ONCOGUIDERX: A DEEP LEARNING-BASED PLATFORM FOR PREDICTING PERSONALIZED CANCER DRUG SENSITIVITY

Project Reference No.: *48S_BE_0789*

College : Mangalore Institute Of Technology And Engineering, Moodbidri
Branch : Department Of Computer Science And Engineering
Guide : Mrs. Suma K
Student(s) : Ms. K Sahana Rao
Mr. Karthik U Shettigar
Mr. Dhanush A
Mr. Dhanush S Shetty

Keywords:

Artificial Neural Network, IC50 Prediction, Genomics, Drug Sensitivity, Personalized Medicine.

Introduction:

Personalized medicine is revolutionizing cancer treatment by tailoring therapies to the genetic makeup of patients. However, predicting how an individual will respond to a particular drug remains a complex task. Our project, titled “Oncoguiderx: A Deep Learning-Based Platform for Predicting Personalized Cancer Drug Sensitivity”, aims to tackle this challenge using machine learning. It focuses on predicting IC50 values a key metric for drug effectiveness based on genomic features like gene mutations, gene expression, and copy number alterations (CNAs) from the Genomics of Drug Sensitivity in Cancer (GDSC) dataset. The project’s deliverables include a high-accuracy Artificial Neural Network (ANN) model developed using PyTorch and a web-based interface developed with Flask for real-time predictions. This system is designed to assist researchers and clinicians in selecting effective cancer treatments using predictive modeling and visualizations.

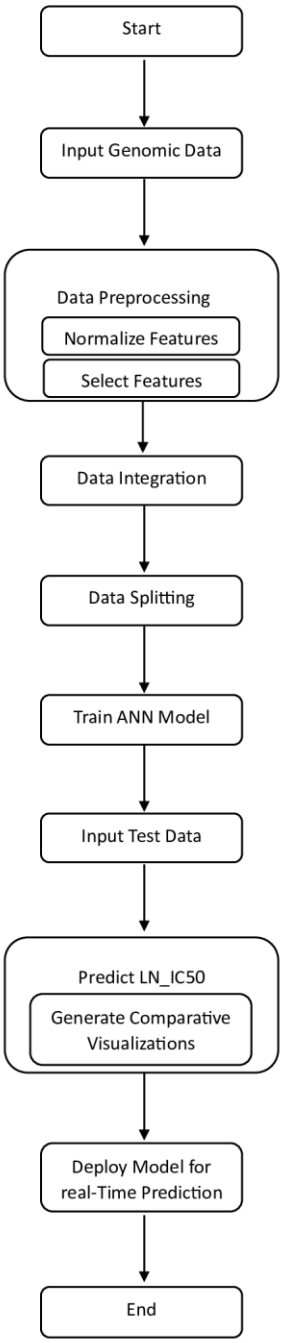

Figure 1: System Architecture

Objectives:

- Merge data from the GDSC, gene expression, mutation profiles, and CNAs to enhance the accuracy and relevance of drug response predictions.
- Build an ANN model to predict IC50 values based on genomic data from multiple datasets.

- Develop a Flask-based interface to facilitate easy data input.
- Enable users to visualize IC50 predictions along with a comparative graph for five selected drugs

Methodology:

This project utilizes genomic and drug response data from the **Genomics of Drug Sensitivity in Cancer (GDSC)** dataset. Three data files were merged to form a comprehensive dataset, incorporating gene mutations, gene expression levels, and copy number alterations (CNAs), all critical to predicting drug sensitivity (IC50).

The preprocessing pipeline included:

- **Normalization** of continuous features to maintain scale uniformity.
- **Feature selection** to identify and retain the most relevant genomic features while minimizing noise.
- **Train-test split** at an 80:20 ratio to ensure unbiased evaluation.

A **PyTorch-based Artificial Neural Network (ANN)** model was developed. The architecture includes:

- An **input layer** for selected genomic features.
- **Multiple hidden layers** using ReLU activation to capture nonlinear relationships.
- A **single output neuron** with a linear activation to predict the LN_IC50 value.

Training was performed using:

- The **Adam optimizer** for weight updates.
- **Mean Squared Error (MSE)** as the loss function.
- Tuned hyperparameters: a moderate **learning rate**, 20 **epochs**, and an optimal **batch size** to ensure convergence and stability.

Model evaluation was based on:

- **MSE** and **MAE** for prediction accuracy,
- **R-squared (R^2)** to measure the variance explained by the model.

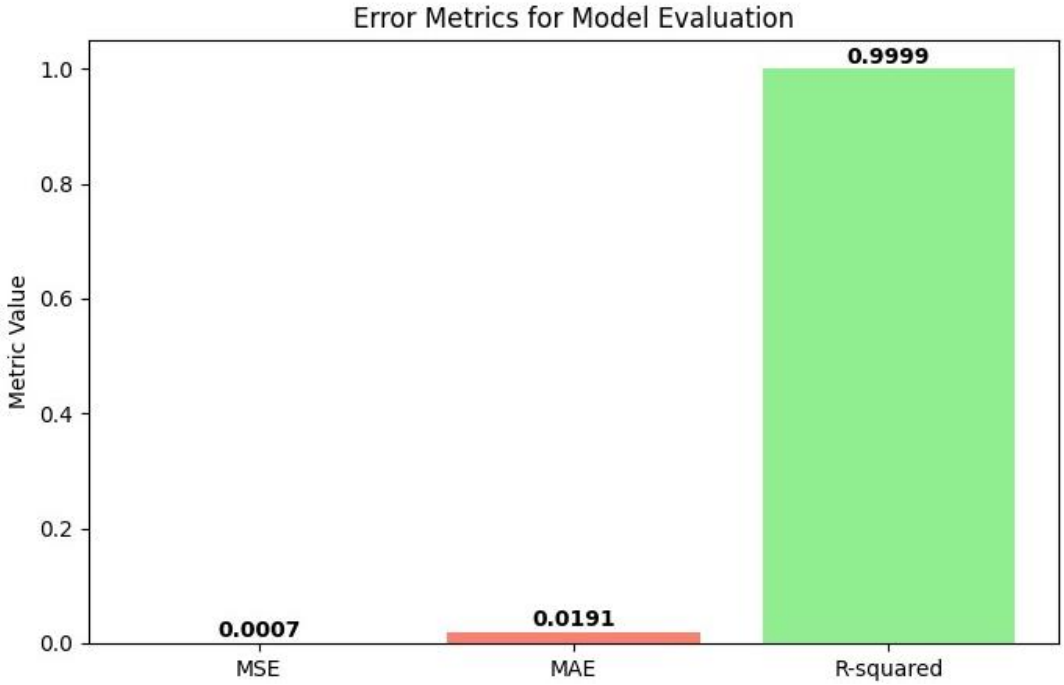

Figure 2: Error Metrics for Model Evaluation

Finally, the trained model was deployed using a **Flask-based web interface**, allowing users to enter genomic data via dropdowns. The application provides **real-time IC50 predictions** and visualizations (e.g., actual vs. predicted IC50 values), making it practical for research and clinical settings.

Result and Conclusion:

The project successfully developed an Artificial Neural Network (ANN) model for predicting cancer drug sensitivity using genomic data from the Genomics of Drug Sensitivity in Cancer (GDSC) dataset. The ANN was trained on key features like gene mutations, tissue descriptors, and copy number alterations (CNAs) to predict IC50 values a critical indicator of drug efficacy.

The model demonstrated excellent predictive power, achieving a high R^2 score (≈0.9999), very low MSE (≈0.0007), and MAE (≈0.0191). These results confirm its

generalizability and robustness across unseen genomic profiles. Hyperparameter tuning, feature selection, and early stopping mechanisms played a vital role in ensuring model stability and performance.

A major outcome of this project is the deployment of a user-friendly Flask web application. The app allows real-time IC50 prediction via a simple interface using dropdown inputs. Users can compare the predicted IC50 values for up to three drugs simultaneously, supporting personalized treatment selection.

In summary, this project validates the practical utility of ANN models in precision oncology. The combination of accurate predictions and an intuitive web interface makes it suitable for both clinical decision support and academic research.

Future Scope:

The future scope of this project includes:

1. Multi-Omics Integration: Future versions can incorporate proteomics, transcriptomics, or metabolomics for a more holistic understanding of tumor biology and drug response.
2. Expanded Dataset Coverage: Integrating additional databases like CTRP and PDX can improve generalizability across diverse cancer types.
3. Dimensionality Reduction: Advanced techniques like PCA or autoencoders could optimize feature selection and reduce model complexity while maintaining performance.
4. Explainable AI: Integration of tools like SHAP or LIME can enhance model transparency and clinical interpretability by showing how features influence predictions.
5. Cloud-Based Deployment: Moving from local deployment to a cloud platform can offer better scalability, accessibility, and support for real-time updates.
6. Personalized Reports: The interface can be extended to generate downloadable, patient-specific drug response reports for use in clinical workflows.
7. Cross-Validation and Regularization: Future training cycles can incorporate K-fold cross-validation and dropout/weight decay to enhance model robustness.