An Intelligent Deep Learning System for Drug Side Effect Prediction

Murali Kanthi[1], V. Gunali[2], V. Sainath[3], B. Shankar Naik[4]
[1]Associate Professor, Dept of CSE (Data Science)
[2, 3, 4]Dept of CSE (Data Science)
[1, 2, 3, 4] CMR Technical Campus

Abstract- *Adverse effects in clinical trials pose serious health risks and financial burdens. Predictive algorithms for side effects can mitigate these issues by informing early drug development. The LINCS L1000 dataset, with its extensive drug-perturbed gene expression data, is a valuable resource for such predictions. However, many current methods only use a limited subset of high-quality experiments, overlooking much of the data. This study leverages the full LINCS L1000 dataset and evaluates five deep learning architectures. A multi-modal model combining drug chemical structure (CS) and gene expression (GEX) shows superior performance among MLP-based methods, with CS features proving more informative. Additionally, a CNN using only SMILES representations achieves the best overall results, improving macro-AUC by 13.0% and micro-AUC by 3.1% over existing approaches. The model also identifies previously unreported drug-side effect associations found in the literature.*

I. INTRODUCTION

Computational methods have shown great potential in reducing the health risks and financial costs associated with drug development by enabling early prediction of side effects before clinical trials. Numerous machine learning approaches have been proposed that utilize features such as drug chemical structures, drug-protein interactions, protein-protein interactions (PPIs), metabolic network activity, biological pathways, phenotypic data, and gene annotations [1].

More recently, deep learning techniques have emerged in this domain. For example, [2] integrates biological, chemical, and semantic information from drugs alongside clinical notes and case reports, while [3] employs deep learning to extract and evaluate various chemical fingerprints for side effect prediction.

Although these models have demonstrated success in predicting adverse drug reactions (ADRs)—used interchangeably with drug side effects—they primarily rely on external drug knowledge and lack cell-type or dosage-specific context. To address this limitation, Wang et al. (2016) utilized the LINCS L1000 dataset [4], which captures gene expression changes in a wide range of human cell lines treated with various drugs and small molecules. Their work marked the first large-scale, unbiased, and cost-effective attempt at ADR prediction using drug-induced gene expression profiles.

Framing the task as a multi-label classification problem, their results highlighted the value of context-specific information. However, their method used only the highest-quality experiment for each drug to reduce noise, thereby leaving a large portion of the 473,647 experiments across 20,338 compounds unexploited. Additionally, their model required manual feature engineering by converting gene expression data into biological term enrichment vectors.

In this study, we propose Deep Side, a deep learning framework for ADR prediction that eliminates the need for manual feature engineering. Deep Side leverages (i) full in vitro gene expression profiling experiments (GEX) along with experimental metadata (such as cell line and dosage), and (ii) the chemical structures of compounds (CS). Our models are trained on the complete LINCS L1000 dataset and validated using the SIDER database as the reference for drug-ADR labels [5]. This approach enables a more comprehensive and context-aware prediction of drug side effects using deep learning. We experiment with five architectures: (i) a multi-layerperceptron (MLP), (ii) MLP with residual connections (Res MLP), (iii) multi-modal neural net- works (MMNN. Concat and MMNN. Sum), (iv) multi-task neural network (MTNN), and finally, (v) SMILES convolutional neural network (SMILES Conv).

We conduct a comprehensive evaluation of the proposed architectures, analyzing the impact of various input features on predictive performance. Our experiments reveal that chemical structure (CS) features are particularly strong predictors of drug side effects. The baseline MLP model using only CS features yields an improvement of approximately 11% in macro-AUC and 2% in micro-AUC over the state-of-the-art method in [6], which combines high-quality gene expression (GEX) data with CS features.

Among MLP-based models, the multi-modal neural network (MMNN.Sum), which integrates CS, GEX, and metadata (META) using summation in the fusion layer, achieves the highest performance—0.79 macro-AUC and 0.877 micro-AUC. Our results also indicate that when chemical structure information is fully leveraged within a complex deep learning architecture, it can surpass the predictive value of gene expression data alone.

Furthermore, a convolutional neural network trained solely on SMILES string representations of drug structures delivers the best overall performance, achieving a 13.0% improvement in macro-AUC and a 3.1% gain in micro-AUC over existing state-of-the-art approaches.

II. EXISTING SYSTEM

A drug-drug interaction (DDI) occurs when the pharmacological effect of one drug is altered by the presence of another. While some DDIs can enhance therapeutic outcomes, negative interactions are a major cause of adverse drug reactions, potentially leading to drug withdrawals and even patient fatalities. As such, accurate identification of DDIs is a critical aspect of both drug development and clinical treatment.

This study presents an existing system, DDI-IS-SL, which predicts DDIs using an integrated similarity approach combined with semi-supervised learning. DDI-IS-SL incorporates chemical, biological, and phenotypic data to compute feature similarities between drugs using cosine similarity. Additionally, Gaussian Interaction Profile (GIP) kernel similarity is calculated based on known DDI data. The interaction probabilities between drug pairs are then estimated using a semi-supervised learning algorithm—specifically, the Regularized Least Squares (RLS) classifier.

Experimental evaluations using 5-fold and 10-fold cross-validation, as well as de novo drug validation, show that DDI-IS-SL outperforms several baseline methods in prediction accuracy. Moreover, the system demonstrates lower average computation time compared to other approaches. Case studies further validate the practical utility and effectiveness of DDI-IS-SL in real-world DDI prediction scenarios.

Disadvantages:

- The complexity of data: Most of the existing machine learning models must be able to accurately interpret large and complex datasets to detect an accurate Drug Side Effect.
- Data availability: Most machine learning models require large amounts of data to create accurate predictions. If data is unavailable in sufficient quantities, then model accuracy may suffer.
- Incorrect labeling: The existing machine learning models are only as accurate as the data trained using the input dataset. If the data has been incorrectly labeled, the model cannot make accurate predictions.

III. PROPOSED SYSTEM

Multi-Layer Perceptron (MLP):

The MLP model [7] processes a concatenated vector of all input features using a sequence of fully connected (FC) layers. Each FC layer is followed by batch normalization [10], ReLU activation [9], and dropout regularization with a dropout rate of 0.2 [8]. The final layer uses a sigmoid activation to output the probability of adverse drug reactions (ADRs) for each class. The model is trained using multi-label binary cross-entropy (BCE) loss, calculated as the sum of negative log-probabilities across all ADR classes. The architecture is demonstrated for drug chemical structure (CS) and gene expression (GEX) inputs.

Residual Multi-Layer Perceptron (ResMLP):

The ResMLP extends the basic MLP by incorporating residual connections between layers. In this design, the input to each intermediate FC layer is added element-wise to its output before being passed to the next layer. This technique helps mitigate the vanishing gradient problem 11], enabling the use of deeper networks that can learn more complex and efficient representations.

Multi-Modal Neural Networks (MMNN):

The MMNN architecture consists of individual MLP sub-networks, each dedicated to learning features from a specific data modality (e.g., CS, GEX, or metadata). The outputs from these sub-networks are fused before classification. Two fusion strategies are explored:

MMNN.Concat: Concatenates outputs from each modality into a single feature vector.
MMNN.Sum: Applies element-wise summation, requiring each sub-network to output vectors of the same size.

Multi-Task Neural Network (MTNN):

This multi-task learning (MTL) architecture leverages the hierarchical taxonomy of side effects from

ADReCS. It includes a shared MLP block that processes concatenated CS and GEX features into a unified embedding. This embedding is then passed through multiple task-specific MLP branches, each responsible for predicting a subset of interrelated ADR classes. This structure enables the model to learn both global and task-specific patterns.

Advantages:

- Diverse Classifier Evaluation: The framework implements and compares multiple machine learning classifiers, offering comprehensive insights into their performance on the dataset.
- Effective Deep Feature Learning: The use of convolutional neural networks (CNNs), known for their ability to learn complex features-especially in image and structured data tasks-contributes to improved prediction accuracy.
- Multi-Modal and Multi-Task Learning: The system supports multi-modal inputs and multi-task learning strategies, enhancing both robustness and generalization across different ADR categories.

IV. SYSTEM ARCHITECTURE

The following modules are used in the system architecture for implementing prediction model.

Service Provider: In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as Train & Test Drug Data Sets, View Trained and Tested Drug Datasets Accuracy in Bar Chart, View Trained and Tested Drug Datasets Accuracy Results, View Drug Side Effect Prediction Type, Find Drug Side Effect Prediction Type Ratio, Download Predicted Data Sets, View Drug Side Effect Prediction Type Ratio Results, View All Remote Users. View and Authorize Users: In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

Remote User: In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGIN, PREDICT DRUG SIDE EFFECT TYPE, and VIEW YOUR PROFILE.

View and Authorize Users: In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

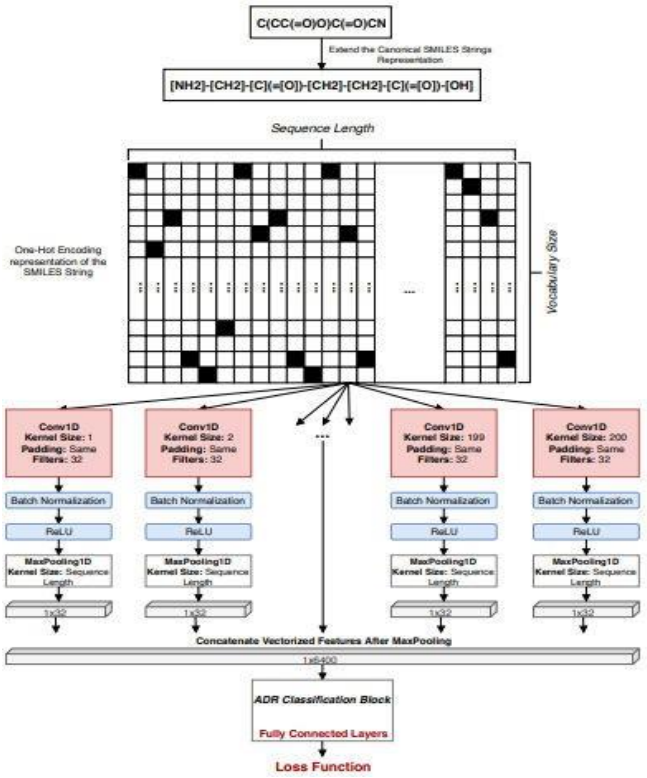

Figure 1: System Architecture

V. ALGORITHIMS USED

Gradient Boosting: Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees.[12][13] When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees; it usually out performs random forest. A gradient-boosted trees model is built in a stage-wise fashion as in other boosting methods, but it generalizes the other methods by allowing optimization of an arbitrary differentiable loss function.

Logistic regression Classifiers: Logistic regression analysis studies [14][15] the association between a categorical dependent variable and a set of independent (explanatory) variables. The name logistic regression is used when the dependent variable has only two values, such as 0 and 1 or Yes and No. The name multinomial logistic regression is usually reserved for the case when the dependent variable has three or more unique values, such as Married, Single, Divorced, or Widowed. Although the type of data used for the dependent variable is different from that of multiple

regressions, the practical use of the procedure is similar. Logistic regression competes with discriminant analysis as a method for analyzing categorical-response variables. Many statisticians feel that logistic regression is more versatile and better suited for modeling most situations than is discriminant analysis. This is because logistic regression does not assume that the independent variables are normally distributed, as discriminant analysis does. This program computes binary logistic regression and multinomial logistic regression on both numeric and categorical independent variables. It reports on the regression equation as well as the goodness of fit, odds ratios, confidence limits, likelihood, and deviance. It performs a comprehensive residual analysis including diagnostic residual reports and plots. It can perform an independent variable subset selection search, looking for the best regression model with the fewest independent variables. It provides confidence intervals on predicted values and provides ROC curves to help determine the best cut off point for classification. It allows you to validate your results by automatically classifying rows that are not used during the analysis.

SVM: In classification tasks a discriminant machine learning technique aims at finding, based on an independent and identically distributed training dataset, a discriminant function that can correctly predict labels for newly acquired instances. Unlike generative machine learning approaches, which require computations of conditional probability distributions, a discriminant classification function takes a data point *x* and assigns it to one of the different classes that are a part of the classification task. Less powerful than generative approaches, which are mostly used when prediction involves outlier detection, discriminant approaches require fewer computational resources and less training data, especially for a multidimensional feature space and when only posterior probabilities are needed.From a geometric perspective, learning a classifier is equivalent to finding the equation for a multidimensional surface that best separates the different classes in the feature space. SVM is a discriminant technique, and, because it solves the convex optimization problem analytically, it always returns the same optimal hyperplane parameter—in contrast to genetic algorithms (GAs) or perceptrons, both of which are widely used for classification in machine learning. For perceptrons, solutions are highly dependent on the initialization and termination criteria. For a specific kernel that transformsthedata from the input space to the feature space, training returns uniquely defined SVM model parameters for a given training set, whereas the perceptron and GA classifier models are different each time training is initialized. The aim of GAs and perceptrons is only to minimize error during training, which will translate into several hyperplanes' meeting this requirement.

RESULTS

Figure 2: Home Page

Figure 3: Trained and Tested Drug Datasets Accuracy Results

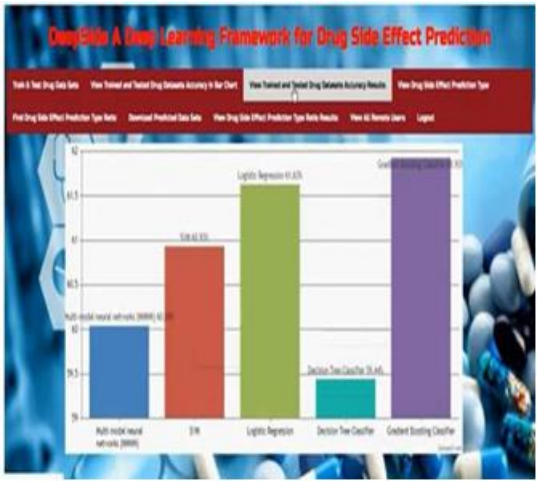

Figure 4: Trained and Tested Drug Datasets Accuracy Results in Bar Chart

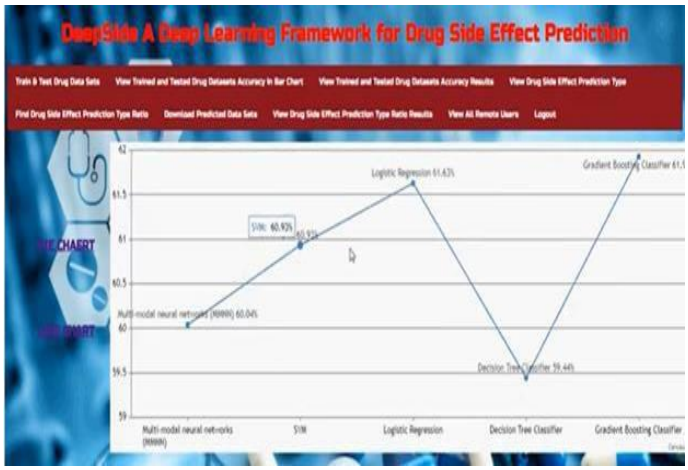

Figure 5: Trained and Tested Drug Datasets Accuracy Results in Line Chart

Figure 6: View Drug Side Effect Prediction Type Ratio Results

VI. CONCLUSION

The pharmaceutical drug development process is a long and demanding process. Unforeseen ADRs that arise at the drug development process can suspend or restart the whole development pipeline. Therefore, the a priori prediction of the side effects of the drug at the design phase is critical. In our Deep Side framework, we use context-related (gene expression) features along with the chemical structure to predict ADRs to account for conditions such as dosing, time interval, and cell line. The proposed MMNN model uses GEX and CS as combined features and achieves better accuracy performance compared to the models that only use the chemical structure (CS) finger- prints. The reported accuracy is noteworthy considering that we are only trying to estimate the condition-independent side effects. Finally, SMILES Conv model outperforms all other approaches by applying convolution on SMILES representation of drug chemical structure.

REFERENCES

[1] Atias, N., Sharan, R.: An algorithmic framework for predicting side e_ects of drugs. Journal of Computational Biology18(3), 207{218 (2011).

[2] Bresso, E., Grisoni,R., Marchetti,G., Karaboga, A.S.,Souchet, M.,Devignes, M.D., Sma•_l-Tabbone, M.: Integrative relational machine-learning for understandingdrugside-e_ectpro_les.BMC bioinformatics 14(1), 207(2013).

[3] Cai,M.C.,Xu, Q.,Pan,Y.J., Pan,W., Ji, N.,Li, Y.B., Jin, H.J.,Liu,K., Ji,Z.L.: Adrecs:anontologydatabase for aiding standardization and hierarchical classi_cationof adverse drugreactionterms. Nucleic acids research 43(D1), D907{D913(2014).

[4] Dey, S., Luo, H., Fokoue, A., Hu, J., Zhang, P.: Predicting adverse drug reactions through interpretable deep learning framework. BMCBioinformatics 19 (12 2018).

[5] Dimitri, G.M., Li_o, P.: Drugclust: A machine learning approach for drugs side effects prediction. Computational Biology and Chemistry 68, 204 (2017).

[6] Groopman, J.E., Itri, L.M.: Chemotherapy-inducedanemiainadults: incidence and treatment. Journal of the NationalCancerInstitute91(19),1616{1634(1999)

[7] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016).

[8] Huang,L.C.,Wu,X.,Chen,J.Y.: Predicting adverse side e_ects of drugs. BMC genomics 12(5), S11 (2011).

[9] Huang,L.C.,Wu,X.,Chen,J.Y.: Predicting adverse drug reaction pro_les by integrating protein interaction networks with drug structures. Proteomics 13(2), 313{324 (2013).

[10] Io_e, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)

[11] 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.: Pubchem substance and compound databases. Nucleic acids research 44(D1), D1202{D1213 (2015)

[12] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classi_cation with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097{1105 (2012)

[13] Kuhn,M.,Letunic,I.,Jensen,L.J.,Bork, P.: The sider database of drugs and side e_ects. Nucleic acids research 44(D1),D1075{D1079(2015)}

[14] Landrum, G., et al.: Rdkit: Open-source cheminformatics (2006).

[15] Lee, W., Huang, J., Chang, H., Lee, K., Lai, C.: Predicting drug side e_ects using data analytics and the integration of multiple data sources. IEEE Access 5, 20449{20462 (2017).