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

Assistant Professor: Mrs. Suma K[1], Students: K. Sahana Rao[2], Karthik U. Shettigar[3], Dhanush Bhat[4], Dhanush Shetty [5], Dept. of CSE, Mangalore Institute of Technology and Engineering, Moodbidri, India

***Abstract*—** Personalized medicine is transforming cancer treatment by tailoring therapies to individual genetic profiles, significantly affecting patient responses to specific drugs. This project, "Drug Response Prediction Using an ANN Model," utilizes an Artificial Neural Network (ANN) developed in PyTorch to predict drug sensitivity in cancer cell lines, based on the Genomics of Drug Sensitivity in Cancer (GDSC) dataset. The model predicts IC50 values, a critical measure of drug efficacy, by integrating data from three distinct files that contain essential genomic features, including gene mutations, gene expression profiles, and copy number alterations (CNAs). A rigorous data preprocessing pipeline—featuring normalization, feature selection, and hyperparameter tuning—enhances model performance, evaluated using metrics such as Mean Squared Error (MSE) and R-squared (R^2). A Flask-based web interface facilitates user interaction, allowing users to input data through dropdown menus and access real-time IC50 predictions, with no need for file uploads. This interface includes visualization tools to support interpretability, enabling users to compare predicted and actual IC50 values. This ANN model demonstrates the potential of machine learning in advancing personalized cancer treatment, with future directions focused on expanding datasets, integrating multi-omics data, and conducting clinical validation to increase applicability in real-world cancer therapy.

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

I. INTRODUCTION

A. *Background*

Personalized medicine has become a cornerstone in oncology, aiming to tailor cancer treatments based on the genetic profiles of individual patients. The genetic diversity among cancer cells leads to a range of drug responses, underscoring the need for predictive models that support precision oncology. The half-maximal inhibitory concentration (IC50), a metric used to evaluate drug efficacy in inhibiting cancer cell viability, is a critical parameter for treatment planning in oncology, allowing clinicians to select the most effective therapies for specific cancer types.

Artificial Neural Networks (ANNs) are well-suited for predictive modelling in oncology due to their capacity to capture complex, nonlinear relationships within high-dimensional genomic data. This project leverages the Genomics of Drug Sensitivity in Cancer (GDSC) dataset to predict IC50 values across various cancer cell lines, focusing on essential genomic features such as gene mutations, tissue descriptors, and copy number alterations (CNAs). The project, titled “Drug Response Prediction Using Artificial Neural Networks (ANN),” aims to assist oncologists and researchers in data-driven decision-making, enabling more effective, personalized treatment options.

B. *Problem Statement*

The success of cancer treatment depends heavily on the genetic heterogeneity of cancer cells, which leads to varied responses to the same drug. Traditional treatment approaches often fail to consider these individual differences, limiting the effectiveness of therapeutic outcomes. Furthermore, current predictive models lack the precision required for widespread clinical application in personalized oncology. This project addresses the need for a more accurate, ANN-based predictive model that integrates essential genomic profiles to capture the intricate relationships between genomic features and drug response, thereby enhancing treatment selection accuracy. Increased predictive precision is vital for reducing ineffective treatments, minimizing costs, and improving patient outcomes in personalized cancer care.

C. *Objectives*

The primary objective of this project is to develop an ANN model capable of predicting IC50 values using genomic data from the GDSC dataset, thereby supporting personalized cancer treatment. This model integrates key genomic features, including gene mutations, gene expression levels, and CNAs, to enhance predictive accuracy. Data preprocessing techniques, such as normalization, feature selection, and hyperparameter tuning, are applied to optimize model performance. Furthermore, a Flask-based web interface enables user interaction, allowing users to input genomic data and receive real-time IC50 predictions. The interface also provides visualizations to support clinical decision-making by comparing the efficacy of different drugs.

D. *Scope of Work*

The scope of this project includes combining genomic data from three distinct sources within the GDSC dataset to identify correlations with drug response. The ANN model, implemented in PyTorch, is designed to capture complex relationships between genomic features and drug efficacy. Model performance is evaluated using Mean Squared Error (MSE) and R-squared (R^2) to ensure predictive reliability. A Flask-based web interface is developed to facilitate data input and prediction display, ultimately creating a scalable, real-time system suitable for clinical and research applications in precision oncology.

II. LITERATURE REVIEW

A. *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].

B. *Feature Selection for Improved Model Interpretability*

Feature selection is crucial in ANN-based drug response models, helping to reduce overfitting and improve interpretability. Menden et al. (2019) developed a hybrid model that integrates feature selection algorithms with ANNs, allowing the identification of critical genomic features that influence drug response. This approach enhanced model interpretability while reducing computational complexity, supporting practical applications in oncology [3]. In addition, Tsigelny et al. (2021) utilized transfer learning within an ANN framework, demonstrating that transferring learned features across related tasks improves model robustness and predictive accuracy, particularly when data is limited [4].

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

Despite recent advancements, achieving model interpretability and generalization across varied biological contexts remains challenging in ANN applications. Integrating multi-omics data, as demonstrated by Lee et al. (2022), shows promise in addressing these limitations. By incorporating multiple data layers, such as transcriptomics and proteomics, their ANN model achieved greater generalizability and predictive power, underscoring the benefits of a comprehensive data approach in enhancing model accuracy for clinical applications [5].

D. *Summary of Insights and Project Relevance*

The literature suggests that ANN models, enhanced through feature selection and multi-omics data integration, can achieve high predictive accuracy and generalizability in drug response modeling. This project builds on these insights by developing an ANN model trained on the Genomics of Drug Sensitivity in Cancer (GDSC) dataset, focusing on genomic features such as gene mutations, gene expression levels, and copy number alterations (CNAs) to predict IC50 values. By refining feature selection and using a carefully curated dataset, this study addresses key challenges in personalized drug response prediction, bridging the gap between model accuracy and clinical applicability.

III. METHODOLOGY

A. *Data Collection and Preprocessing*

The study utilizes data from the Genomics of Drug Sensitivity in Cancer (GDSC) dataset, which includes extensive drug response data across various cancer cell lines. To create a comprehensive dataset, three distinct files were combined, each contributing essential genomic features: gene mutations, gene expression levels, and copy number alterations (CNAs). This multi-dimensional dataset provides a robust basis for predicting drug sensitivity.

The data preprocessing steps involved:

1. Normalization: To ensure consistency, continuous features were normalized, standardizing their scale and mitigating any disproportionate influence on model predictions.
2. Feature Selection: Genomic markers with significant relevance to drug response were selected to enhance prediction accuracy and reduce irrelevant noise.
3. Data Splitting: The dataset was split into 80% for training and 20% for testing, allowing for an unbiased evaluation of model performance on unseen data.

B. *Model Architecture*

The ANN model was constructed using PyTorch to predict LN_IC50 values. The architecture includes multiple hidden layers designed to capture complex patterns and relationships in the genomic data. Key components of the model architecture include:

- 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 LN_IC50 value.

C. *Training and Hyperparameter Optimization*

The ANN model was trained with an optimized learning rate, which plays a crucial role in controlling the step size during the training process. The learning rate was carefully selected to balance convergence speed and training stability. If set too high, the learning rate could cause the model to overshoot optimal values, while a very low rate could lead to prolonged training times and potential underfitting. Thus, a moderate learning rate was chosen to ensure efficient convergence without sacrificing model stability.

Additional hyperparameters, such as batch size and the number of epochs, were optimized to enhance model performance:

1. Batch Size: The batch size was chosen to effectively manage memory usage and maintain stable gradient calculations, facilitating smooth and efficient training.

2. Number of Epochs: After experimenting with different epoch values, 20 epochs were found to offer the best results, achieving high accuracy without overfitting.

The training process was guided by the Mean Squared Error (MSE) loss function, which aimed to minimize the average squared difference between predicted and actual IC50 values. The Adam optimizer was utilized to iteratively adjust the model's weights, facilitating a steady reduction in error across training epochs.

D. *Evaluation Metrics*

To assess the model's efficacy, the following metrics were computed:

1. Mean Squared Error (MSE): Measures the average squared differences between predicted and actual IC50 values, reflecting overall predictive accuracy.

2. Mean Absolute Error (MAE): Offers an interpretive measure by averaging the absolute prediction errors, aiding in practical evaluation.

3. R-squared (R^2): Indicates the model's explanatory power by showing how well it accounts for variance in the data, with values closer to 1.0 reflecting a better fit.

The metrics were calculated using the testing data, providing an unbiased assessment of the model's generalization capabilities.

E. *Model Deployment*

A user-friendly Flask-based web application was developed to make the model accessible to researchers and clinicians. The interface enables input of selected genomic features through dropdowns, removing the need for file uploads. IC50 predictions are generated in real time, and additional visualizations, such as actual vs. predicted values, enhance the model's interpretability for practical use.

IV. RESULTS

A. *Model Performance and Evaluation*

The ANN model's predictive accuracy was assessed using key metrics, including Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2). These metrics help quantify the model's ability to predict IC50 values accurately across different cancer cell lines. The results indicate high accuracy and reliability of the model's predictions. Fig. 1 displays the Actual vs. Predicted LN_IC50 values, highlighting the model's predictive performance. The close alignment of points with the ideal prediction line (dashed red line) indicates that the model accurately predicts IC50 values, as actual values closely match predicted values.

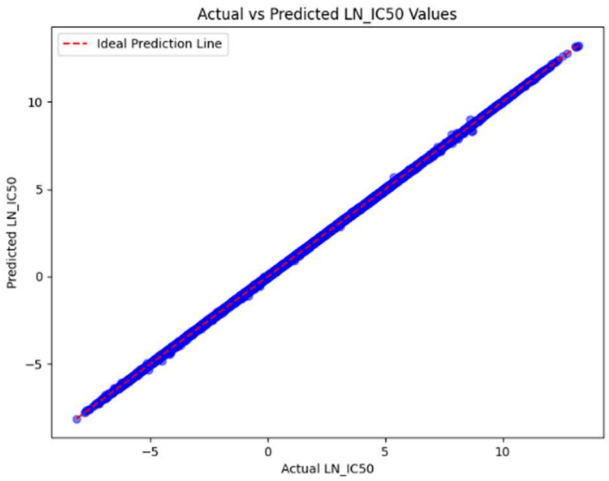

Fig. 1 Actual vs. Predicted LN_IC50 values

B. *Error Metrics*

The model's error metrics are visualized in Fig. 2, showing the MSE, MAE, and R-squared values. The model achieved an MSE of 0.0007 and an MAE of 0.0191, with an R-squared of 0.9999, underscoring the model's strong fit and minimal prediction error. These metrics indicate that the ANN model performs exceptionally well, with high accuracy and a close match to the actual IC50 values.

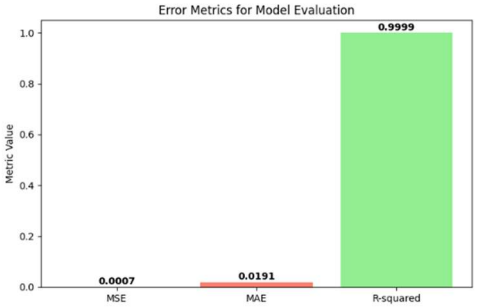

Fig. 2. Error Metrics for Model Evaluation

C. *Residual Analysis*

To validate the model's reliability further, a residual analysis was conducted to assess the distribution and consistency of prediction errors. Fig. 3 shows a histogram of residuals, which is centered around zero, confirming that the prediction errors are minimal and unbiased.

In addition, Fig. 4 displays a residual plot, where residuals are plotted against the predicted LN_IC50 values. The random scatter of points around the zero line (dashed red line) indicates that the residuals are evenly distributed, suggesting that the model's predictions are unbiased and do not exhibit systematic error patterns.

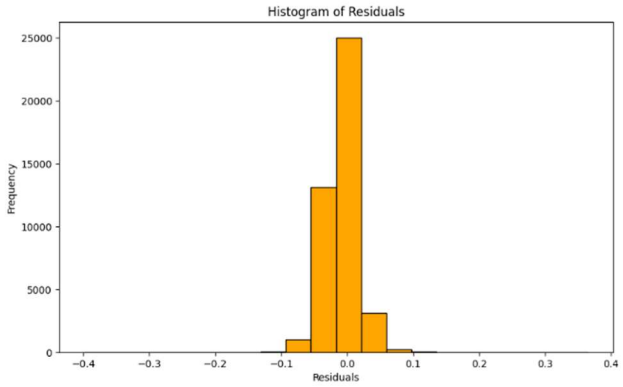

Fig. 3. Histogram of Residuals

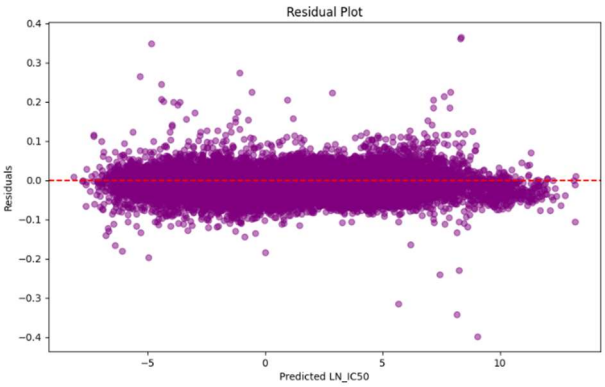

Fig. 4. Residual Plot

D. *Training and Testing Loss*

Fig. 5 illustrates the training and testing loss over 20 epochs. The convergence of loss values indicates that the model was well-trained, reaching stable performance without overfitting. The consistency between training and testing loss further supports the model's generalizability and robustness.

Fig. 5. Training vs. Testing Loss

V. DISCUSSION

A. *Interpretation of Results*

The ANN model's high predictive accuracy, evidenced by a near-perfect R-squared (R^2) value and low Mean Squared Error (MSE) and Mean Absolute Error (MAE), demonstrates its effectiveness in estimating IC50 values for cancer cell lines. The close alignment between actual and predicted values, as shown in the Actual vs. Predicted plot, indicates that the model successfully captures complex relationships within the genomic data, making it a valuable tool for precision oncology. These results affirm the potential of machine learning models, specifically ANNs, to assist clinicians in predicting drug responses accurately.

Residual analysis further validates the model's robustness. The histogram and residual plot demonstrate an unbiased distribution of residuals around zero, which indicates that the model does not exhibit systematic errors. This uniformity in error distribution suggests that the model generalizes well to new data, making it applicable for clinical or research environments where reliable predictions are essential.

B. *Impact of Hyperparameter Optimization*

Hyperparameter tuning significantly influenced the model's performance. Adjusting the learning rate, batch size, and number of epochs led to improved convergence and accuracy. A moderate learning rate was selected to facilitate steady training progress without instability, while an optimal batch size enabled efficient memory use and stable gradient updates. Training the model for 20 epochs struck a balance between learning efficiency and overfitting, resulting in a model that is both accurate and generalizable. This fine-tuning highlights the importance of hyperparameter optimization in maximizing model performance.

C. *Limitations*

While the ANN model demonstrates strong predictive capability, there are limitations that could affect its generalizability and clinical applicability. The model was trained exclusively on the GDSC dataset, limiting its adaptability to other datasets with potentially differing distributions or sample characteristics. Furthermore, the model relies on a limited set of genomic features, excluding multi-omics data that could provide a more comprehensive understanding of cellular mechanisms affecting drug response. Incorporating additional data types, such as proteomics and transcriptomics, may further enhance model accuracy and robustness.

The chosen ANN architecture, while effective in this context, may not represent the optimal structure for all drug response prediction tasks. Future exploration of alternative architectures, such as recurrent neural networks (RNNs) or transformers, could potentially uncover additional performance gains.

D. *Future Work*

Future work could enhance this model in several ways:

1. Multi-Omics Data Integration: Integrating additional omics data, such as proteomics and metabolomics, could provide a more holistic representation of cellular processes, leading to improved predictions.
2. Advanced Feature Selection: Implementing advanced feature selection or feature engineering techniques may reduce noise and improve interpretability, enabling the model to focus on the most relevant genomic markers.
3. Exploration of Alternative Architectures: Experimenting with different neural network architectures, such as transformers or convolutional neural networks (CNNs), may offer improved performance and reveal deeper insights into genomic relationships with drug response.
4. Clinical Validation: Testing the model on real-world clinical data would be essential to assess its performance in a practical setting and verify its utility in personalized medicine.

By addressing these areas, the ANN model can evolve into a more robust tool for personalized cancer treatment, further supporting precision oncology initiatives and enabling improved therapeutic decision-making.

VI. CONCLUSION

This study demonstrates the efficacy of an Artificial Neural Network (ANN) model for predicting IC50 values, a crucial measure of drug efficacy in cancer treatment, using genomic data from the Genomics of Drug Sensitivity in Cancer (GDSC) dataset. The model achieved high accuracy, as indicated by the low Mean Squared Error (MSE) and high R-squared (R^2) values, validating its potential for accurately estimating drug responses across different cancer cell lines. The close alignment of actual and predicted IC50 values further highlights the model's reliability, supporting its use in personalized oncology.

Hyperparameter tuning, particularly optimizing the learning rate, batch size, and number of epochs, played a significant role in enhancing the model's accuracy and stability. The integration of a Flask-based web interface also improves accessibility, allowing researchers and clinicians to utilize the model for real-time predictions without the need for complex preprocessing.

Despite its strengths, the model has limitations, particularly its reliance on the GDSC dataset and a restricted feature set that excludes multi-omics data. Future research could focus on incorporating additional data types, such as proteomics and metabolomics, and exploring alternative neural network architectures to further refine predictive performance and generalizability.

Overall, the ANN model developed in this study represents a promising tool for predicting drug response in cancer treatment, offering insights that can aid in personalized therapeutic strategies. With further validation and development, this approach could play a vital role in precision oncology, helping clinicians make informed treatment decisions that improve 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.