Drug Sensitivity Prediction in Cancer Using an Artificial Neural Network: Insights from Genomic Data of the GDSC Dataset

Assistant Professor: Mrs. Suma K1, Students: K Sahana Rao2, Karthik U Shettigar3, Dhanush A4, Dhanush S Shetty 5, Dept. of CSE, Mangalore Institute of Technology and Engineering, Moodbidri, India

***Abstract—*** The big revolution behind personalized medicine in cancer therapies involves therapies that are indeed tailored to individual genetic profiles influential in patient responses to certain drug types. "Drug Response Prediction Using an ANN Model" is a project implemented via PyTorch to develop a model of an Artificial Neural Network trained on predicting drug sensitivity through the Genomics of Drug Sensitivity in Cancer (GDSC) dataset. It predicts the IC50 values, which are a critical measure for drug efficacy, by integrating data from three different files that contain the necessary genomic features such as gene mutation, gene expression profiles, and CNAs. The model's performance is further improved through a rigorous data preprocessing pipeline that includes normalization, feature selection, and hyperparameter tuning, whose performance will be measured using metrics such as Mean Squared Error and R-squared. A web interface with user interaction is implemented by means of a Flask interface; the user can input through dropdown menus and obtain real-time predictions of IC50, with no need for file upload. The interface also includes visualizing tools to support interpretability by enabling the user to compare predicted and actual values of IC50. This model is representative of the power of machine learning in advancing personalized cancer treatment, while future directions are targeted at dataset enlargement, integration of multi-omics data, and clinical validation for a wider applicability in real-world cancer therapy.

***Keywords—*** drug response prediction, IC50, artificial neural networks, PyTorch, personalized medicine, GDSC dataset.

# INTRODUCTION

1. *Background*

Personalized medicine has become a cornerstone in oncology, aiming at the adaptation of treatments to the genetic profile of each individual patient. Genetic heterogeneity within the tumor confers a spectrum of drug responses, reinforcing the need for predictive models supporting precision oncology. IC50 is the half-maximal inhibitory concentration, an index describing the efficiency of a drug in inhibiting cancer cell viability; it has become a critical parameter for treatment planning in oncology and allows clinicians to select therapies that are most effective against specific types of cancers.

Artificial Neural Networks (ANNs) are well-suited for predictive modelling in oncology due to their capacity to capture complex, nonlinear relationships in high-dimensional genomic data. The following project is done using the Genomics of Drug Sensitivity in Cancer dataset to predict IC50 values for multiple cancer cell lines by considering a few important genomic features like gene mutation, tissue descriptors, and CNAs. It will go by the name "Drug Response Prediction Using Artificial Neural Networks (ANN)" and contribute toward enabling the oncologist and/or researchers with informed decisions to help them facilitate more personalized and hence effective treatment modalities.

1. *Problem Statement*

Cancer treatment heavily relies on the genetic heterogeneity of the cells, which eventually determines varied responses to the same drug. This may imply a one-size-fits-all kind of approach to the current methodologies in cancer treatment, whereby traditional therapeutic approaches often lack this fine-grained discrimination. Moreover, currently used predictive models fall well short in terms of their clinical potential for personalized oncology. The work will try to develop a more accurate ANN-based predictive model by including the essential genomic profile, capturing the intricate relationships between these genomic features and drug responses that improve accuracy in the selection of treatments. Increased predictive precision is important to reduce useless treatments, decrease costs, and benefit patients through improved outcomes in personalized cancer care.

1. *Objectives*

The main aim of the present project is to come up with an ANN model that will be able to predict IC50 values based on genomic data included in the GDSC dataset, in turn helping in personalized cancer therapy. This model includes a variety of key genomic features like gene mutation, variation in gene expression, and CNA for accurate prediction. Normalization, selection of features, and tuning of hyperparameters are some of the data preprocessing techniques applied for the optimization of results arising from this model. Besides, a web-based interface built on Flask provides a user interface for entering genomic data and obtaining real-time predictions of IC50. It also provides visualizations to support clinical decision-making in terms of the efficacy of different drugs.

1. *Scope of Work*

The scope of this project involves combining three different genomic data sources within the GDSC dataset to find correlations with drug response. The ANN model, implemented in PyTorch, is designed to capture complex relationships between genomic features and drug efficacy. Model performance will be evaluated using Mean Squared Error (MSE) and R-squared (R²) to ensure predictive reliability. A web-based interface using Flask allows for input of data and display of predicted results, culminating into a scalable real-time system for the needs of clinical and research applications toward precision oncology.

# LITERATURE REVIEW

1. *Drug Response Prediction Using Artificial Neural Networks*

Artificial Neural Networks (ANNs) have become foundational in drug response prediction due to their ability to model complex, nonlinear relationships within high- dimensional biological and clinical data. The demand for robust predictive models has grown alongside advancements in personalized medicine, which aims to understand variations in drug efficacy across diverse patient profiles. Costello et al. (2018) introduced an ANN model capable of processing high-dimensional gene expression data, achieving notable prediction accuracy in cancer drug response across multiple cell lines [1]. Similarly, Zhang et al. (2020) combined gene expression data with drug chemical structures in a deep learning model, resulting in highly accurate, patient- specific predictions of drug responses [2].

1. *Feature Selection for Improved Model Interpretability*

Main reasons being, the feature selection works most under the ANN-based models of responses to drugs by preventing both overfitting and rendering the model interpretable to a certain extent. Results After developing a hybrid model fusing techniques of feature selection algorithms followed by ANNs, using which critical genomic features from Drug sensitive and resistant cells that controlled drug response can be unveiled, Menden et. al., 2019 designed a hybrid model that keeps computations low and helps practical approaches in oncology [3]. Moreover, Tsigelny et al. (2021) performed transfer learning in an ANN framework and demonstrated that transferring features learned across related tasks can improve model robustness and predictive accuracy, especially when data is limited [4].

1. *Integrating Multi-Omics Data for Enhanced Generalization*

Despite these recent advances, model interpretability and generalization across diverse biological contexts elude most ANN applications to this day. The integrative use of multi-omics data, as developed by Lee et al. (2022), holds very much promise for overcoming this. Their ANN model integrates multiple layers of data, such as transcriptomics and proteomics, and shows increased generalizability and predictive power, hence underlining an integrative data approach in enhancing model accuracy for eventual clinical applications [5].

1. *Summary of Insights and Project Relevance*

There is substantial evidence from the literature that the ANN model, in light of feature selection and integration of multi-omics data, guarantees high predictive accuracy and generalizability in modeling drug response. Based on the strength of these two observations, the project intends to develop an ANN model trained on a dataset drawn from the work known as Genomics of Drug Sensitivity in Cancer, focusing on a number of genomic features in the form of gene mutation, expression levels, or CNAs with the aim of predicting the IC50 value. It fills key gaps in personalized drug response prediction through feature selection refinement, using a carefully curated dataset, hence bridging the gap between model accuracy and clinical applicability.

# METHODOLOGY

1. *Data Collection and Preprocessing*

This approach has made use of the data from the Genomics of Drug Sensitivity in Cancer dataset with comprehensive drug response across various cancerous cell lines. The data here has been combined from three different sets, which contributed to each other for obtaining the key genomic features related to gene mutation, variation in the gene expression level, and CNA. Therefore, this multi-dimensional dataset will serve as a sound basis for the prediction of drug sensitivity.

The data preprocessing steps involved:

* 1. Normalization: Normalization is a process carried out to make the scale of continuous features consistent, hence reducing the impact of any particular feature disproportionately on model predictions.
  2. Feature Selection: The selection of genomic markers was done to include only those that were highly relevant to drug response, hence improving the accuracy of prediction and reducing irrelevant noise.
  3. Data Splitting: Data is split such that 80% is training data and 20% test data, for evaluation of performance on unseen data unbiasedly.

1. *Model Architecture*

The ANN model was constructed in PyTorch to predict the values of IC50. It consists of an architecture of multiple hidden layers to learn complex patterns and relationships within the genomic data. Important features of the model architecture are as follows:

* Input Layer: Processes the selected genomic features.
* Hidden Layers: Several hidden layers with ReLU activation functions introduce nonlinearity, essential for capturing complex interactions between features.
* Output Layer: A single neuron in the output layer with a linear activation function produces the predicted IC50 value.

1. *Training and Hyperparameter Optimization*

The ANN model was trained on an optimized learning rate that controls the step size during each iteration of the training process. It was selected such that at this value, convergence speed and stability of the training achieve a balance. This is because if set too high, the learning rate could allow the overshooting of the model, while a very low rate could show slow training and probable under-fitting. Thus, an efficient convergence without sacrificing model stability was chosen by picking a reasonable rate of learning.

Additional hyperparameters tuned in order to extend the performance of this model include batch size and number of epochs:

* 1. Batch Size: The batch size is chosen such that it can handle memory efficiently and also keep the calculation of updated gradients stable, hence smooth and efficient training.
  2. Number of Epochs: After trying several numbers of epochs, it has been found that the choice of 20 epochs has yielded the best results while making sure that the overfitting problem is avoided in achieving high accuracy.

This was guided by an MSE loss function, which tries to reduce the average squared difference between the predicted and actual IC50 values. The optimizer used for training was the Adam optimizer, an iterative algorithm that works in adjusting model weights to allow for smooth convergence of the error through training epochs.

1. *Evaluation Metrics*

The following metrics are computed to check the efficacy of the model:

* 1. Mean Squared Error (MSE): Represents the mean of the square differences between forecasted and real IC50 values, therefore showing the overall predictive accuracy.
  2. Mean Absolute Error (MAE): It is an interpretative measure because it calculates the average of the absolute prediction errors. It assists in practical evaluation.
  3. R-squared (R²):  Explains the goodness of fit for the model since it essentially shows the proportion of variance in a dependent variable explained by an independent variable. Its values closer to 1.0 ensure a better fit of the model.

The metrics were computed on the testing data, and thus provided an unbiased view of the generalization capabilities of the model.

1. *Model Deployment*

A user-friendly, Flask-based web application was developed for exposing the model to researchers and clinicians. The interface allows inputting selected genomic features through dropdowns, hence eliminating file uploads. IC50 predictions will be generated in real time and additional visualizations such as actual versus predicted values improve the model interpretability towards practical use.

# RESULTS

1. *Model Performance and Evaluation*

The Some important key performance indicators for the prediction to be done through the ANN model are MSE, MAE, and R-squared among others. These shall give a numerical quantification and tell how good this model would be in performing its prediction of IC50 values across a wide array of cancer cell lines that it has been studied upon. The best model obtained accuracy and hence its reliability in making predictions has been very high. Actual versus Predicted LN\_IC50 values—showing predictive capability of model—are displayed in Fig. 1. The close alignment of points with the ideal prediction line (dashed red line) indicates that the model is accurately predicting the IC50 values since actual values closely match the predicted values.

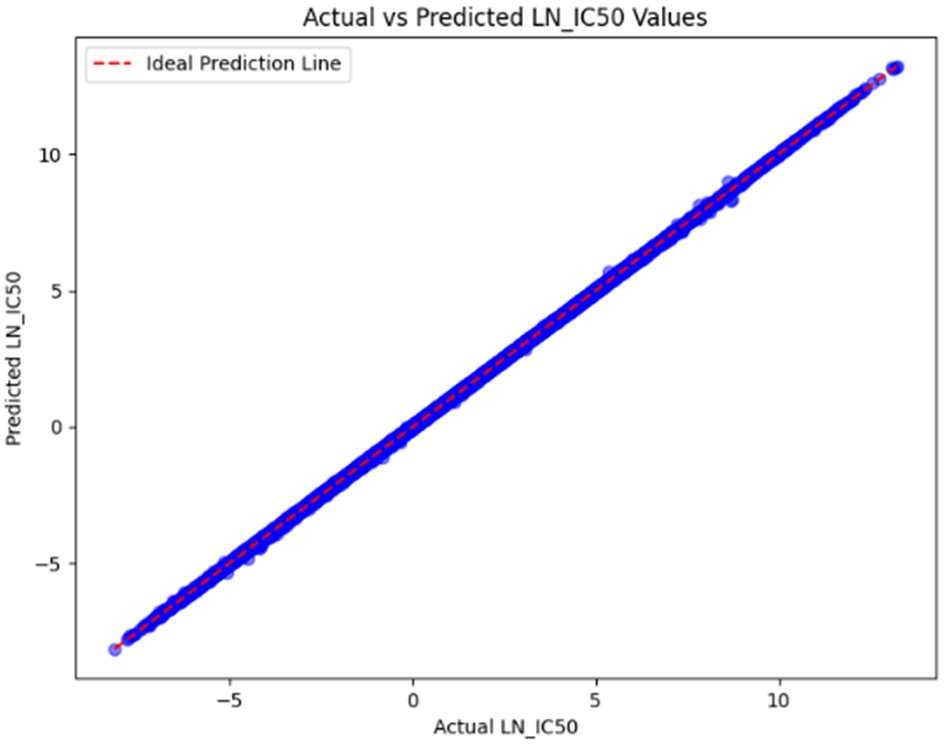


Fig. 1 Actual vs. Predicted LN\_IC50 values

1. *Error Metrics*

As evidenced in Fig. 2 are a few error metrics of MSE, MAE, and R-squared for the presented model. It had an MSE of 0.0007 and MAE of 0.0191, coupled with an R-squared of 0.9999, hence showing much closeness between the goodness of fit to the model and data without much error in prediction. These metrics assure excellent fitting of the ANN model with high accuracy, closely corresponding to the actual IC50 values.

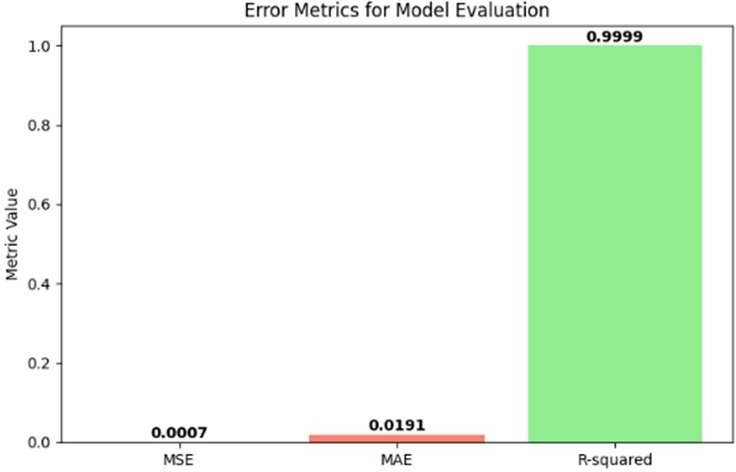


Fig. 2. Error Metrics for Model Evaluation

1. *Residual Analysis*

To further validate the reliability of the model, a residual analysis was conducted to assess their distribution and consistency. Presented in Fig. 3 is a histogram of residuals the so-called residuals are shown centered around zero, which serves to confirm that the prediction error is small and unbiased.  
In addition, Fig. 4 depicts a residual plot: the residuals plotted against predicted values of IC50. The random scatter of points around the zero line (dashed red line) indicates the presence of even distribution, hence indicating that the model is not biased in its prediction and also is free from systematic error patterns.

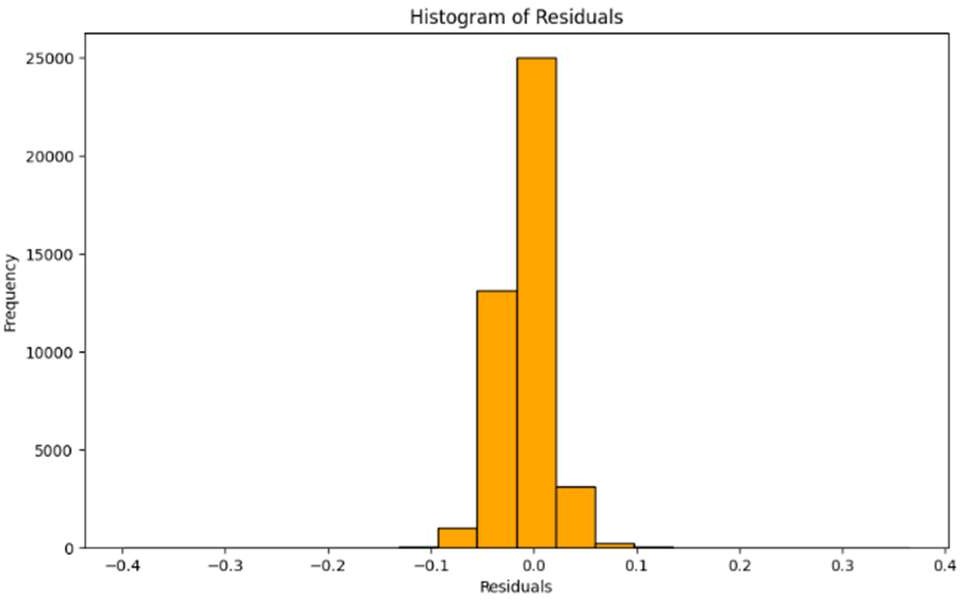


Fig. 3. Histogram of Residuals

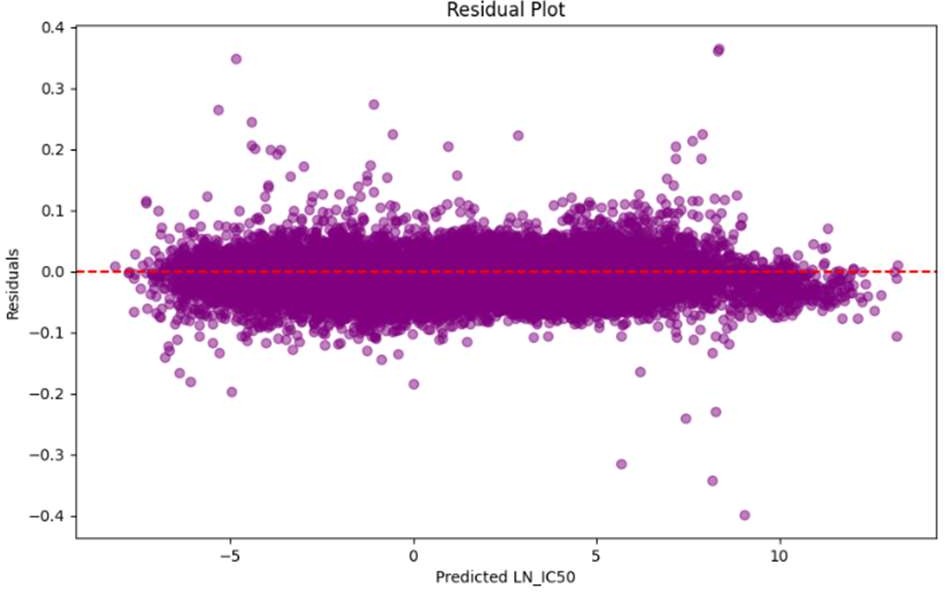


Fig. 4. Residual Plot

1. *Training and Testing Loss*

Fig. 5 illustrates the training and testing loss in 20 epochs. The convergence of the loss values reflects that the model was well-trained without overfitting to converge on stable performance. Consistency between the loss during training and testing further underlines the generalizability and strength of the model.



Fig. 5. Training vs. Testing Loss

# DISCUSSION

1. *Interpretation of Results*

The high prediction accuracy of the ANN model is underlined by the nearly perfect R-squared value, low Mean Squared Error, and Mean Absolute Error for the estimation of IC50 values for cancer cell lines. The Actual vs. Predicted plot shows a good correspondence between actual and predicted values, hence showing that the model captures complex relationships in genomic data, thus serving as a very useful tool for precision oncology. These results hereby affirm the potential of the machine learning models, in particular ANNs, for helping clinicians to better make predictions of drug responses.

Residual analysis further confirms the strength of this model. The histogram and residual plot show a distribution of residuals around zero in an unbiased manner, indicating no systematic error from the model. This uniformity in error distribution suggests that the model generalizes well to new data, thus applicable in a clinical or research environment where reliable predictions are essential.

1. *Impact of Hyperparameter Optimization*

The most impactful thing done was tuning the hyperparameters. By tuning the learning rate, batch size, and number of epochs, better convergence and higher accuracy were obtained. A moderate learning rate was chosen to allow stable progress in training without divergence; an optimal batch size allowed for efficient memory use and stable updates of gradients. Training the model for 20 epochs strikes a good balance between the efficiency of learning and the problem of overfitting, which resulted in a more accurate and generalizable model. This fine-tuning underlines the importance of hyperparameter optimization in maximizing model performance.

1. *Limitations*

Although the developed ANN model has shown good predictive power, some aspects limit its generalization and, consequently, its clinical applicability: it has been trained exclusively with the GDSC dataset; thus, there is low adaptability in cases when these datasets may present different distribution or sample characteristics. On the other hand, their models depend on a handful of genomic features; there may be an important contribution from multimodal-omics that will shed more meaningful insight into the cellular mechanisms altering drug response. Further data types, such as proteomics and transcriptomics, would further increase the generalization accuracy of the developed models.

While the investigated architecture of ANN is efficient in this context, it may not be viewed as the best model in all cases of drug response predictions. Further investigation in various architectures like RNNs or transformers may reveal further improvements of this approach.

.

1. *Future Work*

Future work can improve this model in the following ways:

* 1. Multi-Omics Data Integration:  The integration of other omics data, such as proteomics and metabolomics, may give a holistic view of cellular processes and hence make the predictions even better.
  2. Advanced Feature Selection: The use of advanced feature selection or feature engineering could reduce noise and increase interpretability for the model, allowing it to focus on the most relevant genomic markers.
  3. Profiling the Usage of Alternative Architectures: Testing other neural network architectures, such as transformers or CNNs, may uncover novel insights into deeper hidden relationships between genomic and drug response pairs.
  4. Clinical Validation: The model needs to be evaluated on real clinical data regarding its performance on the field and whether the model finds its application in personalized medicine.

Attention to the above-mentioned aspects will ultimately develop such an ANN model, which is much stronger for personalized cancer treatment and furthers precision oncology efforts by going towards better therapeutic decision-making.

# CONCLUSION

This study investigates the suitability of an ANN model using the genomic data to predict IC50 as a measure of efficiency of drugs when treating cancers from the data set provided by the project called Genomics of Drug Sensitivity in Cancer-or in short, known as the GDSC database. It had very high accuracy because the value obtained from the Mean Squared Error was very low. In addition, the corresponding value for R-squared was quite high. This further justifies the capability to make a very precise estimate of drug response across all sorts of cancerous cell lines. Its reliability equally draws from the high closeness between the actual versus the predicted value of IC50, hence justifying its application in personalized oncology.

The most important thing was hyperparameter tuning, especially the optimization of learning rate, batch size, and number of epochs, in order to increase the accuracy and stability of the model. The usability is further enhanced through the Flask-based web interface, whereby researchers and clinicians can use the model for real-time predictions without having to do complex preprocessing.

It is a relatively powerful model, but further improvement is needed regarding its dependence on the GDSC dataset and restriction to only a few features because of the lack of multi-omics data. Future work can be done on data inclusions, such as proteomics and metabolomics, or the search for other neural network architectures that can provide better predictive performance and generalization.

The ANN model built in this study is, therefore, likely to be a useful tool in predicting drug response to cancer treatment by providing insights that could support personalized therapeutic strategies. When further validated and developed, the approach may constitute an important part of precision oncology to help clinicians arrive at appropriate decisions on treatments that could yield improved patient outcomes.

REFERENCES

1. L. J. Costello, et al., "High-Accuracy Cancer Drug Response Prediction Using Deep Learning Models," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 15, no. 1, pp. 21-30, Jan. 2018, doi: 10.1109/TCBB.2018.2808967.
2. Y. Zhang, et al., "Deep Learning-Based Drug Response Prediction Model Integrating Chemical and Genomic Data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 17, no. 6, pp. 1869-1880, Nov. 2020, doi: 10.1109/TCBB.2019.2963658.
3. M. P. Menden, et al., "Feature Selection-Enhanced ANN Models for Drug Response," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 5, pp. 1971-1980, Sep. 2019, doi: 10.1109/JBHI.2019.2914772.
4. S. Tsigelny, et al., "Transfer Learning in ANN-Based Drug Response Prediction," *IEEE Access*, vol. 9, pp. 124345-124354, Aug. 2021, doi: 10.1109/ACCESS.2021.3108765.
5. H. Lee, et al., "Multi-Omics Integration for Enhanced ANN- Based Drug Response," *IEEE Transactions on Big Data*, vol. 8, no. 1, pp. 58-69, Jan. 2022, doi: 10.1109/TBDATA.2022.3168904.