**Personalized Medicine Using Genomics: A Machine Learning Approach to Drug Response Prediction**

**Group Members**

Wajeeh Ahmed 21L-7565

Laraib Akhtar 21L-5294

Ahmer Zaidi 21L-5284

**Abstract**

Personalized medicine leverages genetic data to optimize drug prescriptions for individual patients. This project aims to develop a machine-learning-based drug response prediction system using pharmacogenomics data. Traditional drug prescription methods are often generalized, leading to adverse drug reactions or inefficacy in some patients. By analyzing genetic markers, our model will predict whether a patient will respond positively or negatively to a specific drug.

The system will be a full-stack web application comprising a Next.js frontend, FastAPI backend, and a machine learning model trained on genomic datasets (e.g., GDSC, PharmGKB, or TCGA). The model will classify patients as "responders" or "non-responders" based on gene expression data. We will use supervised learning techniques like Random Forest, XGBoost, or deep learning-based models such as DNABERT for classification.

The expected outcome is an interactive platform where users (researchers, clinicians, or patients) can upload genetic data and receive personalized drug response predictions. This system has the potential to enhance drug efficacy, minimize side effects, and contribute to the growing field of precision medicine.

**Introduction & Literature Review**

Traditional drug prescription follows a "one-size-fits-all" approach, often neglecting individual genetic differences. This leads to **adverse drug reactions (ADRs)** and reduced treatment effectiveness. **Pharmacogenomics** is the study of how genetic variations influence an individual’s response to drugs. By analyzing gene expression data, researchers can determine which genetic markers are linked to positive or negative drug responses, paving the way for personalized medicine.

Several studies have demonstrated the potential of machine learning in pharmacogenomics. Research utilizing large-scale genomic databases such as **Genomics of Drug Sensitivity in Cancer (GDSC)**, **The Cancer Genome Atlas (TCGA)**, and **PharmGKB** has shown that predictive models can classify patients into responders and non-responders based on gene expression profiles. Machine learning approaches such as **Random Forest, Support Vector Machines (SVM), XGBoost, and deep learning models** have been employed to identify genetic biomarkers that correlate with drug efficacy.

One significant challenge in this domain is data preprocessing, as genomic data is often high-dimensional and contains noise. Feature selection methods, such as **Principal Component Analysis (PCA), Recursive Feature Elimination (RFE), and LASSO regression**, are used to extract the most relevant genetic features for model training. Another challenge is the interpretability of machine learning models in clinical settings, which can be addressed using techniques like **SHAP (Shapley Additive Explanations)** and **LIME (Local Interpretable Model-agnostic Explanations)** to understand how genetic factors influence drug response predictions.

By integrating these methodologies, our project aims to provide an accessible, AI-driven solution to assist researchers and clinicians in making informed drug prescription decisions based on individual genetic profiles.

**Proposed Methodology & Tools**

1. **Dataset Selection:** GDSC, PharmGKB, or TCGA datasets.
2. **Preprocessing:** Feature selection (SNPs, gene expression), normalization, handling missing data.
3. **Model Development:** Train a **Random Forest/XGBoost classifier** or a deep learning model (**DNABERT**) on labeled drug response data.
4. **Backend:** FastAPI-based API for ML inference.
5. **Frontend:** Next.js-based UI for user interaction (file upload, result display).
6. **Database:** MongoDB/PostgreSQL for storing past predictions and user data.

**Expected Outcomes & Deliverables**

* A machine-learning model capable of predicting drug response based on genetic markers.
* A functional **web-based application** allowing users to upload genetic data and receive predictions.
* A dataset preprocessing and feature extraction pipeline for genomic data.
* A comprehensive report detailing model performance, methodology, and results.

**References**

1. PharmGKB Database – https://www.pharmgkb.org/
2. Genomics of Drug Sensitivity in Cancer (GDSC) – https://www.cancerrxgene.org/
3. The Cancer Genome Atlas (TCGA) – https://portal.gdc.cancer.gov/
4. Liu, R., et al. "Machine learning approaches for drug response prediction." *Briefings in Bioinformatics*, 2021.