Henry gas solubility and Harris hawk optimizer for feature selection: Case study drug design and discovery. *Artif. Intell. Rev.* **2021**, *54*, 4685–4730. [CrossRef]
25. Liu, M.; Wu, Y.; Chen, Y.; Sun, J.; Zhao, Z.; Chen, X.W.; Matheny, M.E.; Xu, H. Large-Scale Prediction of Adverse Drug Reactions using Chemical, Biological, and Phenotypic Properties of Drugs. Available online: https://academic.oup.com/jamia/article-abstract/19/e1/e28/2909247 (accessed on 11 April 2023).
26. Mizutani, S.; Pauwels, E.; Stoven, V.; Goto, S.; Yamanishi, Y. Relating Drug–Protein Interaction Network with Drug Side Effects. Available online: https://academic.oup.com/bioinformatics/rticle-abstract/28/18/i522/246017 (accessed on 14 April 2023).
27. Jamal, S.; Goyal, S.; Shanker, A.; Grover, A. Predicting neurological Adverse Drug Reactions based on biological, chemical and phenotypic properties of drugs using machine learning models. *Sci. Rep.* **2017**, *7*, 872. [CrossRef]
28. Dey, S.; Luo, H.; Fokoue, A.; Hu, J.; Zhang, P. Predicting adverse drug reactions through interpretable deep learning framework. *BMC Bioinform.* **2018**, *19*, 476. [CrossRef]
29. Wang, C.-S.; Lin, P.-J.; Cheng, C.-L.; Tai, S.-H.; Yang, Y.-H.K.; Chiang, J.-H. Detecting Potential Adverse Drug Reactions Using a Deep Neural Network Model. *J. Med. Internet Res.* **2019**, *21*, e11016. Available online: https://www.jmir.org/2019/2/e11016/ (accessed on 14 April 2023). [CrossRef] [PubMed]
30. Lee, C.Y.; Chen, Y.-P.P. Prediction of drug adverse events using deep learning in pharmaceutical discovery. *Brief. Bioinform.* **2020**, *22*, 1884–1901. [CrossRef] [PubMed]
31. Zhou, H.; Cao, H.; Matyunina, L.; Shelby, M.; Cassels, L.; McDonald, J.F.; Skolnick, J. MEDICASCY: A Machine Learning Approach for Predicting Small-Molecule Drug Side Effects, Indications, Efficacy, and Modes of Action. *Mol. Pharm.* **2020**, *17*, 1558–1574. [CrossRef] [PubMed]
32. Mohsen, A.; Tripathi, L.P.; Mizuguchi, K. Deep Learning Prediction of Adverse Drug Reactions in Drug Discovery Using Open TG–GATEs and FAERS Databases. *Front. Drug Discov.* **2021**, *1*. [CrossRef]
33. Jiang, M.; Zhou, B.; Chen, L. Identification of drug side effects with a path-based method. *Math. Biosci. Eng.* **2022**, *19*, 5754–5771. [CrossRef] [PubMed]
34. Liang, X.; Fu, Y.; Qu, L.; Zhang, P.; Chen, Y. Prediction of drug side effects with transductive matrix co-completion. *Bioinformatics* **2023**, *39*, btad006. [CrossRef]
35. SIDER Side Effect Resource. Available online: http://sideeffects.embl.de/ (accessed on 16 August 2023).
36. Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B.A.; Thiessen, P.A.; Yu, B.; et al. PubChem 2019 update: Improved access to chemical data. *Nucleic Acids Res.* **2019**, *47*, D1102–D1109. Available online: https://academic.oup.com/nar/article-abstract/47/D1/D1102/5146201 (accessed on 6 June 2023). [CrossRef]
37. Wishart, D.S.; Feunang, Y.D.; Guo, A.C.; Lo, E.J.; Marcu, A.; Grant, J.R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074–D1082. Available online: https://academic.oup.com/nar/article-abstract/46/D1/D1074/4602867 (accessed on 6 June 2023). [CrossRef]

38. Brooks, T.; Holynski, A.; Efros, A.A. InstructPix2Pix: Learning to Follow Image Editing Instructions. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 18392–18402. Available online: https://arxiv.org/abs/2211.09800v2 (accessed on 1 August 2023).
39. Mirza, M.; Osindero, S. Conditional Generative Adversarial Nets. Available online: https://arxiv.org/abs/1411.1784v1 (accessed on 1 August 2023).
40. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. Available online: http://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html (accessed on 23 March 2023).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.