Journal of Theoretical and Applied Information Technology
15th January 2024. Vol.102. No 1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

GRAPH CONVOLUTIONAL NEURAL NETWORK FOR IC50 PREDICTION MODEL WITH DRUG SMILES GRAPHS AND GENE EXPRESSIONS OF AMYOTROPHIC LATERAL SCLEROSIS

DEVIPRIYA S[1], VIJAYA M S[2]

[1] Research Scholar, Department of Computer Science, PSGR Krishnammal College for Women, Coimbatore, India

[2]Associate Professor, Department of Computer Science, PSGR Krishnammal College for Women, Coimbatore, India

E-mail: [1]devipriya041996@gmail.com, [2]msvijaya@psgrkcw.ac.in

ABSTRACT

IC50 prediction for neurodegenerative disorders like Amyotrophic Lateral Sclerosis is crucial in biomedical studies. Traditional machine learning models that use molecular descriptors and gene expression for building IC50 prediction models produce less accuracy and also most of the descriptors created by different tools are irrelevant and undefined. In this paper, a Graph Convolutional Neural Network, a deep learning algorithm, is employed for constructing a more precise IC50 prediction model. The model leverages the structural properties of drug molecules represented in graph format, and incorporates gene expression data as global features. So, the model is able to learn drug-gene interactions better. The drug-gene interactivity is learned by the model without drug-induced gene expressions as it is not found for most of the diseases. The work is implemented with well-known and most relevant 80 drugs related to ALS based on the pIC50 values of 32 protein targets of ALS disorder. The Canonical Smiles graph and their corresponding IC50 values of 80 drugs have been derived from the ChEMBL databases. Based on information from the Repurposing Hub in the Depmap database gene expression data for drug-related genes connected with ALS-related conditions is collected. The predictive results show that the proposed GCNN model with fine-tuned hyperparameters achieves MAE of 0.18, RMSE of 0.16 and R2 Score of 0.90.

Keywords: *IC50, Gene Expression, Graph Convolutional Neural Network, SMILES, Prediction*

1. INTRODUCTION

Lou Gehrig's disease, often known as Amyotrophic Lateral Sclerosis (ALS), is a chronic neurological disorder which damages the neurons that control voluntarily operated muscles [1]. It results in progressive paralysis, loss of mobility, and lack of muscle strength. The cells in the spinal cord and brain are largely impacted by ALS, which causes their degradation and mortality.

Drug discovery and drug design are crucial processes in the pharmaceutical industry. Drug discovery involves identifying potential drug candidates for treating diseases. Drug design focuses on rational design of drug molecules to interact with specific targets in the body. Challenges in drug discovery and design include the complexity of diseases, safety concerns, drug resistance, and the high cost and time involved [2]. The goal of drug development for ALS is to create treatments that can halt the progress of the illness, reduce symptoms, and enhance the standard of life for patients with the condition.

IC50 enables the identification of promising compounds for drug discovery, allowing for the optimization of therapeutic effects and the prioritization of drug candidates with higher potential. The prediction of IC50 values (half-maximal inhibitory concentration) is a crucial task in drug discovery and development. It helps to identify the potency of a drug compound in inhibiting a specific biological target, such as an enzyme or a receptor. Accurate IC50 prediction is vital in optimizing lead compounds and prioritizing potential drug candidates for further experimental validation [3].

A number of biological targets are crucial in the progression of Amyotrophic Lateral Sclerosis. These targets represent specific chemicals or biological pathways that, when disrupted, aggravate motor neuron degeneration. Researchers are focusing on numerous key targets for ALS inhibition, such as Superoxide Dismutase 1 (SOD1): A significant proportion of cases of familial ALS are caused by mutations in the SOD1 gene. Inhibiting the aberrant activity of the mutant SOD1 protein or target pathways interrupted by SOD1 failure could be one method for delaying the progression of the illness. TDP-43 (TAR DNA-binding Protein 43), C9orf72 Repeat Expansion, FUS (Fused in Sarcoma), Proteostasis, and Protein Quality Control, Excitotoxicity of Glutamate, Axonal Transport, Mitochondrial Dysfunction, and RNA Metabolism with Preprocessing are some of the additional disorders.

Currently, Machine learning and deep learning algorithms are extensively utilized in IC50 prediction due to their ability to analyze complex relationships between compound features and their corresponding inhibitory concentrations. Machine learning algorithms perform extremely well when dealing with small sized IC50 prediction datasets and avoid overfitting. Machine learning algorithms fail when capturing natural properties of drug molecules. But Deep Learning algorithms are employed when large IC50 datasets in different data forms are used in model building. It includes data like graphical data, sequential data and image data each with its own corresponding algorithms like GCNN, GRU and CNN [4].

Some of the below references provide a comprehensive literature review on the application of graph-based models, such as GCNNs, in drug discovery, molecular property prediction, and protein-ligand interactions. They establish the foundation for integrating drug SMILES graphs and gene expression features in our proposed GCNN model for IC50 prediction.

Kearnes et al. [5] explores the use of graph convolutional networks (GCNNs) for molecular property prediction. It highlights the limitations of traditional fingerprint-based approaches and demonstrates the advantages of GCNNs in capturing structural information and atom relationships in molecular graphs. GCNNs operate directly on the graph representation of molecules, enabling them to consider local and global features. They learn from raw molecular graphs without the need for handcrafted features, improving predictive accuracy of MSE with 0.46 ± 0.08. The research highlights the possibilities of GCNNs in discovering drugs through its property prediction tasks.

The MoleculeNet standard is a framework for comparing several artificial intelligence models used in drug development, according to Wu et al. [6]. The paper makes reference to a number of datasets, including QM7, QM8, QM9, ESOL, FreeSolv, Lipophilicity, and others. It concentrates on graphical models, such GCNNs, and evaluates how well they perform across a range of tasks. The paper addresses the possibility of graphical models in accomplishing this and emphasizes the significance of precisely predicting molecular features, especially IC50 values. It offers a thorough overview of the difficulties and possibilities in using machine learning with molecular data. With these kinds of data, several metrics are applied, and GCNN achieves greater precision as compared to other machine learning models.

A thorough summary of deep learning models used in computational chemical science, notably in the field of drug development, is given by Goh et al. [7]. The use of graph convolutional networks (GCNNs) as a potent method for generating graph representations of drug graphs is highlighted in the article. The use of GCNNs is noted for its capacity to accurately predict numerous drug properties, including IC50 values, by capturing complex structural details and atom interactions. The authors offer suggestions for possible future advances as well as a discussion of the benefits and difficulties of using deep learning algorithms in drug design.

A paper on protein-ligand reactions by Ragoza et al. [8] shows the value of convolutional neural networks (CNNs) in determining binding capacity of medicinal compounds. It explained the manner in which CNNs are good at detecting regional trends and spatial correlations in complexes made from protein-ligand . When compared with the autodock vina framework, the CNN model achieves an AUC greater than 0.8. When compared with the autodock vina framework, the CNN model achieves an AUC greater than 0.8. The principles and insights presented in this paper can be relevant for understanding the application of graph-based models in drug discovery.

Gilmer et al. [9] introduced the concept of neural message passing, a framework for combining graph neural networks with quantum chemistry data. It demonstrates the efficacy of this approach in predicting molecular properties, including IC50

values, by leveraging the structural information encoded in molecular graphs. The study highlights the potential of graph-based models for accurate and interpretable predictions in drug discovery. The QM9 dataset was employed, and it was discovered that simultaneously training on all 13 targets continuously outpaced training a single model for each target. In certain instances, the improvement reached 40%.

In order to predict drug response, Liu et al. [10] developed a unique technique that incorporates multi-omics data of cancer cells and their drug structures. The DeepCDR composite graph convolutional network that the authors propose combines a graph convolutional network with several subnetworks. The hidden depiction of geometrical arrangements between the bonds and atoms in drugs is automatically learned by DeepCDR, in contrast to past studies that depended on manually created drug characteristics. Using the CCLE and GDSC datasets, a spearman correlation of 0.903 0.004 and an RMSE of 1.058 0.06 were achieved.

In the existing research irrelevant molecular descriptors are used in building machine learning based IC50 prediction models. GCNN were used to predict IC50 with only structural properties without using gene expression. The gene expression used in few works with CNN algorithms performs classification tasks and does not make accurate predictions. Most of the hybrid algorithms that use CNN fail to capture its properties completely. In this paper, a GCNN [11] based approach for developing a more accurate IC50 prediction model using chemical and gene expression features is proposed. Protein targets of ALS related conditions are chosen and their associated drug attributes, gene expression data are retrieved from databases. They are converted into suitable formats for training and building IC50 prediction models.

2. A GRAPH DATA FOR ALS DRUGS

The graphical representation of ALS drugs is modeled through two tasks. The Pharmacogenomics collection [12] is the first task and featurization is the second. During Pharmacogenomics collection, the Canonical SMILES, gene expression and IC50 values are retrieved from databases with a series of processes such as pathway analysis of protein targets, drug attributes retrieval, gene expression retrieval. Featurization is the process of converting the raw data obtained during the Pharmacogenomics collection phase into a suitable format that can be used for model building and analysis. The drug attributes are used in the featurization process while gene expression is normalized and then featured.

2.1 Pharmacogenomics Collection

The ALS targets are searched in the UniProt [13] database and retrieved. Through pathway analysis it is found that all the thirty-two protein targets are observed in ALS disease. The protein targets are mapped with ChEMBL [14] databases for finding the drug attributes associated with 32 ALS targets. From the identified drugs, most relevant and approved drugs are selected. Around 80 drugs are considered in this study. The process of the Pharmacogenomics collection is depicted in Figure 1 and explained below.

Figure 1: Pharmacogenomics Collection Process.

Protein target pathway analysis identifies drugs and their targets related to ALS. A list of 32 ALS-associated protein targets is obtained from UniProt based on their relevance to ALS pathology. Mapping these targets with ChEMBL enables to identify associated drugs and similar proteins. The pathway analysis for few drug targets in Amyotrophic Lateral Sclerosis disease is given in Table 1.

Journal of Theoretical and Applied Information Technology
15th January 2024. Vol.102. No 1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1: Pathway Analysis of few ALS drug Targets.

Drug Target	Drug Name	Pathway Analysis		
		Term	Adjusted P-value (lower)	Effect size (larger)
KCNK10, SLC7A11, SCN9A	Riluzole [15]	Ferroptosis	0.016498	0.25
SLC6A4, ANO1, HTR2B	Fluoxetine [16]	Serotonergic synapse	0.000567	0.66667
KCNH1, HTR7, HTR1D, DRD1, GRIN2B	Haloperidol [17]	Neuroactive ligand-receptor interaction	8.70E-06	0.8
TRPC5, ADRA1A, SMPD1, DRD1, CALM1	Chlorpromazine [18]	Calcium signaling pathway	0.000411	0.6
S1PR1; S1PR5	Fingolimod [19]	Neuroactive ligand-receptor interaction	0.000427	1.0

The drug attributes are retrieved by collecting drug SMILES and pIC50 values from ChEMBL database. Drug SMILES provide a concise and standardized way to represent the structure of a drug molecule using a specific set of characters. pIC50 is a measure of the potency or inhibitory activity of a drug compound. It represents the negative logarithm of base 10 of the concentration of a drug required to inhibit a target by 50%. The pIC50 value is commonly used in pharmacology and drug discovery to quantify and compare the potency of different compounds. A higher pIC50 value indicates a higher potency, as it corresponds to a negative logarithm for inhibitory activity.

Gene expression data associated with the 80 drugs of ALS are collected from the DepMap [20] database. DepMap is a comprehensive resource that provides genomic and transcriptomic data, including gene expression profiles, for a wide range of disease-related cell lines. The query is submitted to the DepMap database, specifying ALS-related cell lines in PRISM Drug repurposing hub. Therefore, the gene expression data contains information about the expression levels of specific genes in ALS-related cell lines treated with the selected drug SMILES. It signifies how cancer gene expressions are related to neuroinflammation in ALS and applied on drug repurposing from cancer to ALS [21]. The drugs are repurposed based on the PRISM Drug repurposing hub in Depmap. A sample gene expression data corresponding to Dalfampridine drug in heatmap format is given in

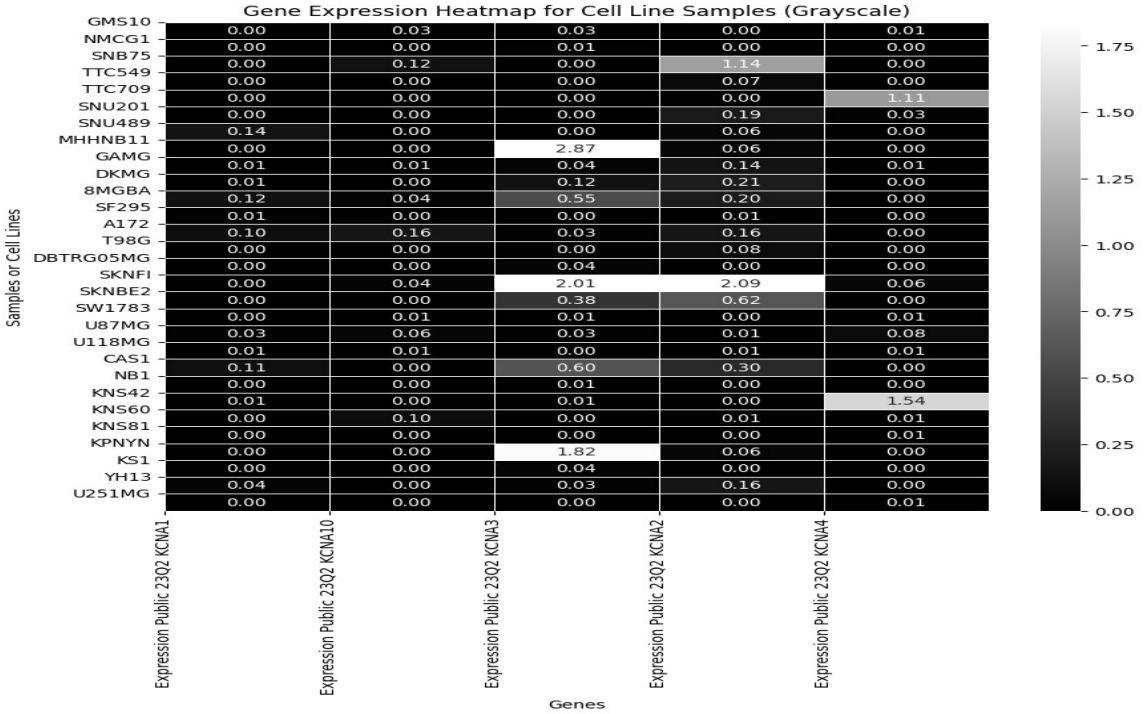


Figure 2: Sample Gene Expression Data of Dalfampridine Drug

Figure 2.

A complete approach of combining protein target pathway analysis and ChEMBL mapping is used to find candidate drugs and their targets. The gene expression is finally collected for drug targets. The three procedures above assist in gathering data - drug SMILES, pIC50 values and gene expression of drug targets on 80 drugs. The description of a few drug samples are shown in Table 2.

Table 2: Description of Drug Samples

.

Drug Target	No. of cell-lines	Drug Name	Drug SMILES	PIC50
KCNK10, SLC7A11, SCN9A	25	Riluzole	Nc1nc2ccc(OC(F)(F)F)cc2s1	6
SLC6A4, ANO1, HTR2B	28	Fluoxetine	CNCCC(Oc1ccc(C(F)(F)F)cc1)c1ccccc1	6.301899
KCNH1, HTR7, HTR1D, DRD1, GRIN2B	40	Haloperidol	O=C(CCCN1CCC(O)(c2ccc(Cl)cc2)CC1)c1ccc(F)cc1	8.248721
TRPC5, ADRA1A, SMPD1, DRD1, CALM1	34	Chlorpromazine	CN(C)CCCN1c2ccccc2Sc2ccc(Cl)cc21	6.345823
S1PR1; S1PR5	46	Fingolimod	CCCCCCCCc1ccc(CCC(N)(CO)CO)cc1	7

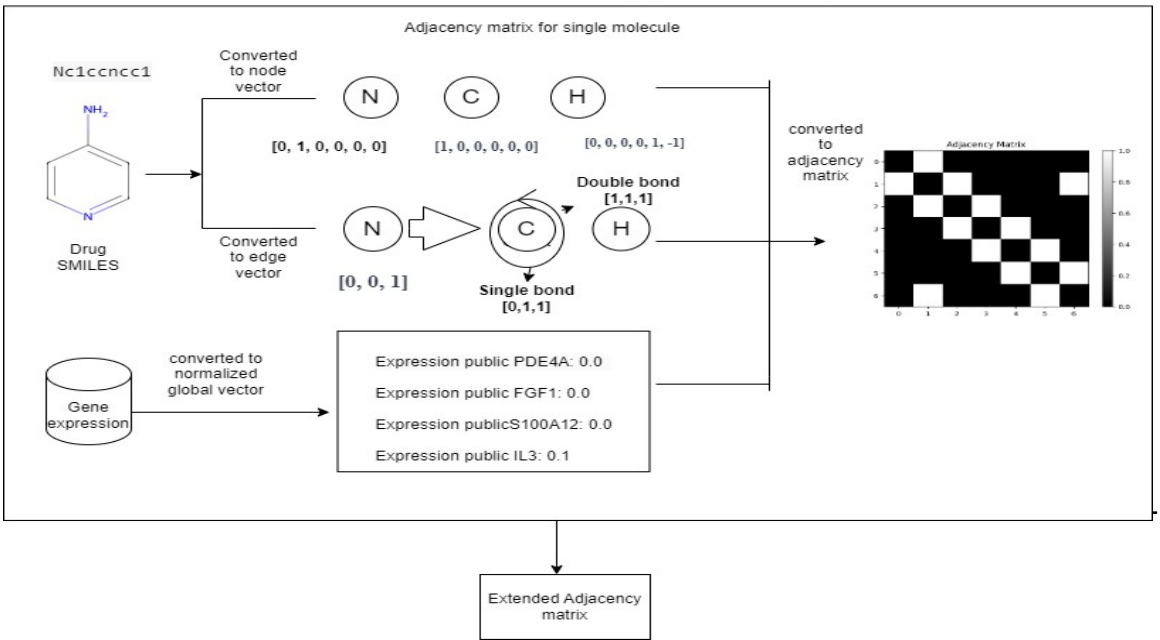
As a result, the Pharmacogenomics collection phase yields the gene expression and drug SMILES relevant to 80 drugs of ALS disease, which will then be transformed into an adjacency matrix. The adjacency matrix is created through featurization, which is described below.

2.2 Featurization

Drug SMILES are transformed into feature vectors during the featurization process and the gene expressions are normalized. The Drug SMILES and gene expression together are converted into feature vectors. The process is looped for all drug SMILES and their corresponding gene expression to create an extended adjacency matrix [22]. This adjacency matrix is finally passed to GCNN for training. Figure 3 depicts adjacency matrix creation for a single molecule.

A molecular graph for each drug is constructed where atoms are nodes, and bonds are edges. Each node is associated with a set of features, which can include atom type, hybridization state, chirality and formal charge. Each edge has features such as bond type and bond order. A node feature vector is created for each atom in the molecular graph with encoded characteristics of atom such as atom type, hybridization state, chirality and formal charge. Bond attributes including bond type, bond order is used to create edge feature vectors.

For example, consider the Dalfampridine molecule. The feature vector construction [23] for this molecule having three atoms Nitrogen (N), Carbon (C), and Hydrogen (H) is given in Table 3 and Table 4. In the node feature vectors, each element represents a different atom type in the order of [C, N, O, S, H, Other]. In the edge feature vector, the first element is bond order, second element is source atom type and third element is target atom type. So, the vector is composed of [Bond order,



source atom type, target atom type].

Table 3: Node Feature Vector Description

Node	Hybridization	chirality	Formal charge	vector
N	sp2	None	0	[0,1,0,0,0,0]
C	sp2	None	0	[1,0,0,0,0,0]
H	None	None	0	[0,0,0,0,1,0]

Table 4: Edge Feature Vector Description

Edge	Bond Type	Bond Order	vector
NC	single	0	[0,0,1]
CC	double	1	[1,1,1]
CC	single	0	[0,1,1]

Gene expression data corresponding to each drug is normalized using Min-Max normalization and added to the feature vector corresponding to the drug SMILES graph as a global feature. These global features are updated concurrently with the node and edge features in each graph convolutional layer and are added as an additional feature dimension to any node in the graph. Gene expression data is utilized as a general feature to describe the impact of a certain gene on the drug's activity. Each drug SMILES is provided with its corresponding gene expression data as global feature vector.

Finally, the node feature, edge feature and global feature vectors are transformed into a single adjacency matrix. Each row and column of the matrix represents an atom in the molecule, and the entries in the matrix indicate whether there is a bond between the corresponding atoms. The adjacency matrix is generated for all molecules and they are combined and padded with the largest atom number size. This extended adjacency matrix represents graphical data for all 80 drugs and it is used by GCNN architecture to build an IC50 prediction model.

3. IC50 PREDICTION MODEL BUILDING

The GCNN layers perform graph convolutions to propagate information through the drug graph, capturing local structural patterns. The global gene expression features are added to each drug separately as each drug has a unique gene expression. The combined features are then fed into fully connected layers for building the IC50 prediction model. This model integrates both structural and gene expression information to enhance the predictive performance for IC50 values. The GCNN model for IC50 prediction takes drug SMILES as input in graph format along with gene expression as global features. The drug molecules are represented as graphs, capturing the topological information of atoms and bonds. IC50 Prediction model building process is shown in Figure 4.

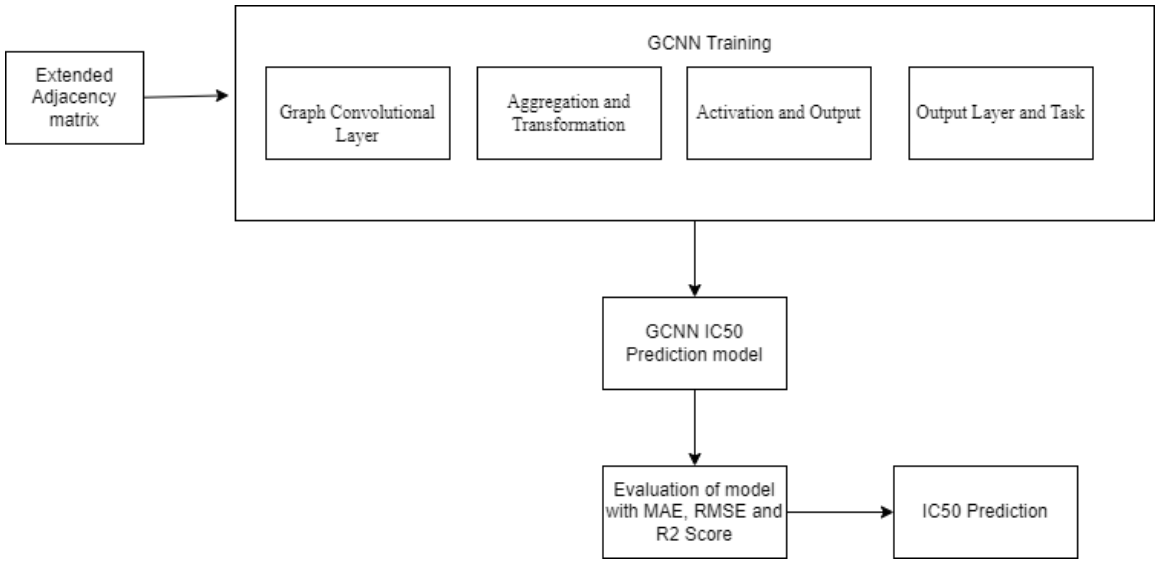


Figure 4: IC50 Prediction Model Building Process

An extended adjacency matrix is typically used to define the graph's structure. Graph convolution, which involves gathering data from nearby nodes for each node in the graph, is the fundamental operation of a GCNN. By taking into account the features of its neighbors, a graph convolutional layer creates new feature vectors for each node. The adjacency matrix of the graph is used to identify adjacent nodes. Using methods like weighted summation or attention processes, the properties of nearby nodes are combined. The new feature vectors for each node are then computed from the aggregated data using learnable parameters such as weights and biases. In order to introduce non-linearity, the updated feature vectors of the nodes are typically passed via activation functions like ReLU. The updated feature matrix corresponding to the new feature vectors for the nodes is the GCNN layer's final output. This updated feature matrix is used for IC50 prediction by GCNN.

Thus, the IC50 prediction model is built with hyperparameters tuning and k-fold cross validation and trained with an extended adjacency matrix. The hyperparameters like filter size, learning rate and optimizer plays an important role in model building. In convolutional neural networks, the filter size refers to the dimensions of the filter used to slide over the input data. It determines the receptive field, impacting what features the network can learn from the input. Learning rate is a hyperparameter that controls the step size during the optimization process of training a machine learning model. It influences the rate at which the model's parameters are updated to minimize the loss function. An optimizer is used to update the model's parameters during training to minimize the loss function. K-fold cross-validation is used to assess the performance and generalization ability of a GCNN based IC50 prediction model. The typical choice for k is 7 here. The model is evaluated for its prediction efficiency using metrics MAE, RMSE, R2 Score.

4. EXPERIMENTS AND RESULTS

DeepChem [24, 25] serves as a front-end framework specialized for cheminformatics and drug discovery tasks, while TensorFlow is the backend framework responsible for the actual computation and execution of the models. The IC50 prediction model has been built by implementing GCNN using Python and training the extended adjacency matrix of Drug SMILES and gene expression. The experiment is carried out for various epoch sizes and by setting other hyperparameters as shown below in Table 5. The output layer is designed with one unit and hidden layers and GCNN layers are defined with 32 units. Adam optimizer is used here to reduce the error and increase efficiency.

Table 5: Hyperparameters Setting for GCNN Model

GCNN MODEL	
Epoch	500
GCNN layer	1
Filter size	32
Learning rate	0.001
Output size	1
Optimizer	Adam
k-fold	7

The iteration of training the GCNN network starts from epoch 10 and converges with epoch 500 at fold 7. The outcomes of the GCNN prediction with regard to MAE, RMSE, and R2 Score over several epochs are observed and shown below. The highest accuracy produced by GCNN is R2 Score of 0.90 with loss percentages of 0.18 by MAE and 0.16 by RMSE. The error functions MAE, RMSE, and R2 Score are calculated for each epoch in the intervals of 100. The accuracy and error rate gradually shows improvement and finally produces maximum accuracy with respect to R2 score and minimum error values with respect to MAE and RMSE. The maximum value is recorded as 0.51 and 0.403 respectively by MAE and RMSE. Similarly, the maximum accuracy produced by R2 Score is 0.90 which signifies 90 percent of accuracy and the minimum accuracy is 0.43 at epoch 100. The Performance Results of GCNN based IC50 Prediction Model using Drug SMILES and gene expression is shown in Table 6 and depicted in Figure 5.

Table 6: Performance Results of GCNN based IC50 Prediction Model for Various Epochs

Epoch	MAE	RMSE	R2 Score
100	0.51	0.403	0.43
200	0.38	0.32	0.53
300	0.36	0.30	0.58
400	0.35	0.25	0.65
500	0.18	0.16	0.90

Journal of Theoretical and Applied Information Technology
15th January 2024. Vol.102. No 1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

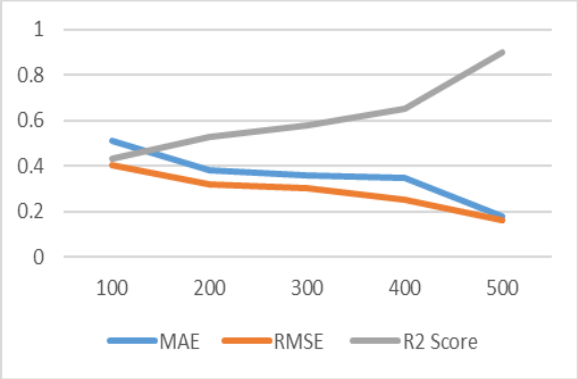


Figure 5: Performance Results of GCNN based IC50 Prediction Model using Drug SMILES and Gene Expression

In our previous work, a GCNN model was built with drug SMILES alone. The iteration of training the GCNN network started from epoch 10 and converged with epoch 500. The outcomes of the GCNN prediction with regard to MAE, RMSE, and R2score over several epochs were observed and reproduced here in Table 7 and Figure 6. The highest accuracy produced by GCNN was 0.73 by R2 Score, with loss percentages of 0.31 by MAE and 0.21 by RMSE. The error functions MAE, RMSE, and R2 score were calculated for each epoch in the intervals of 100. The accuracy and error rate gradually showed improvement and finally produced maximum accuracy with respect to R2 score and minimum error values with respect to MAE and RMSE. The minimum MAE and RMSE was recorded as 0.25 and 0.211 while the maximum value was recorded as 0.41 and 0.303 respectively. Similarly, the maximum accuracy produced by R2 score was 0.73 which signifies 73 percent of accuracy and the minimum accuracy was 0.42 at epoch 100.

Table 7: Performance Results of GCNN based IC50 Prediction Model Built with Drug SMILES

Epoch	MAE	RMSE	R2 Score
100	0.41	0.303	0.42
200	0.36	0.31	0.52
300	0.35	0.29	0.54
400	0.32	0.24	0.63
500	0.31	0.21	0.73

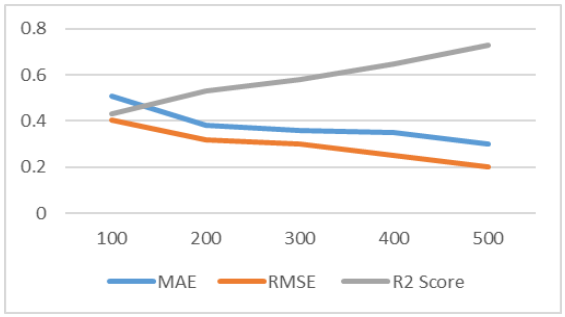


Figure 6: Performance Results of GCNN based IC50 Prediction Model using drug SMILES

The performance of the GCNN based IC50 prediction model built with Drug SMILES and gene expression is compared with the performance of the GCNN IC50 prediction model trained only with Drug SMILES. The highest accuracy is achieved at Epoch 500. GCNN using Drug SMILES and gene expression obtains an R2 score of 0.90, RMSE of 0.16 and MAE of 0.18. GCNN using drug SMILES obtains R2 score of 0.73, RMSE of 0.2 and MAE of 0.3. Thus, from the results GCNN using Drug SMILES and gene expression achieves more accuracy than GCNN using drug SMILES alone. The comparative results are also given in Table 8 and Figure 7.

Table 8: Comparative Results of IC50 Prediction Models based on Two Datasets

Dataset	MAE	RMSE	R2 Score
Drug SMILES and Gene Expressions	0.18	0.16	0.90
Drug SMILES	0.3	0.2	0.73

Figure 7: Comparative Results of IC50 Prediction Models Based on Two Datasets

4.1 Method Contrast

The adopted methodology distinguishes the proposed work from existing literature. In the present literature, only gene expressions or drug SMILES are used to determine IC50. However, in our technique, both were taken into account and fed into the model to precisely learn drug-gene interactions. CNN [26] has been used to estimate binding affinity in a few studies using protein structures and drug pictures. While most studies focus on binding affinity, our methodology focuses on gene expression and IC50 prediction. The approach utilized in our study was principally compared to DRUGGCN [27], which uses drug-induced gene expression to determine IC50. As assessment metrics in DRUGGCN RMSE, Pearson Correlation Coefficient (PCC) and Spearman's Rank Correlation Coefficient (SCC) are utilized, whereas in the proposed methodology in this paper better metrics are considered.

Several other publications [28] have used data obtained from the Genomics of Drug Sensitivity in Cancer (GDSC) dataset to create complex networks. Given a novel cell line, these models are used for predicting the cell line's reaction to the evaluated GDSC medicines. In a similar way provided a novel medication, these models are used to assess the probable response of tested GDSC cell lines to this drug. The model classifies the cell lines as resistant or sensitive to drugs incorporated which is a classification problem whereas other method uses pathway analysis to find sensitive genes and cell lines of drugs therefore giving high results with different metrics.

The majority of the research focuses on cancer pathology [29, 30] and does not take into consideration mutations in diseases such as Amyotrophic Lateral Sclerosis. However, previous research has addressed all data types, involving copy number variation and mutation data. Various models, such as heterogeneous models and transformer models, have been developed and employed in the available literature. However, all of these models contribute to cancer disease and have not been tested on neurological disorders or other pathologies. Our findings demonstrate that IC50 may be predicted using mutations in any disorder like ALS.

The main disadvantage of adopting drug-induced gene expressions is that disease gene expressions must also be considered if they are included. The main assumption is that drug-induced gene expression should reverse illness expression, which is not accomplished in many models [31, 32].

Finally, the proposed GCNN model using Drug SMILES and gene expression demonstrates improved IC50 prediction performance. By integrating the structural information encoded in drug SMILES graphs and the contextual influence of gene expression patterns the model captures the complex relationships between drugs and gene expression profiles, resulting in enhanced predictive accuracy. The quantification of certain structural and gene expression characteristics that substantially contribute to IC50 is made possible with feature selection. The model's ability to make accurate predictions also implies that it has learned meaningful representations of drug-gene interactions. The integration of gene expression data as global features enriches the predictive capabilities of the model. The GCNN recognition rate supports the model's ability to forecast IC50. When compared with other Machine learning or deep learning algorithms, the GCNN based IC50 prediction model captures the natural depiction of drug molecules effectively.

5. CONCLUSION

A novel GCNN architecture for building an IC50 prediction model by integrating drug SMILES in graph format with gene expression as global features has been proposed in this paper. The model leverages the power of graph convolutional networks (GCNNs) to capture the structural information and relationships within drug molecules represented as graphs. The UniProt, ChEMBL and DepMap databases have been used in this study. The drug SMILES and gene expression data, which have been collected for 80 drugs are featurized and used as an adjacency matrix for training the GCNN. The GCNN has been implemented with the Deepchem framework and the experiments were carried out with proper setting of hyperparameters. The IC50 prediction model has been tested for its efficiency with standard metrics and produced promising results in predicting IC50 values. In future the same study can be done with EC50, or half-maximal effective concentration. EC50 [33] is used to explore drug induction effect while IC50 capture inhibitory effect of a drug. This holistic approach enhances understanding of the factors influencing IC50 values and enables more accurate predictions, thus facilitating the process of drug discovery and development.

REFERENCES

[1] B. Marin et al, "Variation in worldwide incidence of amyotrophic lateral sclerosis: a meta-analysis", Int. J. Epidemiol, May 2016, p. dyw061.

[2] S.-F. Zhou and W.-Z. Zhong, "Drug Design and Discovery: Principles and Applications", Molecules, vol. 22, no. 2, Feb. 2017, p. 279.

[3] S. Aykul and E. Martinez-Hackert, "Determination of half-maximal inhibitory concentration using biosensor-based protein interaction analysis", Analytical Biochemistry, Vol. 508, Sep. 2016, pp. 97–103.

[4] S. Kolluri, J. Lin, R. Liu, Y. Zhang and W. Zhang, "Machine Learning and Artificial Intelligence in Pharmaceutical Research and Development: a Review", The AAPS Journal, Vol. 24, No. 1, Jan. 2022, p. 19.

[5] S. Kearnes, K. McCloskey, M. Berndl, V. Pande and P. Riley, "Molecular graph convolutions: moving beyond fingerprints", Journal of Computer-Aided Molecular Design, Vol. 30, No. 8, Aug. 2016, pp. 595–608.

[6] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing and V. Pande, "MoleculeNet: a benchmark for molecular machine learning", Chemical Science, Vol. 9, No. 2, 2018, pp. 513–530.

[7] G. B. Goh, N. O. Hodas and A. Vishnu, "Deep learning for computational chemistry", Journal of Computational Chemistry, Vol. 38, No. 16, Jun. 2017, pp. 1291–1307

[8] M. Ragoza, J. Hochuli, E. Idrobo, J. Sunseri and D. R. Koes, "Protein–Ligand Scoring with Convolutional Neural Networks", Journal of Chemical Information and Modeling, Vol. 57, No. 4, Apr. 2017, pp. 942–957.

[9] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals and G. E. Dahl, "Neural message passing for quantum chemistry", Proceedings of the 34th International Conference on Machine Learning, PMLR, Jul. 2017, pp. 1263–1272.

[10] Q. Liu, Z. Hu, R. Jiang and M. Zhou, "DeepCDR: a hybrid graph convolutional network for predicting cancer drug response", Bioinformatics, Vol. 36, No. Supplement_2, 2020, pp. I911–i918.

[11] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks", arXiv preprint, arXiv:1609.02907, 2016 Sep 9.

[12] G. Ermak, Emerging medical technologies. New Jersey: World Scientific, 2016.

[13] https://www.uniprot.org/

[14] https://www.ebi.ac.uk/chembl/

[15] J.-L. Wang *et al.*, "Effect of the neuroprotective agent riluzole on intracellular Ca 2+ levels in IMR32 neuroblastoma cells ", *Archives of Toxicology*, vol. 75, no. 4, 2001, pp. 214–220.

[16] J. Wu and G. Qin, "The efficacy and safety of fluoxetine versus placebo for stroke recovery: a meta-analysis of randomized controlled trials", *International Journal of Clinical Pharmacy*, vol. 45, no. 4, 2023, pp. 839–846.

[17] Granger B and Albu S. The haloperidol story. Annals of Clinical Psychiatry. 2005 Jan 1;17(3):137-40.

[18] R. Mlambo, J. Liu, Q. Wang, S. Tan, and C. Chen, "Receptors Involved in Mental Disorders and the Use of Clozapine, Chlorpromazine, Olanzapine, and Aripiprazole to Treat Mental Disorders ", *Pharmaceuticals*, vol. 16, no. 4, 2023, p. 603.

[19] A. Hatem *et al.*, "Current and future trends in multiple sclerosis management: Near East perspective ", *Multiple Sclerosis and Related Disorders*, vol. 76, 2023, p. 104800.

[20] https://depmap.org/

[21] D. M. Freedman, R. E. Curtis, S. E. Daugherty, J. J. Goedert, R. W. Kuncl, and M. A. Tucker, "The association between cancer and amyotrophic lateral sclerosis", Cancer Causes & Control, Vol. 24, No. 1, 2013, pp. 55–60.

[22] P. Paramadevan and S. Sotheeswaran, "Properties of adjacency matrix of a graph and it's construction", Journal of Science, Vol. 12, No. 1, 2021, p. 13.

[23] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, and P. Riley, "Molecular graph convolutions: moving beyond fingerprints", J Comput Aided Mol Des, vol. 30, no. 8, pp. 595–608, Aug. 2016,

[24] https://deepchem.io/

[25] B. Ramsundar, P. Eastman, P. Walters, V. Pande, K. Leswing, and Z. Wu, "Deep Learning for the Life Sciences", O'Reilly Media, 2019.

[26] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology", Insights Imaging, vol. 9, no. 4, pp. 611–629, Aug. 2018.

[27] S. Kim, S. Bae, Y. Piao, and K. Jo, "Graph Convolutional Network for Drug Response Prediction Using Gene Expression

Data", Mathematics, vol. 9, no. 7, p. 772, Apr. 2021.

[28] Z. Stanfield, M. Coşkun, and M. Koyutürk, "Drug Response Prediction as a Link Prediction Problem", Sci Rep, vol. 7, no. 1, p. 40321, Jan. 2017.

[29] J. Shin, Y. Piao, D. Bang, S. Kim, and K. Jo, "DRPreter: Interpretable Anticancer Drug Response Prediction Using Knowledge-Guided Graph Neural Networks and Transformer", IJMS, vol. 23, no. 22, p. 13919, Nov. 2022.

[30] H. Zhang, Y. Chen, and F. Li, "Predicting Anticancer Drug Response With Deep Learning Constrained by Signaling Pathways", Front. Bioinform., vol. 1, p. 639349, Apr. 2021.

[31] M. Tan, O. F. Özgül, B. Bardak, I. Ekşioğlu, and S. Sabuncuoğlu, "Drug response prediction by ensemble learning and drug-induced gene expression signatures", Genomics, vol. 111, no. 5, pp. 1078–1088, Sep. 2019.

[32] S. Chawla et al., "Gene expression based inference of cancer drug sensitivity", Nat Commun, vol. 13, no. 1, p. 5680, Sep. 2022.

[33] X. Jiang and A. Kopp-Schneider, "Summarizing EC50 estimates from multiple dose-response experiments: A comparison of a meta-analysis strategy to a mixed-effects model approach", Biometrical J, vol. 56, no. 3, pp. 493–512, May 2014.