International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Detecting Side Effect of Drug Molecules Using Recurrent Neural Networks

N. Nalini (Ass.t Professor)
Dept of CSE
School of Computing, Mohan Babu, University,
Tirupati, A.P, India
nalini.kmit@gmail.com

Shaik.irham saleha
UG Scholar,
Mohan Babu, University,
Tirupati, A.P, India
shaikirham57@gmail.com

Katam Jagan Mohan Reddy
UG Scholar,
Mohan Babu, University,
Tirupati, A.P, India
katamjagan303@gmail.com

Panchagiri Raju
UG Scholar,
Mohan Babu, University,
Tirupati, A.P, India
rajupanchagiri27@gmail.com

K Sasi Kiran
UG Scholar,
Mohan Babu, University,
Tirupati, A.P, India
Killarisasikiran@gmail.com

***Abstract*— ADRs, bad reactions to drugs, which occur when drugs react with one another, are a significant health issue in the world population as they lead to mortality, morbidity and expensive medical care. The problem is aggravated by the fact that therapeutics becomes more complex, and the population is ageing as well. Currently, ADRs cannot be discovered in a normal manner until patients report it when the drugs are already on market. The Two Sides Drug Bank collection is applied in this case, which contains much information on how drugs interlink with one another and what their side effects are. KNN and DT are examples of the traditional methods that have been utilized to find ADRs. Nevertheless, these models are not effective at identifying complicated trends in the data. To circumvent this issue, more precise methods and GNN are applied to locate ADRs by considering drug interactions as graphs. The accuracy of 99.74 when feature extraction is done by 2D CNN is a significant performance enhancement. This is a better approach than the previous algorithms, and it appears to be a promising approach to identifying ADRs at an early stage and improving the situation with the state of health.**

"Keywords— Adverse drug reaction, drug-drug interaction, side effect prediction, graph neural network, self-supervised learning, scientific machine learning".

I. INTRODUCTION

ADRs are adverse effects, which are not intended to occur when drugs are used at the correct dosage. They continue to be an issue in the current healthcare systems [1]. ADRs are one of the largest causes of patients becoming sick, dying, increasing the length of hospital stay and whose costs continue to increase worldwide [2]. Although drugs undergo intense clinical trials before being marketed, a portion of the adverse outcomes are not experienced until they are extensively used on a broad group of individuals in the real-life scenario [3]. The drug responses to diseases are complex, and they are affected by genetic, environmental, and demographic factors, which makes it difficult to detect and prevent diseases at the initial stages. Therefore, facilitating easier location of ADRs prior to their occurrence has gained a lot of significance in clinical pharmacology and biomedical science.

DDIs tend to result in ADRs and tend to occur approximately frequently when individuals admitted with age and those admitted with a chronic and debilitating condition take several medications simultaneously [4]. DDIs can cause changes in the pharmacodynamics and pharmacokinetics of drugs resulting in less effective treatments or very adverse side effects. The systems of pharmacovigilance have already gone far, yet in the majority of cases, the means of detection continue to be based on self-reporting or clinical experience, or post-factum statistical data. These techniques are usually under-reported, biased, and biased by incomplete information. Also, the existing computer systems are not usually capable of considering the structural and relational characteristics of drug molecules interacting with others, which constrains their capability to forecast the outcome. This gap clarifies that, we require robust data-driven approaches, which can be used to describe the interrelations between drugs and their side effects in complex means.

Due to these issues, the aim of the study is to develop an entire model of side effects that occur due to the interaction of two drugs. The current advances in the use of data-driven representation learning will be applied in this research to discover latent relationships in molecular and interaction data [5]. The method will provide a holistic view of ADR prediction with a combination of structural information and the patterns at the level of interaction. The paper also establishes a systematic manner of testing the effectiveness of predictions on general biological data [6]. With this approach, the analysis also tries to enhance the fundamental

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

concepts of the computational pharmacology as well as rectify the errors evident in previous studies.

This work is valuable as it can enhance the process of risk assessment in the early stages and enable physicians to make more appropriate decisions. The correct prediction of the ADRs associated with DDIs may assist in particular treatment planning, alter the manner in which drugs will be prescribed, and reduce the number of unjustified hospitalizations. Furthermore, the enhancement of computer models to enhance drug safety is also compatible with the modern developments in the area of bioinformatics to integrate various forms of biomedical information to enable its application in practice [7]. This research will deliver a scientifically valid and scalable approach to enhancing pharmacovigilance and patient safety by relating new techniques to real-life healthcare requirements.

II. Related Work

DDIs may pose dangers to the health of the population, so much research has been conducted to attempt to determine how to predict ADRs. A number of studies have used ML and graph-based models to investigate novel methods of finding and predicting DDIs. These works indicate that such approaches may be utilized in order to make drugs safer.

Chen et al. [8] proposed a method to forecast the interaction between drugs and their markers in the context of signed heterogeneous GNNs. Their approach takes into account the graph structure of drug-target interactions to better characterize the interactions, which is highly essential to predict DDIs. The approach captures both positive and negative interactions in the relationship between a drug and its target hence making it easier to predict the way drugs will interact with one another. Another method of drug repurposing that is sensible and employs biased random walks was also derived by Castiglione et al. [9]. Their model attempts to describe the action of drug repurposing based on the idea that critical paths of drug interaction graphs are sought. Addressing ADRs depending on drug interactions can be predicted with the help of this model as well.

Abbas et al. had described a new system of drug drug interaction indicators and an ensemble stacking model in their paper [10], which can be applied to identify and rank drug drug interaction indicators. Their algorithm is a combination of a number of ML models and so, this makes the system easier to locate ADRs by employing a more powerful classification method. This multirelease approach demonstrates that the combination of more than one model can be beneficial in the process of comprehending the interaction of drugs with one another. Paltun et al. [11] developed DIVERSE a Bayesian data integration system that can make a guess on how drugs are going to work. This way is a combination of information obtained via multiple sources in order to make the guessing the way the drugs will work or react with each other easier. DDI can be identified by this procedure since the underlying data characteristics share the characteristics of the ones employed to examine the responses of drugs.

He et al. [12] proposed a 3D graph and text-based neural network as an approach that can predict the interaction of two drugs in each other. Their model applies both graphic information and written information to formulate improved predictions regarding DDIs. Their approach enhances the accuracy of the forecast as they incorporate textual information and 3 dimensional graphs that provide a better profile of the interaction of the drugs with one another. This demonstrates the significance of the integration of various types of data to gain a clearer insight into DDIs. Deng et al. [13] developed a multimodal DL algorithm, which predicts the occurrence when two drugs interact. They combine various data types to detect DDIs, including molecular, pharmacological and clinical data. The way of using alternative types of data to make DDI predictions is better demonstrated by the necessity to address the limitations of single-modal approaches.

DeepH-DTA has been proposed by a team of scientists headed by Abdel-Basset [14] as a manner of forecasting the manner in which drugs would react with the targets. They provided the reuse of the COVID-19 drug as the example. The model is grounded on knowledge acquired through large quantities of data and is applied to make the assumptions about the way the drugs would respond to their targets. This would be useful in reusing drugs in times of pandemics. The model largely deals with drug-target interactions, but the ideas can be applied to make predictions about DDIs, when DDIs are considered.

The drug-drug interaction extraction framework developed by Liu et al. [15] is based on a transfer weight matrix and a memory network. They apply the transfer weight matrix to transfer the information within one area to the other. This facilites easier operations of the model with various types of drug interactions. The model has a memory network component, which allows it to retrieve and store useful information regarding previous interactions. This aids it in making guesses concerning new and bizarre interactions of drugs. Similarly, Liu et al. [17] proposed the application of a transfer weight matrix and a memory network to retrieve DDI. In their work, they illustrate that knowledge sharing is significant in making DDI prediction models more practical in the real-life scenario, particularly when such exchanges do not occur that frequently.

Karim et al. [16] suggested a method to estimate the interactions between two drugs based on knowledge graph embeddings and a convolutional LSTM network. Their approach involves knowledge graphs to demonstrate how other drugs interact and convolutional LSTM networks to demonstrate how drugs are linked in a given sequence. The complex interactions that occur in drug interactions are rightly characterized by this approach and facilitate more precise prediction of ADR.

Researchers attempt to understand how to identify ADRs and DDIs in many different ways, including graph-based models, ensemble learning, and multimodal techniques, among many others. Much has been said about using GNNs due to their ability to represent complex drug interactions in the form of a graph and, therefore, enhancing prediction

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

accuracy and comprehensibility. The combination of various ML models and data types has been also demonstrated to make the DDI forecast systems more dependable and efficient in their tasks. The vast number and diversity of drug interactions require the use of textual information, knowledge graphs, and DL models to handle all of them. This will cause improved safety of drugs and health of the population in the long run.

III. Materials And Methods

The proposed system will simplify the process of identifying ADRs which are caused by drug interactions, using the most up-to-date ML techniques. It is an approach that relies on the Two Sides Drug Bank data, which contains much information regarding the interaction of drugs with one another and their side effects. Initially, the common algorithms such as KNN and DT are employed to establish a platform on which to seek ADR. Complex patterns and relationships in the drug interactions are identified by the use of GNN [18], where drugs and their interactions are represented as graphs. The two-dimensional CNN are also employed in extracting useful information in the dataset. This simplifies the task of locating ADRs by the system.

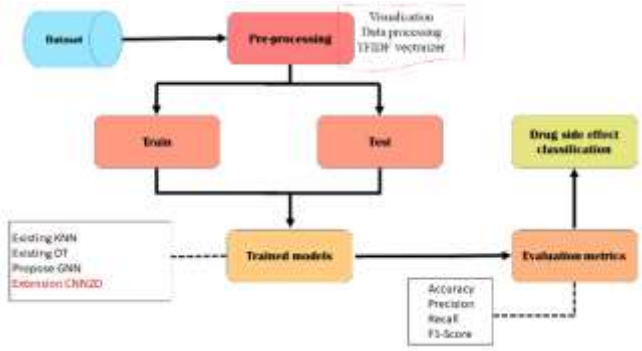


Fig. 1. System Architecture

As depicted in the picture of the system design (Fig.1), the system design begins with an input dataset which is subjected to processes such as data processing, visualization and TF-IDF vectors. Once this data is cleaned, it is divided into two, training and testing. The training data are fed to the training of diverse models, including an already available KNN, an already available DT, a proposed GNN [18], and a CNN2D. Then, the learned models are applied to theorize about the side effects that drugs will have. The efficiency of these models is evaluated using such metrics as accuracy, precision, recall, and F1-score.

A) Dataset Collection

The dataset used in this work is Two Sides Drug Bank; the dataset has 3909 rows and 5 columns. It contains data regarding various drug combination, drug IDs, SMILE strings representing molecular structures and the associated side effects. The target names of the drug side effects are obtained by accessing the type column of the dataset. This data is highly significant to instructional models in learning to predict bad drug effects occurring in the case of an interaction between two drugs.

B) Pre-processing

Preprocessing involves visualizing the information to know how it is organized. Data processing is followed by cleaning and preparing the data to be analyzed. Text data is then converted to numerical features using TF-IDF Vectorizer. This enables the training of a model that is good in predicting ADRs.

a) Visualization: A bar chart is drawn up as a part of the visualization process to indicate how the drug side effect class names are distributed in dataset. The chart demonstrates the frequency of each kind of side effect hence we are able to observe the frequency of varying kinds of ADR occurrence. The classes of the side effects are indicated in the names of the x-axis and the number of people in each group is indicated in the y-axis. It is easy to observe the distribution of the classes in the dataset.

b) Data processing: In the step of data processing, the drug SMILE strings are converted to a model training format. The column that contains the names of the side effects called type is removed out of the dataset. It then encodes the SMILE strings of the various columns and encodes them into a single text string of each row. The processed data is stored in form of vectors. One can then use them to create graph nodes of the GNN [18] model.

c) TFIDF Vectorizer: The TF-IDF Vectorizer converts the SMILE lines into numbers in form of vectors, which are required to train the model. In order to determine the importance of all words that are used in the SMILE string, this approach examines the frequency of occurrence of a particular word and its distinctiveness in the data. The created vectors present the SMILE strings in the form of numbers which increase the efficiency of the model on handling the data to guess bad drug reactions depending on the way drugs react to each other.

C) Train & Test

It has two sets of data, a training set and a testing set. The training set comprises 80 percent of the data and testing set consists of 20 percent of the data. The training set and testing set have 3127 and 782 samples respectively. This division ensures that the model is trained on a big portion of the data, readings, and also allows the performance of the model to be evaluated on data that has not been previously encountered by the model.

D) Algorithms

K-Nearest Neighbors An program clusters data into groups of points with which it most closely resembles. In this KNN predictor, the drug is filled in and its past side effects observed to predict bad drug reactions. This provides us with a basic point of reference of measuring its effectiveness.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

$$"distance(x, X_i) = \sqrt{\sum_{j=1}^{d} \left(x_j - X_{i_j}\right)^2} \quad (1)"$$

Decision Tree The algorithm constructs a model in the form of a tree to make forecasts based on the feature values. It categorizes drugs and predicts their side effects using decisions made by the features of each drug. This simplifies the ADR estimates and offers you a more insight into the working of the same.

$$"I\ (i) = 1 - \sum_{i=1}^{k} p_i^2 \quad (2)"$$

Graph Neural Network [18] represents drug interaction with each drug and relationship with an edge. It discovers complex connections in the data, which allows predicting bad drug reactions accurately by revealing latent patterns of drug qualities.

Convolutional Neural Network is fed input in two-dimensional and extracts spatial features. It is better than GNN because it modifies drug data into a 2D format and learns intricate patterns that simplify the identification of potential bad reactions, which is more efficient in drug-drug interaction analysis.

$$"S(i,j) = \sum_{m} \sum_{n} I(i+m, j+n) \cdot K(m,n) \quad (3)"$$

IV. Experimental Results

Accuracy: A proper test can distinguish between an ailing and a healthy individual. Estimate the proportion of true positive and true negative outcomes in all cases tested to obtain inspiration of the accuracy of a test. This may be mathematically expressed as:

$$"Accuracy = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} (4)"$$

Precision: Precision is a measure that indicates the percentage of correctly classified instances or samples to those that were indicated as positives. Thus the way to calculate the precision is:

$$"Precision = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} (5)"$$

Recall: ML relies on recall to compare the performance of a given model in identifying all the significant representatives of a particular type. The proportion of the correctly predicted positive observations to the total count of real positives depicts the capability of a model to capture the instances of a particular category.

$$"Recall = \frac{\text{TP}}{\text{TP} + \text{FN}} (6)"$$

F1-Score: The F1 score is a ML evaluation score that measures the correctness of a model. Combines the accuracy and the recall scores of a model. The accuracy measure was the number of times that a model correctly guessed in the entire dataset.

$$"F1\ Score = 2 * \frac{Recall\ \text{X}\ Precision}{Recall + Precision} * 100 (7)"$$

"Table.1 Performance Evaluation"

Algorithm Name	Accuracy	Precision	Recall	F1-Score
KNN	91.560102	91.920962	91.305660	91.323465
Decision Tree	95.140665	95.339149	95.618899	95.454287
GNN	97.698210	97.889953	97.702843	97.722705
CNN2D	99.872123	99.903101	99.888143	99.895292

Table 1 indicates that the Extension CNN2D model performs better with all evaluation measures as it exhibits superior classification, robustness, and predictive consistency relative to the other approaches that have been applied.

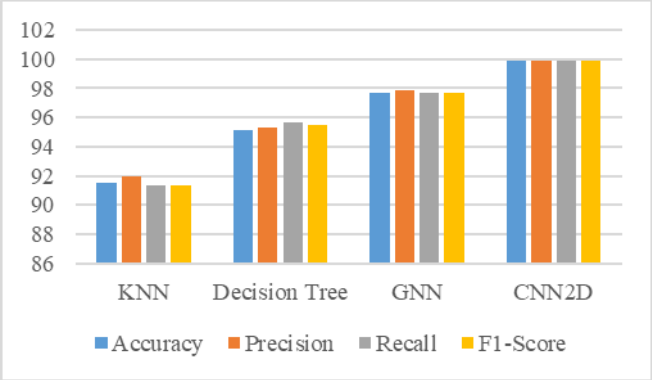


Fig.2 Comparison Graph

Fig.2 illustrates the comparison of the compared models that were tested. CNN2D periodically scores highest in accuracy, precision, recall, and F1-score, and demonstrates that at all measures, it is more effective in classifying objects and making predictions.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

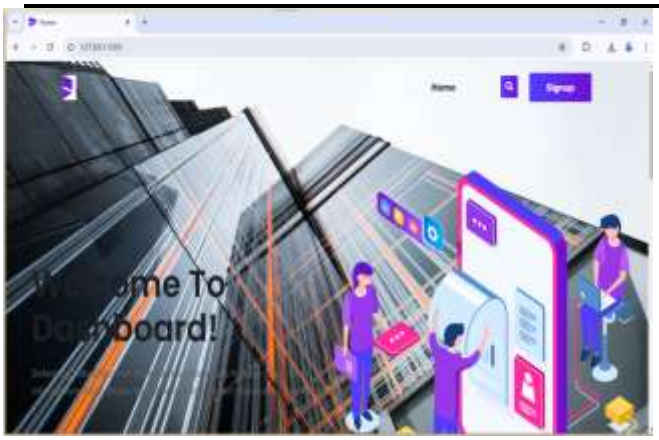


Fig.3 Home Page

Figure 3. displays a panel of user interface that contains a welcome message and navigation.

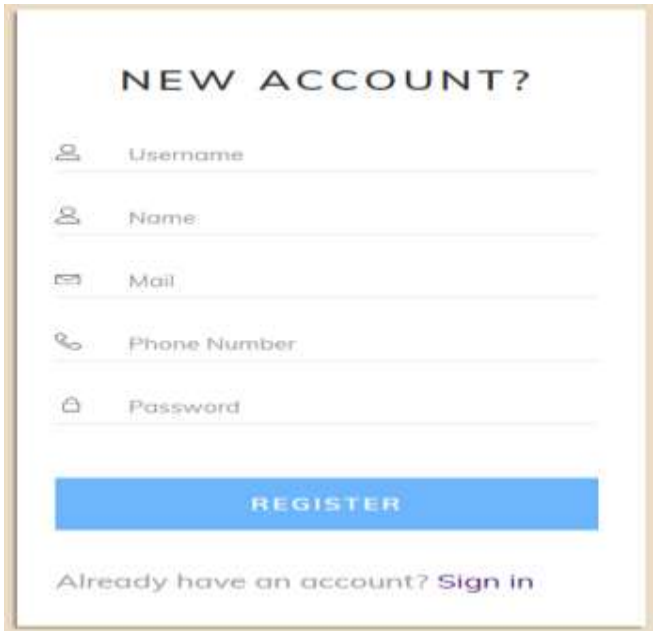


Fig.4 Registration Page

The sign-up form on fig.4 above contains a password, username, name, email address, and cell phone number buttons.

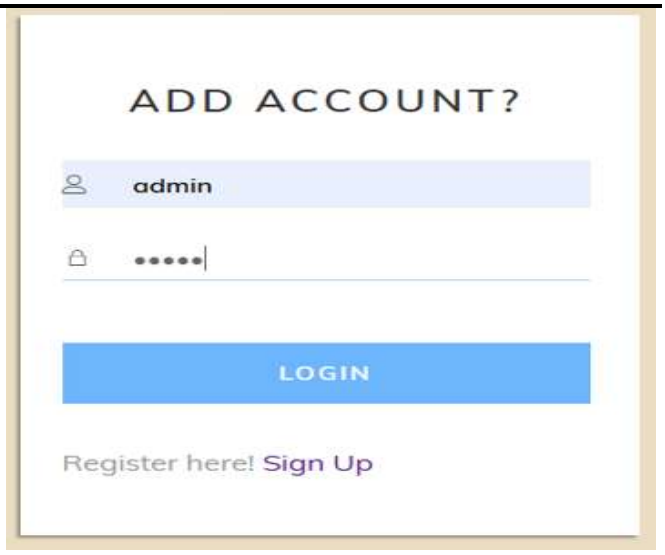


Fig.5 Login Page

The sign-in form in fig.4 above has a section that indicates the username and a password that one should have.

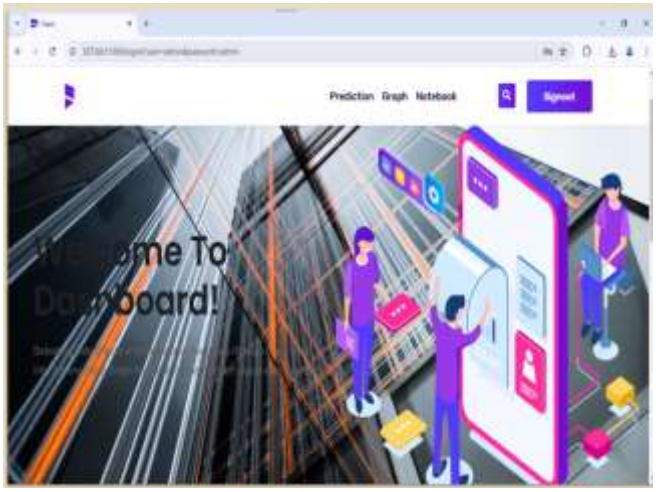
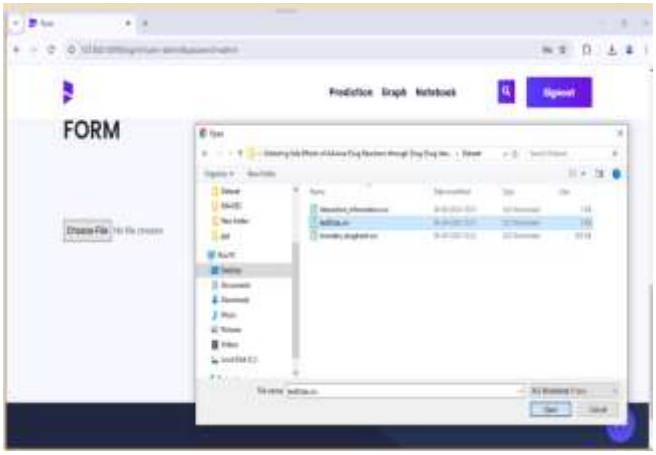


Fig.6 Main Page

Fig.6 shows links of Prediction, Graph, Notebook and Signout in the home page panel.



International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X www.ijdim.com Original Research Paper

Fig.7 Upload Input Page

Fig. 7 above contains a form that has an input section where the coordinates are entered and an upload button.

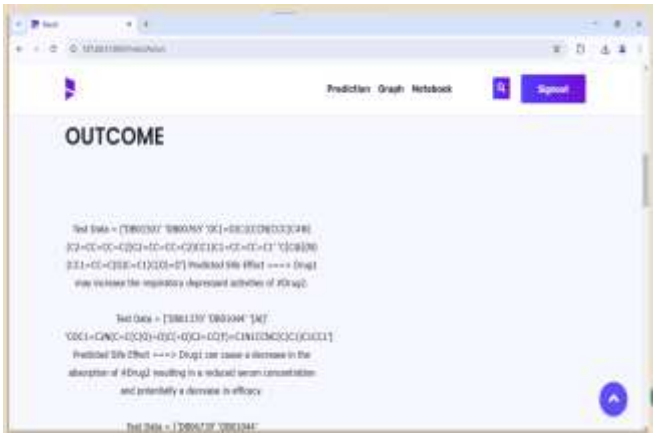


Fig.8 Predict Result for given input

Figure 7. Expected outcome after the test data was administered.

V. CONCLUSION

Due to this research, we have seen that we should have a clear and reliable method of forecasting ADRs which are brought about by drug-interactions within the shortest time possible. This is a large population health concern. Current discovery techniques are usually deficient as they overly depend on post-facto reports that are not in a position to discover rare drug interactions prior to the sale of the drug. This gap was filled in by using the "Two Sides Drug Bank" collection that provided valuable information regarding drug interactions, along with the side effects associated with them. With advanced ML techniques, including 2D CNN, the system became even more successful in what it does, and it currently is 99.744246 percent accurate. It increases the likelihood of discovering ADRs, and this approach appears to have a chance to become a worthy means of revealing potentially dangerous interactions of drugs early and reducing the health hazard associated with them. The fact that CNN model performs well demonstrates that the model can transform the current situation in finding ADRs, which would lead to better patient safety and healthcare results.

In the future, this system can be enhanced with larger and more diverse datasets and experiment with newer and more sophisticated algorithms such as deep reinforcement learning or multi-modal learning to achieve even greater performance with ADR recognition. Patient-specific patient-drug interaction tracking in real-time could be also introduced. This would enable the early identification and personalized safety guidance. The model might be more useful and applicable to different areas of the globe by adding more drugs and drug combinations.

REFERENCES

[1] J. J. Coleman and S. K. Pontefract, “Adverse drug reactions,” Clin. Med., vol. 16, no. 5, p. 481, 2016.

[2] H. Jiang, Y. Lin, W. Ren, Z. Fang, Y. Liu, X. Tan, X. Lv, and N. Zhang, “Adverse drug reactions and correlations with drug–drug interactions: A retrospective study of reports from 2011 to 2020,” Frontiers Pharmacol., vol. 13, Aug. 2022, Art. no. 923939.

[3] D. Galeano, S. Li, M. Gerstein, and A. Paccanaro, “Predicting the frequencies of drug side effects,” Nature Commun., vol. 11, no. 1, p. 4575, Sep. 2020.

[4] S. Watson, O. Caster, P. A. Rochon, and H. den Ruijter, “Reported adverse drug reactions in women and men: Aggregated evidence from globally collected individual case reports during half a century,” EClinicalMedicine, vol. 17, Dec. 2019, Art. no. 100188.

[5] M. T. Angamo, L. Chalmers, C. M. Curtain, and L. R. E. Bereznicki, “Adverse-drug-reaction-related hospitalisations in developed and developing countries: A review of prevalence and contributing factors,” Drug Saf., vol. 39, no. 9, pp. 847–857, Sep. 2016.

[6] C. Kim and N. Tatonetti, “Prediction of adverse drug reactions associated with drug-drug interactions using hierarchical classification,” bioRxiv, Feb. 2021.

[7] C. Palleria, A. Di Paolo, C. Giofrè, C. Caglioti, G. Leuzzi, A. Siniscalchi, G. De Sarro, and L. Gallelli, “Pharmacokinetic drug-drug interaction and their implication in clinical management,” J. Res. Med. Sciences, vol. 18, no. 7, p. 601, 2013.

[8] M. Chen, Y. Jiang, X. Lei, Y. Pan, C. Ji, and W. Jiang, “Drug target interactions prediction based on signed heterogeneous graph neural networks,” Chin. J. Electron., vol. 33, no. 1, pp. 231–244, Jan. 2024.

[9] F. Castiglione, C. Nardini, E. Onofri, M. Pedicini, and P. Tieri, “Explainable drug repurposing approach from biased random walks,” IEEE/ACM Trans. Comput. Biol. Bioinf., vol. 20, no. 2, pp. 1009–1019, Mar. 2023.

[10] S. Abbas, G. Avelino Sampedro, M. Abisado, A. S. Almadhor, T.-H. Kim, and M. Mohamed Zaidi, “A novel drug-drug indicator dataset and ensemble stacking model for detection and classification of drug-drug interaction indicators,” IEEE Access, vol. 11, pp. 101525–101536, 2023.

[11] B. G. Paltun, S. Kaski, and H. Mamitsuka, “DIVERSE: Bayesian data integrative learning for precise drug response prediction,” IEEE/ACM Trans. Comput. Biol. Bioinf., vol. 19, no. 4, pp. 2197–2207, Jul. 2022.

[12] H. He, G. Chen, and C. Yu-Chian Chen, “3DGT-DDI: 3D graph and text based neural network for drug–drug interaction prediction,” Briefings Bioinf., vol. 23, no. 3, May 2022, Art. no. bbac134.

[13] Y. Deng, X. Xu, Y. Qiu, J. Xia, W. Zhang, and S. Liu, “A multimodal deep learning framework for predicting drug–drug interaction events,” Bioinformatics, vol. 36, no. 15, pp. 4316–4322, Aug. 2020.

[14] M. Abdel-Basset, H. Hawash, M. Elhoseny, R. K. Chakrabortty, and M. Ryan, “DeepH-DTA: Deep learning for predicting drug-target interactions: A case study of COVID-19 drug repurposing,” IEEE Access, vol. 8, pp. 170433–170451, 2020.

[15] J. Liu, Z. Huang, F. Ren, and L. Hua, “Drug-drug interaction extraction based on transfer weight matrix and memory network,” IEEE Access, vol. 7, pp. 101260–101268, 2019.

[16] M. R. Karim, M. Cochez, J. B. Jares, M. Uddin, O. Beyan, and S. Decker, “Drug-drug interaction prediction based on knowledge graph embeddings and convolutional-LSTM network,” in Proc. 10th ACM Int. Conf. Bioinf., Comput. Biol. Health Informat., Sep. 2019, pp. 113–123.

[17] J. Liu, Z. Huang, F. Ren, and L. Hua, “Drug-drug interaction extraction based on transfer weight matrix and memory network,” IEEE Access, vol. 7, pp. 101260–101268, 2019.

[18] P. Bongini, F. Scarselli, M. Bianchini, G. M. Dimitri, N. Pancino, and P. Lió, “Modular multi-source prediction of drug side-effects with DruGNN,” IEEE/ACM Trans. Comput. Biol. Bioinf., vol. 20, no. 2, pp. 1211–1220, Mar. 2023.