ABSTRACT

Personalized medicine is transforming cancer treatment by tailoring therapies to the genetic variations found in cancer cells, significantly improving patient responses to specific drugs. This project, titled "Leveraging Ann for Targeted Drug Sensitivity Prediction on GDSC Data", utilizes a machine learning approach with Artificial Neural Networks (ANN) to predict drug sensitivity in cancer cell lines using data from the Genomics of Drug Sensitivity in Cancer (GDSC) dataset. The model focuses on predicting IC50 values, a key measure of drug effectiveness, by integrating genomic features such as gene mutations, gene expression profiles, and copy number alterations (CNAs) to identify potential biomarkers for personalized therapies. The system employs an ANN architecture that captures complex nonlinear relationships within the genomic data, optimizing predictions with techniques like normalization, feature selection, and hyperparameter tuning. The approach uses a Flask-based web interface that enables users to easily input genomic data through labeled form fields, ensuring accessibility without the need for file uploads. The interface displays IC50 predictions in real-time, making it suitable for both clinical and research use. Model performance is evaluated using key metrics like Mean Squared Error (MSE), Mean Absolute Error (MAE) and R-squared (R^2). Visualization tools further enhance interpretability, displaying the comparison between predicted and actual IC50 values. This project highlights the potential of ANN models in advancing personalized oncology treatment by providing actionable drug response insights based on genomic data. Future work will focus on integrating additional datasets, enhancing multi-omics data integration, and conducting clinical validation to ensure broader applicability in real-world cancer therapies.

CHAPTER 1

INTRODUCTION

Personalized medicine is gaining prominence in cancer treatment due to the diverse genetic makeup of cancer cells, which leads to varied responses among patients undergoing similar therapies. This project, titled "Leveraging Ann for Targeted Drug Sensitivity Prediction on GDSC Data", aims to create a model that leverages genomic data to predict how different cancer cell lines respond to specific drugs. The complexity of cancer biology and the uniqueness of each patient's genetic profile necessitate computational tools that can support precision oncology by identifying the most effective drug treatments. This predictive model, based on the ANN architecture, seeks to aid oncologists and researchers by providing tailored predictions of drug efficacy based on specific cancer genomics, thus advancing the approach toward personalized therapies.

Our project uses the Genomics of Drug Sensitivity in Cancer (GDSC) dataset, which contains essential genomic features and drug response data across over 1,000 cancer cell lines. Focusing on IC50 prediction, a critical metric for assessing drug potency, the model leverages features such as gene mutations, expression profiles, and cancer-type descriptors to capture relevant patterns in drug sensitivity. A user-friendly web-based interface developed with Flask allows researchers to input data for IC50 predictions seamlessly. By simplifying the interaction between the model and users, this web application supports data input through dropdown menus and displays prediction outputs in an accessible, clinician-oriented manner, aiming to streamline and support decision-making in cancer research and clinical practice.

1.1 Rationale

In cancer treatment, variability in patient responses to the same drug presents a major obstacle in effective therapy. Traditional statistical models often struggle with the complex and nonlinear nature of biological data. ANNs, with their ability to learn from vast datasets and uncover hidden relationships, offer a superior approach to predicting drug responses. By utilizing this model, we can provide insights into which drugs are more likely to be effective for specific patients based on their biological data, leading to more precise and personalized treatments.

1.2 Problem Statement

Cancer treatment efficacy is significantly impacted by the genetic diversity of cancer cells, causing variable responses to the same drug among patients. Traditional treatment approaches often do not account for these individual differences, leading to inconsistent therapeutic outcomes. Furthermore, the lack of reliable predictive models for drug response limits the scope of precision medicine. This project addresses the following key issues.

- **Limited personalized treatment options:** Conventional treatments do not consider individual genetic profiles.
- **Challenges in predicting drug efficacy:** The relationship between genomic features and drug response is complex and not fully understood.
- **High failure rates in therapy selection:** The absence of predictive accuracy can lead to ineffective treatments, higher costs, and increased side effects.

To tackle these challenges, our project proposes an ANN-based model trained on genomic data to predict drug sensitivity, helping oncologists make better-informed treatment decisions. By analyzing genomic features relevant to drug response, the model aims to provide more accurate predictions that align with the unique profiles of cancer patients.

1.3 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.

CHAPTER 2

LITERATURE REVIEW

1. Discovering the Mechanism of Action of Drugs with a Sparse Explainable Network

- Introduces SparseGO, an explainable neural network for predicting drug responses in cancer using gene expression data from GDSC1 and GDSC2.
- Offers high interpretability but has high computational requirements.

2. Deep Learning Methods for Drug Response Prediction in Cancer: Predominant and Emerging Trends

- Reviews 61 deep learning models that predict cancer response to drug treatments using data from GDSC.
- Emphasizes the lack of standardized evaluation frameworks, which makes comparing models difficult.

3. Machine Learning Approaches to Predict Drug Efficacy and Toxicity in Oncology

- Evaluates machine learning techniques for predicting drug sensitivity in cancer cell lines using GDSC data.
- Highlights high computational complexity and challenges with multi-omics data integration.

4. Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells

- Discusses machine learning models like random forests for predicting drug sensitivity using GDSC genomic data.
- Provides accurate predictions but lacks generalizability across different cancer types.

CHAPTER 3

METHODOLOGY

3.1. 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 strong 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.

3.2. Model Architecture

The ANN model was built 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.

3.3. 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 the actual and the predicted IC50 values. The Adam optimizer was utilized to iteratively adjust the model's weights, facilitating a steady reduction in error across training epochs.

3.4. Evaluation Metrics

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

1. Mean Squared Error (MSE): Measures the average squared differences between the actual and the predicted 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.

3.5. 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.

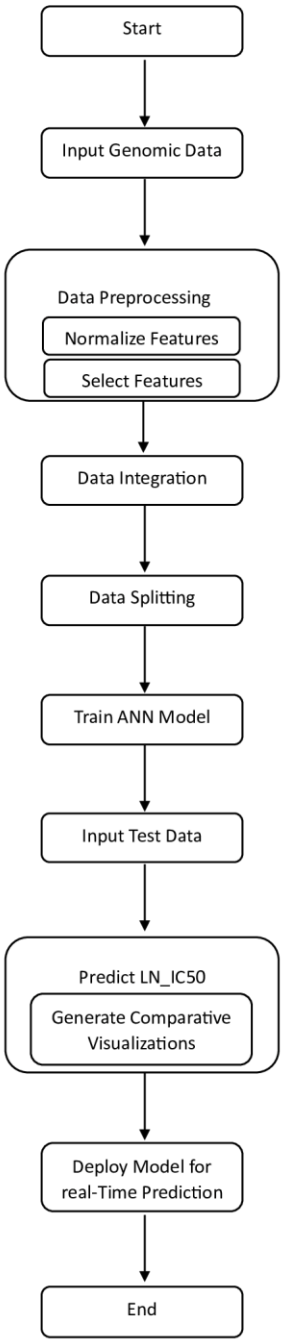

Figure 1. System Architecture

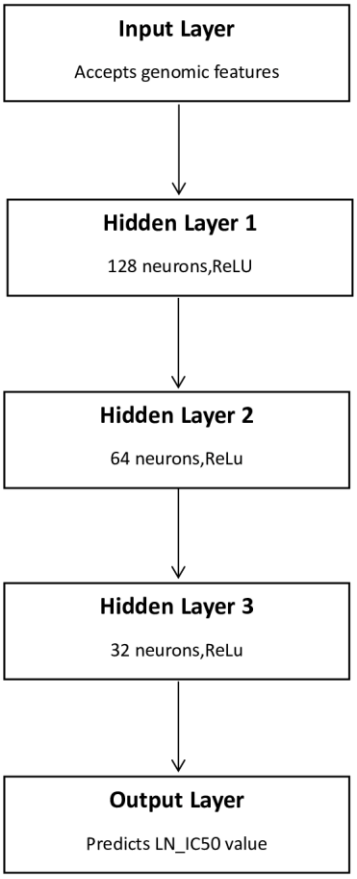

Figure 2. ANN Architecture

3.6 Software Requirements

- **Processor** : Intel Core i5 or equivalent
- **RAM** : Minimum 8 GB
- **Storage** : 50 GB for datasets and application files
- **GPU (Optional)** : NVIDIA GTX 1650 or equivalent for accelerated model training

3.7 Hardware Requirements

- **Operating System** : Windows 10 or compatible

- **Programming Language** : Python 3.x, HTML, CSS, and JavaScript
- **Libraries** : Pytorch, Pandas, Scikit-Learn, Matplotlib and Flask for backend
- **Development Environment** : Visual Studio Code & Jupyter Notebook

CHAPTER 4

OUTCOMES

- **Accurate Drug Response Predictions**: The Artificial Neural Network (ANN) model demonstrated high accuracy in predicting drug sensitivity based on biological and genomic data. The use of IC50 values allows for precise predictions of drug efficacy in various cell lines.
- **Improved Precision in Treatment Planning**: By accurately predicting drug responses, this system aids in the development of personalized treatment plans. Clinicians can tailor drug therapies based on a patient's specific biological profile, leading to more effective treatments.
- **Reduced Time for Drug Sensitivity Analysis**: The automated nature of the ANN model speeds up the process of drug sensitivity analysis, reducing the manual effort and time required for clinicians to assess potential drug efficacy.
- **Insights into Nonlinear Relationships in Biological Data**: The ANN's ability to learn complex, nonlinear patterns in biological data provides deeper insights into how gene expression and mutations influence drug responses, offering a more nuanced understanding than traditional methods.
- **User-Friendly Interface for Clinical Use**: The model has been integrated into a user-friendly interface, allowing healthcare professionals to input patient data and receive accurate drug response predictions with minimal technical expertise.
- **Potential for Scaling to Larger Datasets**: The project demonstrates that ANNs can be effectively scaled to handle large datasets, paving the way for future implementations on even more extensive biomedical data.

REFERENCES

[1] Katyna Sada Del Reala and Angel Rubio. (2023). "Discovering the Mechanism of Action of Drugs with a Sparse Explainable Network", *Journal of Biomedical Informatics*, 128, 104019.

[2] Alexander Partin, Thomas S. Brettin, Yitan Zhu, Oleksandr Narykov, Austin Clyde, Jamie Overbeek, and Rick L. Stevens. (2023). "Deep Learning Methods for Drug Response Prediction in Cancer: Predominant and Emerging Trends", *Nature Reviews Drug Discovery*, 22(4), 317-334.

[3] Bara A. Badwan, Gerry Liaropoulos, Efthymios Kyrodimos, Dimitrios Skaltsas, Aristotelis Tsirigos, and Vassilis G. Gorgoulis. (2023). "Machine Learning Approaches to Predict Drug Efficacy and Toxicity in Oncology", *Computational and Structural Biotechnology Journal*, 21, 3187-3202.

[4] Brent M. Kuenzi, Jisoo Park, Samson H. Fong, Kyle S. Sanchez, John Lee, Jason F. Kreisberg, Jianzhu Ma, and Trey Ideker. (2020). "Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells", *Cell Reports*, 32(6), 108053.