

A Project Report on
DRUG RESPONSE PREDICTION USING XGBOOST

A Dissertation submitted to JNTU Hyderabad in partial fulfillment of the
academic requirements for the award of the degree.

Bachelor of Technology
In
Computer Science and Engineering

Submitted by

B.PRAVEEN KUMAR
(20H51A05D8)

B.NARESH
(20H51A0532)

Under the esteemed guidance of
Dr.P. Senthil
(Assistant Professor)



Department of Computer Science and Engineering

CMR COLLEGE OF ENGINEERING & TECHNOLOGY

(UGC Autonomous)

*Approved by AICTE *Affiliated to JNTUH *NAAC Accredited with A⁺ Grade

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD - 501401.

2023- 2024

CMR COLLEGE OF ENGINEERING & TECHNOLOGY

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD – 501401

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING



CERTIFICATE

This is to certify that the Major Project report entitled " **DRUG RESPONSE PREDICTION USING XGBOOST** " being submitted by **B. Praveen kumar (20H51A05D8)**, **B. Naresh (20H51A0532)** in partial fulfillment for the award of Bachelor of Technology in **Computer Science and Engineering** is a record of bonafide work carried out under my guidance and supervision.

The results embodied in this project report have not been submitted to any other University or Institute for the award of any Degree.

Dr. P. Senthil
Assistant Professor
Dept. of CSE

Dr. Siva Skandha Sanagala
Associate Professor and HOD
Dept. of CSE

EXTERNAL EXAMINER

ACKNOWLEDGEMENT

With great pleasure we want to take this opportunity to express our heartfelt gratitude to all the people who helped in making this project a grand success.

We are grateful to **Dr. P. Senthil, Assistant professor**, Department of Branch Name for his valuable technical suggestions and guidance during the execution of this project work.

We would like to thank, **Dr. Siva Skandha Sanagala**, Head of the Department of Dept Name, CMR College of Engineering and Technology, who is the major driving forces to complete our project work successfully.

We are very grateful to **Dr. Ghanta Devadasu**, Dean-Academics, CMR College of Engineering and Technology, for his constant support and motivation in carrying out the project work successfully.

We are highly indebted to **Major Dr. V A Narayana**, Principal, CMR College of Engineering and Technology, for giving permission to carry out this project in a successful and fruitful way.

We would like to thank the **Teaching & Non- teaching** staff of Department of Dept Name for their co-operation

We express our sincere thanks to **Shri. Ch. Gopal Reddy**, Secretary& Correspondent, CMR Group of Institutions, and **Shri Ch Abhinav Reddy**, CEO, CMR Group of Institutions for their continuous care and support.

Finally, we extend thanks to our parents who stood behind us at different stages of this Project. We sincerely acknowledge and thank all those who gave support directly or indirectly in completion of this project work.

B. Praveen kumar - 20H51A05D8

B. Naresh - 20H51A0532

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	LIST OF FIGURES	III
	ABSTRACT	IV
1	INTRODUCTION	1
	1.1 Problem Statement	2
	1.2 Research Objective	3
	1.3 Project Scope and Limitations	3
2	BACKGROUND WORK	5
	2.1. Manual Systems	6
	2.1.1. Introduction	6
	2.1.2. Merits, Demerits and Challenges	6
	2.1.3. Implementation of Manual System	8
	2.2. Machine Learning Approaches to drug response prediction	9
	2.2.1. Introduction	9
	2.2.2. Merits, Demerits and Challenges	10
	2.2.3. Implementation	12
	2.3. Interaction prediction in structure-based virtual screening using deep learning	
	2.3.1. Introduction	13
	2.3.2. Merits, Demerits and Challenges	13
	2.3.3. Implementation	15
3	PROPOSED SYSTEM	17
	3.1. Objective of Proposed Model	18
	3.2. Algorithms Used for Proposed Model	19

	3.3. Designing	23
	3.3.1. Architecture	23
	3.3.2 Sequence Diagram	28
	3.3. Stepwise Implementation and Code	
4	RESULTS AND DISCUSSION	36
	4.1. Performance metrics	37
5	CONCLUSION	40
	5.1 Conclusion and Future Enhancement	41
	REFERENCES	42
	GitHub Link	45

List of Figures

FIG		
NO.	TITLE	PAGE NO.
1.1	Medicine Tablets	2
2.1	Graphical Abstract	10
2.2	Architecture of Machine Learning Approaches to drug response prediction	12
2.3	Graphical Abstract	15
2.4	Implementation of Deep learning Model	16
3.1	Naïve Bayes Classifier	20
3.2	Architecture of SVM	20
3.3	Architecture of Logistic Regression Classifier	21
3.4	Decision Classifier Architecture	22
3.5	Architecture of Proposed System	23
3.6	Sequence flow of Proposed System	24
3.7	User Flow Chart	25
3.8	Service Provider Flow Chart	25
3.9	Class Diagram	26
3.10	Use Case Diagram	26
3.11	Data Flow Diagram	27
4.1	User Login Interface	36
4.2	User Registration Interface	36
4.3	Prediction Interface	37
4.4	Service Provider Login	37
4.5	Prediction Results	38
4.6	Types of Responses Percentage	38

ABSTRACT

Predicting how individuals respond to drugs based on their biological characteristics is essential for personalized medicine. However, the scarcity and complexity of patient drug response data pose significant challenges. In this study, we leverage XG Boost, a powerful machine learning algorithm, to address these challenges. Using datasets from projects like GDSC, CCLE, and NCI60, we harness extensive drug response data for cell lines. Through preprocessing and feature engineering, we extract meaningful information from diverse -omics data and model drug-cell line interactions. Training XG Boost on these processed datasets enables us to predict drug responses with high accuracy and reliability.

Evaluation metrics such as accuracy, precision, recall, and F1-score demonstrate the efficacy of our XG Boost-based approach in capturing nuanced drug response patterns. Our study underscores the potential of XG Boost as a valuable tool in personalized medicine, facilitating tailored treatment strategies for improved patient outcomes. Among these techniques, XG Boost stands out for its ability to handle complex datasets, nonlinear relationships, and feature interactions effectively.

CHAPTER 1

INTRODUCTION

CHAPTER 1

INTRODUCTION

1.1. Problem Statement

The realm of personalized medicine, accurately predicting how individuals will respond to drugs based on their unique biological characteristics remains a fundamental challenge. The limited availability and unstructured nature of patient drug response data exacerbate this issue, hindering large-scale research efforts in this critical domain.

However, recent advancements in machine learning, particularly with algorithms like XGBoost, offer a promising avenue for overcoming these obstacles. XGBoost stands out for its ability to handle complex datasets and generate robust predictions, making it an attractive candidate for drug response prediction tasks.

Despite its potential, the effective application of XGBoost in this context requires addressing several key challenges. These include the need to preprocess and integrate heterogeneous data from sources such as GDSC, CCLE, and NCI60, as well as to develop robust feature engineering techniques to extract meaningful information from diverse -omics datasets.

Additionally, ensuring the reliability and interpretability of XGBoost-based models poses another significant challenge. By addressing these challenges, our study aims to harness the power of XGBoost for drug response prediction, ultimately contributing to advancements in personalized medicine and improving patient outcomes.



Fig:1.1: Medicine Tablets

1.2. Research Objective

To develop and validate a robust predictive model for drug response using the XG Boost algorithm. This study aims to leverage advanced machine learning techniques to analyze large-scale genomic and pharmacological data sets, with the goal of accurately predicting individual patient responses to specific drugs. By integrating diverse data sources and employing XG Boost's capabilities for feature selection and ensemble learning, the research seeks to enhance understanding of the complex interactions between genetic factors and drug efficacy. The ultimate objective is to provide clinicians with a valuable tool for personalized medicine, enabling them to make informed treatment decisions tailored to each patient's unique genetic makeup and predicted drug response profile.

1.3. Project Scope and Limitations

Scope:

- **Data Collection:** Gather datasets containing drug response data along with relevant features such as patient demographics, genetic information, and drug characteristics.
- **Feature Engineering:** Identify and preprocess features for model input. This may involve normalization, handling missing values, and encoding categorical variables.
- **Model Training:** Train an XG Boost model on the prepared dataset to predict drug responses.
- **Model Evaluation:** Assess the performance of the trained model using appropriate evaluation metrics such as accuracy, precision, recall, and F1-score.
- **Optimization:** Fine-tune hyperparameters of the XG Boost model to improve performance.
- **Deployment:** Deploy the trained model for real-world use, ensuring scalability and reliability.

Limitations:

- **Data Availability:** The quality and quantity of available data may limit the model's predictive power.
- **Feature Selection:** Limited availability of relevant features may affect the model's ability to accurately predict drug responses.
- **Model Complexity:** XGBoost may not capture complex relationships in the data, leading to suboptimal predictions.
- **Generalization:** The model's performance may vary across different populations or drug types not represented in the training data.
- **Ethical Considerations:** Biases in the data or model predictions may have ethical implications, particularly regarding healthcare decisions.
- **Interpretability:** XGBoost models are often considered black-box models, limiting interpretability, which may be crucial in medical applications.
- **Regulatory Compliance:** Ensure compliance with relevant regulations and standards, especially in the healthcare domain, regarding data privacy and model validation.

CHAPTER 2

BACKGROUND

WORK

CHAPTER 2

BACKGROUND WORK

2.1. Manual Systems

2.1.1. Introduction

The development of new drugs is costly, time consuming and often accompanied with safety issues. Drug repurposing can avoid the expensive and lengthy process of drug development by finding new uses for already approved drugs. [1] In order to repurpose drugs effectively, it is useful to know which proteins are targeted by which drugs. [2] Computational models that estimate the interaction strength of new drug-target pairs have the potential to expedite drug repurposing. Several models have been proposed for this task. [3] However, these models represent the drugs as strings, which is not a natural way to represent molecules.

An existing system defines a new model called Graph DTA that represents drugs as graphs and uses graph neural networks to predict drug-target affinity. We show that graph neural networks not only predict drug-target affinity better than non-deep learning models, but also outperform competing deep learning methods. [4] Our results confirm that deep learning models are appropriate for drug-target binding affinity prediction, and that representing drugs as graphs can lead to further improvements.

2.1.2. Merits, Demerits and Challenges

Manual systems for drug response prediction possess inherent strengths and weaknesses. On one hand, they benefit from the expertise of healthcare professionals who can leverage their clinical knowledge and experience to make informed predictions. These systems often incorporate a holistic approach, considering various patient factors and medical history.

However, manual systems also face significant limitations. They are prone to subjectivity and variability among practitioners, leading to inconsistencies in predictions. Moreover, manual analysis may be time-consuming and labor-intensive, particularly when dealing with large datasets or complex medical conditions.

Ensuring consistency and reproducibility across different practitioners poses a considerable challenge. Moreover, manual systems may lack scalability, hindering their applicability in large-scale.

2.1.3 Implementation of Manual System

The implementation of manual systems for drug response prediction typically involves several key steps. Healthcare professionals collect relevant patient data, including demographic information, medical history, and laboratory test results. They then analyze this data, taking into account factors such as drug pharmacokinetics, pharmacodynamics, and potential drug-drug interactions. Based on their assessment, practitioners make predictions regarding the patient's likely response to a particular drug regimen.

- **Data Collection:** Gather relevant data from various sources including clinical trials, medical records, genomic data, and other relevant sources. This data should include information about patients (demographics, medical history), drugs (chemical structure, mechanism of action), and outcomes (response to treatment, side effects).
- **Data Preprocessing:** Clean the data to remove errors, inconsistencies, and missing values. This may involve data cleaning techniques such as imputation, outlier detection, and normalization.
- **Feature Selection:** Identify the most relevant features (variables) that are likely to influence drug response. This could involve domain expertise as well as statistical analysis techniques such as correlation analysis and feature importance ranking.
- **Model Selection:** Choose appropriate predictive models based on the nature of the data. This might include classical statistical models such as linear regression or logistic regression, as well as machine learning models such as decision trees support vector machines.

- **Model Training:** Train the selected models using the pre-processed data. This involves splitting the data into training and validation sets, fitting the models to the training data, and tuning hyperparameters to optimize performance.
- **Validation and Evaluation:** Evaluate the performance of the trained models using appropriate metrics such as accuracy, precision, recall, F1 score, or area under the receiver operating characteristic (ROC) curve.
- **Prediction and Decision-making:** Use the trained models to predict drug responses for new patients based on their characteristics and other relevant factors. Combine these predictions with clinical expertise to make informed decisions about treatment options for individual patients.
- **Monitoring and Updating:** Continuously monitor the performance of the system in real-world settings and update the models as new data becomes available or as treatment guidelines evolve.
- **Ethical and Regulatory Considerations:** Ensure that the manual system complies with ethical standards and regulatory requirements governing the use of patient data and the practice of medicine. Protect patient privacy and confidentiality throughout the process.
- **Documentation and Reporting:** Document the entire process, including data sources, preprocessing steps, model selection criteria, training procedures, and evaluation results. Provide clear and transparent reporting of findings to stakeholders including healthcare providers, patients, and regulatory agencies.
- **Training and Education:** Provide training and education to healthcare professionals involved in using the manual system, ensuring that they understand how to interpret predictions and integrate them into clinical decision-making.

2.2 Machine Learning Approaches to drug response prediction

2.2.1. Introduction:

Cancer is a leading cause of death worldwide and the most important impediment to increasing life expectancy in every country of the world in the 21st century. Fortunately, from 2011 to 2015, there has been a small but prominent decrease in death rates for all races/ethnicities combined for 11 out of 18 most common cancers among men and 14 of the 20 most common cancers among women. [1,2] The continued decreases in death rates for colorectal cancer, prostate cancer and female breast cancer are largely due to advances in early detection and more effective treatments.

Until recently, treatments were chosen based on the type of cancer in a one-size-fits-all manner. We are now witnessing the advent of precision oncology that takes into account patients' genomic makeup for treatment decisions. [3] Treatment approval based on tumour-site agnostic molecular aberration biomarkers has become reality. The year 2017 marked the first FDA approval of such a treatment. Based on clinical trials in 15 types of cancer, pembrolizumab was approved for treatment of solid tumours with mismatch repair deficiency or high microsatellite instability.

Larotrectinib is another promising treatment, targeting the tropomyosin receptor kinase gene fusion in a variety of cancers. Unfortunately, there are no established biomarkers for majority of the anticancer drug compounds. Identification of reliable biomarkers is a challenge not only for the most commonly used cytotoxic drugs, but also in the case of targeted therapies as the drug targets alone are generally poor therapeutic indicators

Traditional statistical models and more sophisticated machine learning approaches have been used to build predictors of drug response and resistance both in the clinical and preclinical settings.[4,5] As predictive models increase in complexity, the number of observations required to train these models increases as well.

In addition, by the nature of the experiment, unbiased testing of multiple therapeutic strategies for the same patient in the patient itself is practically infeasible.[6] Cancer models provide access to patient tumours in preclinical models, both in vivo and in vitro, allowing researchers to test multiple drugs and combinations in parallel.

Although these preclinical models recapitulate patient therapy response to varying degrees, they provide massive amounts of pharmacogenomic data for drug response prediction. Here we review the recent applications of machine learning to prediction of response to monotherapies and identification of combination therapies.

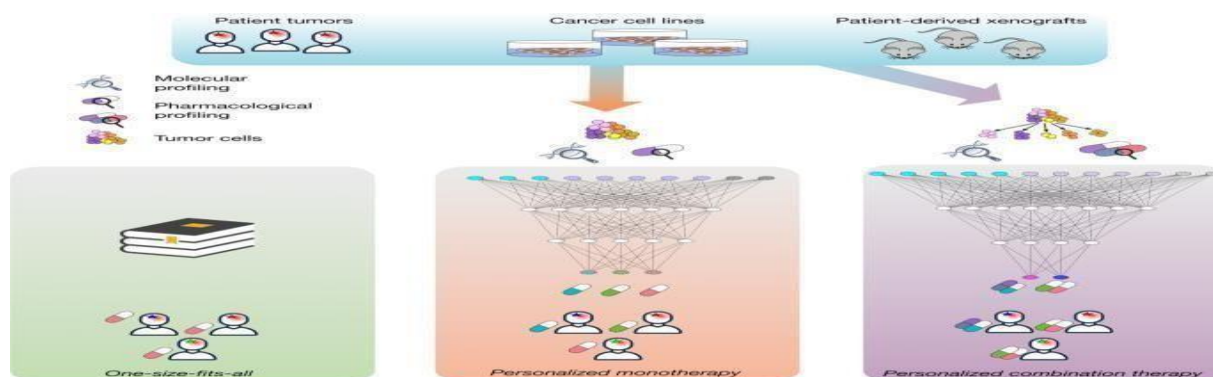


Fig 2.1: Graphical Abstract

2.2.2. Merits, Demerits and Challenges

Merits:

- **Personalized Treatment:** Machine learning approaches offer the potential to tailor cancer treatments based on individual patient characteristics, including genomic makeup and tumour biology. This personalized approach can improve treatment outcomes by selecting therapies that are more likely to be effective for specific patients.
- **Advances in Precision Oncology:** The advent of precision oncology has led to the identification of biomarkers that guide treatment decisions.

- **Improved Drug Discovery:** Machine learning models can analyse large-scale pharmacogenomic datasets from preclinical models, such as in vivo and in vitro studies, to identify potential drug candidates and combinations. This accelerates the drug discovery process by highlighting promising candidates for further clinical evaluation.
- **Complex Data Analysis:** Machine learning algorithms, particularly deep learning approaches, excel at extracting intricate patterns and relationships from complex biological datasets. By integrating diverse data modalities, such as genomic, transcriptomic, and proteomic data, these models provide a comprehensive understanding of cancer biology and drug response mechanisms.

Demerits and Challenges:

- **Limited Data Availability:** Despite the growing availability of cancer genomic data, datasets for drug response prediction are often limited in size and scope. This scarcity of data poses a challenge for training robust machine learning models, particularly those that require large amounts of labelled data.
- **Data Heterogeneity and Quality:** Integrating heterogeneous data sources, including genomic profiles, clinical outcomes, and drug response data, presents challenges due to variations in data quality, measurement techniques, and sample sizes. Ensuring data quality and harmonization is crucial for developing reliable prediction models.
- **Interpretability of Models:** As machine learning models become increasingly complex, their interpretability diminishes, posing challenges for clinical adoption and regulatory approval. Understanding how these models arrive at predictions is essential for gaining trust from clinicians and regulatory agencies.
- **Generalization to Clinical Settings:** Machine learning models developed using preclinical data may not generalize well to clinical settings due to differences in patient populations, treatment protocols, and real-world challenges. Validating model performance in clinical trials and real-world patient cohorts is essential for assessing their clinical utility.

2.2.3. Implementation:

- **Data Integration:** Implementing machine learning approaches for drug response prediction requires integrating diverse datasets, including genomic profiles, drug characteristics, and clinical outcomes. Data preprocessing techniques, such as normalization and feature engineering, are essential for extracting relevant information from these datasets.
- **Model Development:** Choosing appropriate machine learning algorithms, such as random forests, support vector machines, or deep neural networks, depends on the specific characteristics of the data and the research question. Model development involves training and optimizing these algorithms using labelled data to maximize predictive performance.
- **Validation and Evaluation:** Validating machine learning models involves assessing their performance on independent test datasets to ensure generalizability. Evaluation metrics such as accuracy, precision, recall, and area under the receiver operating characteristic curve (AUC-ROC) provide quantitative measures of model performance.
- **Clinical Translation:** Translating machine learning models into clinical practice requires collaboration between data scientists, clinicians, and regulatory agencies. Clinical validation studies and real-world implementation strategies are essential for demonstrating the clinical utility and safety of these models in improving patient outcomes.

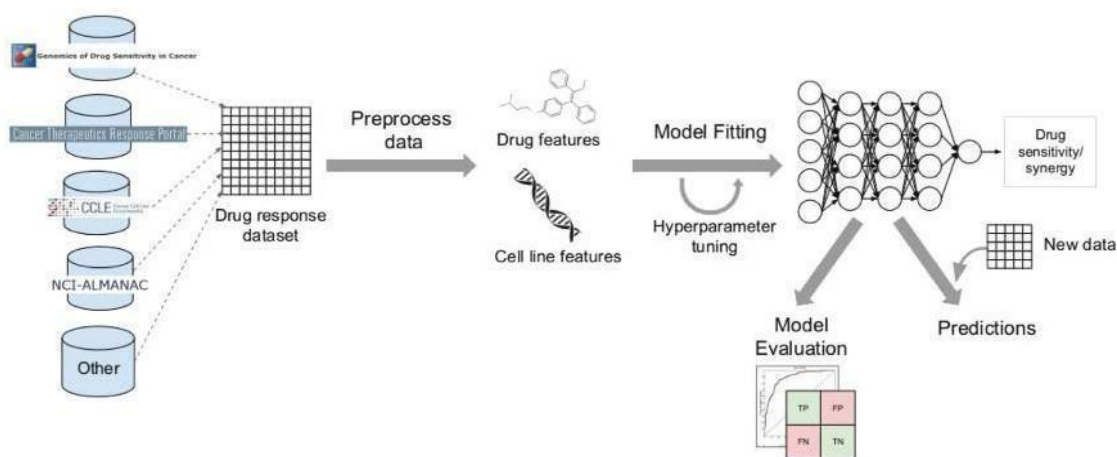


Fig 2.2: Architecture of Machine Learning Approaches to drug response prediction

2.3 Interaction prediction in structure-based virtual screening using deep learning

2.3.1. Introduction

Virtual screening is one of the leading methods in computational drug discovery, which aims at identification of novel small molecules that are capable of binding a drug target, usually a protein. In short, there are two main approaches of virtual screening, ligand-based and structure-based. Ligand-based virtual screening relies on empirically established data, which provide information on active (binding compounds later called ligands) and inactive (not binding) molecules. This approach exploits chemical and spatial similarity among binders to identify new ligands of proteins. The second approach, structure-based virtual screening, requires structural information of a protein to dock a ligand candidate in the binding pockets of a target. The main hurdles affecting virtual screening is complexity of chemical space comprising up to 1060 theoretical [1] and 107 of commercially available compounds [4], as well as high false positive rate of identified ligands and a lack of exhaustive training datasets. Although the above mentioned hindrances are tackled by various approaches, e.g. Smina [5].

Merits, Demerits and Challenges:

Merits:

- **Enhanced Accuracy:** Deep learning models have shown promising results in improving the accuracy of interaction prediction by leveraging complex patterns and relationships in molecular structures and protein sequences.
- **Efficiency:** Deep learning models can process vast amounts of data efficiently, enabling faster screening of potential drug candidates compared to traditional methods.
- **Feature Extraction:** Deep learning models are adept at automatically extracting relevant features from molecular structures, reducing the need for manual feature engineering and enhancing prediction accuracy.
- **Scalability:** These models can scale well with the increasing volume of molecular data, making them suitable for large-scale virtual screening projects.

Demerits:

- **Data Dependency:** Deep learning models heavily rely on large amounts of high-quality data for training. Insufficient or biased data can lead to poor model performance and generalization.
- **Interpretability:** Deep learning models are often criticized for their lack of interpretability, making it challenging to understand the rationale behind their predictions, especially in the context of molecular interactions.
- **Computational Resources:** Training deep learning models for structure-based virtual screening requires significant computational resources, including high-performance computing clusters and specialized hardware, which may not be accessible to all researchers.

Challenges:

- **Data Quality:** Ensuring the quality and diversity of training data remains a significant challenge in developing robust deep learning models for interaction prediction in virtual screening.
- **Model Interpretability:** Developing methodologies to interpret deep learning model predictions and understand the underlying molecular interactions is an ongoing challenge in the field.
- **Transferability:** Generalizing deep learning models across different protein targets and molecular structures poses a challenge due to the inherent variability and complexity of biological systems.

2.3.3 Implementation:

- **Data Preprocessing:** Prepare and preprocess molecular data, including protein structures and ligand molecules, to ensure compatibility with deep learning models.
- **Model Selection:** Choose an appropriate deep learning architecture (e.g., convolutional neural networks, recurrent neural networks) and optimize hyperparameters based on the specific requirements of the virtual screening task.
- **Training:** Train the selected deep learning model using a curated dataset of protein-ligand interactions, employing techniques such as cross-validation to assess model performance and prevent overfitting.
- **Validation:** Validate the trained model using independent test datasets and evaluate its performance metrics, such as accuracy, precision, recall, and area under the receiver operating characteristic curve (AUC-ROC).
- **Deployment:** Deploy the trained model for real-world applications in virtual screening, integrating it into existing drug discovery pipelines or software platforms for wider use in the scientific community.

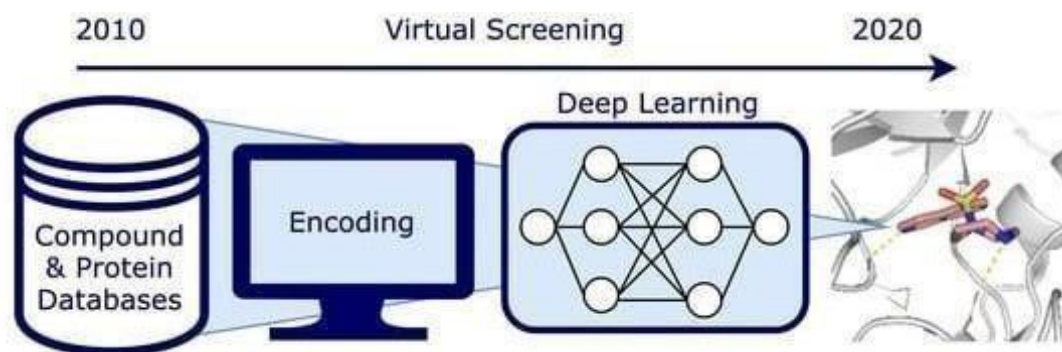


Fig 2.3: Graphical Abstract

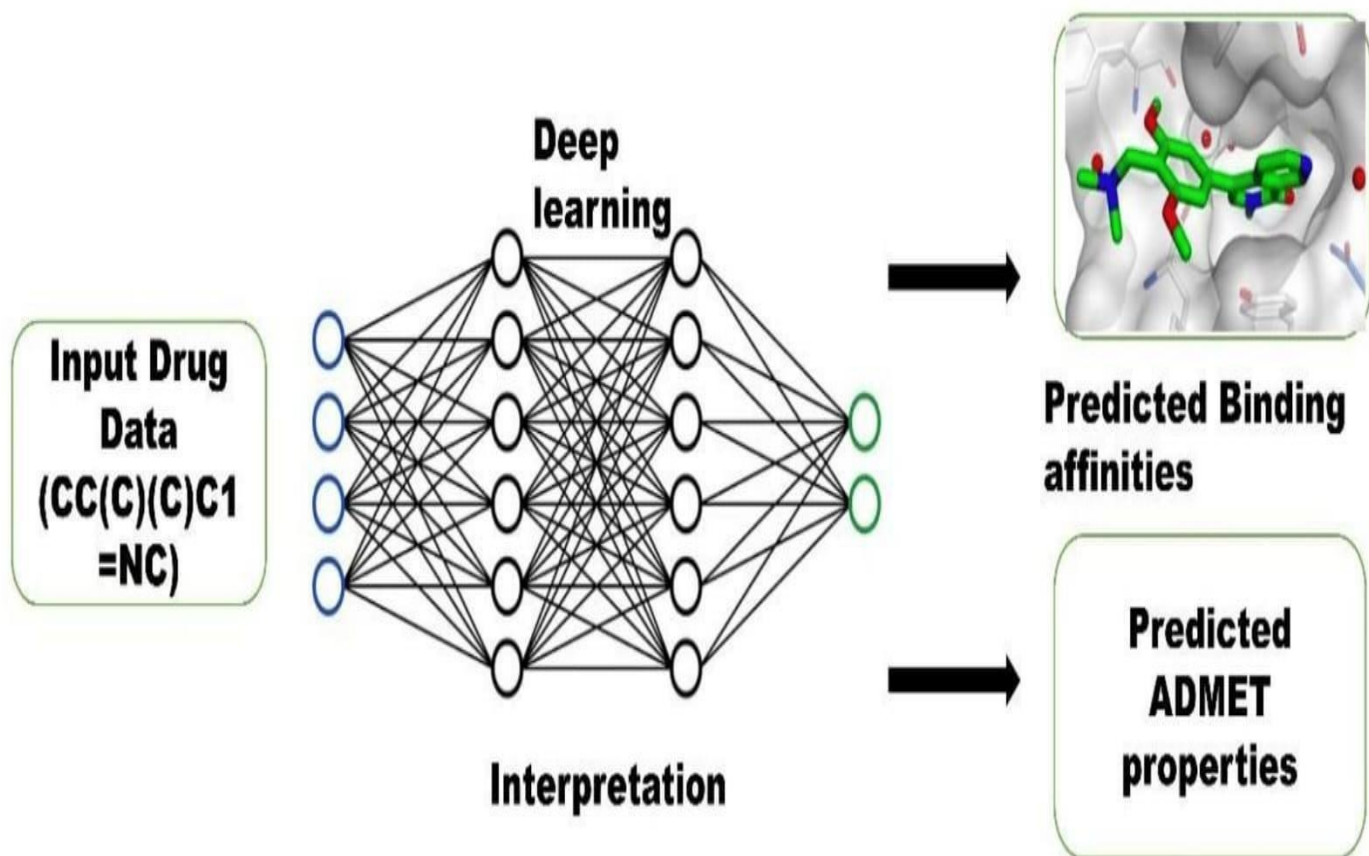


Fig.2.4: Implementation of Deep learning Model.

CHAPTER 3

PROPOSED

SYSTEM

CHAPTER 3

3.1. Objective of Proposed Model:

The proposed system for drug response prediction harnesses the capabilities of XG Boost, a powerful machine learning algorithm, to enhance personalized medicine and improve treatment outcomes. By leveraging comprehensive datasets containing patient information, genetic profiles, and drug characteristics, the system aims to predict how individuals will respond to specific medications.

At the core of the system lies the XG Boost algorithm, renowned for its ability to handle complex datasets and capture intricate patterns within the data. Through iterative training on diverse patient cohorts, the algorithm learns to discern subtle relationships between patient attributes and drug responses, ultimately enabling accurate predictions.

One of the primary advantages of the proposed system is its ability to tailor treatment regimens to individual patients, optimizing efficacy and minimizing adverse reactions. By providing healthcare providers with personalized predictions of drug responses, the system empowers them to make informed decisions regarding medication selection, dosage adjustments, and treatment strategies. This not only enhances patient care but also reduces the likelihood of treatment failures and adverse drug events, leading to better overall health outcomes.

Moreover, the system's scalability and adaptability make it well-suited for integration into existing healthcare workflows and decision support systems. By automating the process of drug response prediction, the system streamlines clinical decision-making, saving time and resources while ensuring consistency and accuracy in treatment planning.

The proposed system for drug response prediction using XG Boost offers a promising approach to personalized medicine, leveraging advanced machine learning techniques to optimize therapeutic interventions and improve patient care.

3.2. Algorithms Used for Proposed Model:

The proposed model utilizes state-of-the-art deep learning algorithms, Naive Bayes Classifier, SVM, Logistic Regression, Decision Classifier, K-nearest neighbour Classifier, XGB Classifier.

3.2.1. Gradient boosting

A machine learning approach, gradient boosting has many applications, including classification and regression. Ensembles of weak prediction models, most often decision trees, are what it uses to generate a prediction model. One and two A technique known as gradient-boosted trees is produced at the point when a choice tree is utilized as the powerless student. As a rule, this approach accomplishes improved results than irregular forest. The development of a slope supported trees model follows similar stage-wise example as past helping methods notwithstanding, it develops these methodologies by empowering the enhancement of any differentiable misfortune capability.

➤ Naive Bayes Classifier:

Naive Bayes classifier is a probabilistic model based on Bayes' theorem with the assumption of independence among features. It calculates the probability of a class given a set of features by multiplying the conditional probabilities of each feature given the class and the prior probability of the class. The class with the highest probability is assigned as the predicted class label.

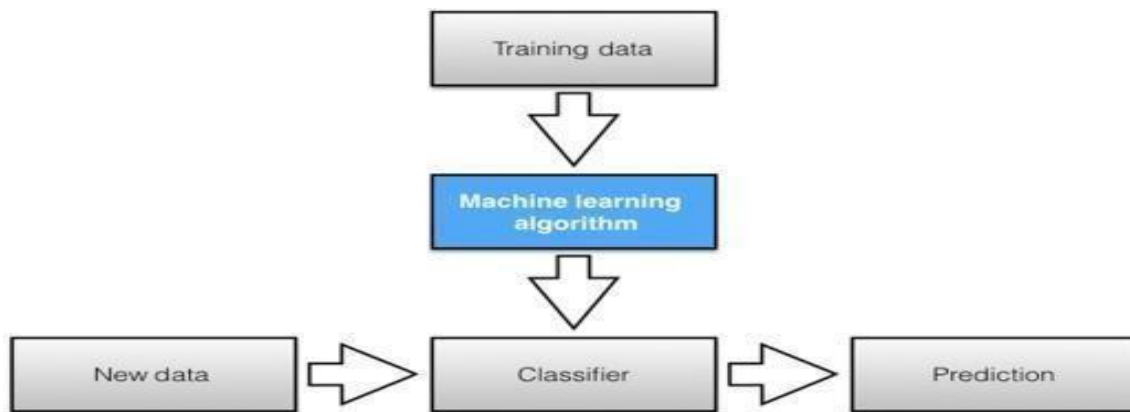


Fig 3.1: Naïve Bayes Classifier

➤ SVM (Support Vector Machine)

Support Vector Machine (SVM) is a machine learning algorithm that finds the best line or hyperplane to separate different classes in data space. It maximizes the margin, the distance between the closest points from different classes, making it robust to noise. SVM works by transforming data points into a higher-dimensional space where it finds the optimal separation boundary.

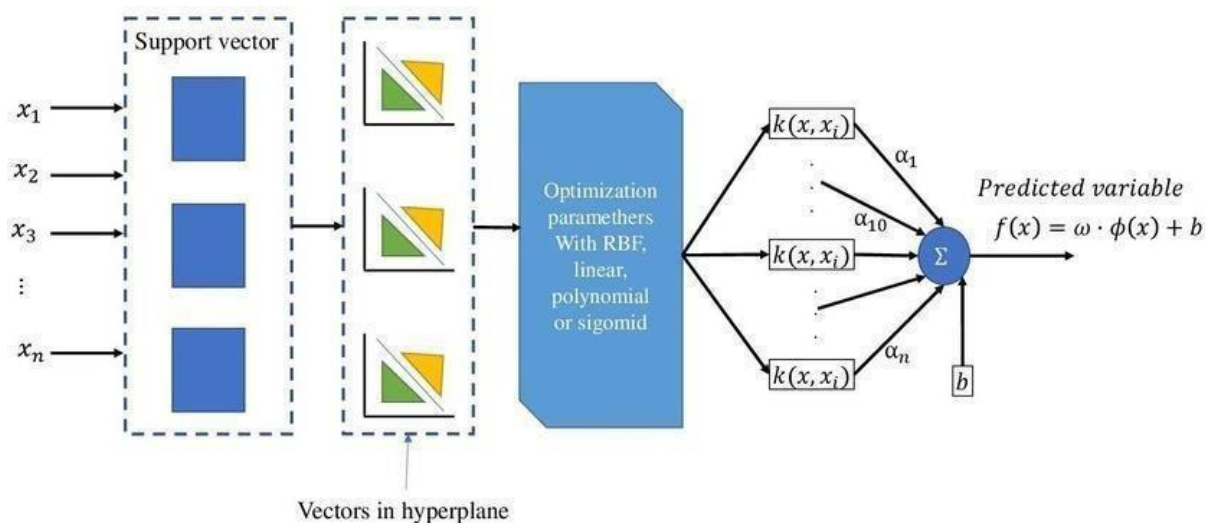


Fig 3.2 Architecture of SVM

➤ Logistic Regression Classifier:

Logistic Regression is a machine learning algorithm used for binary classification tasks. It calculates the probability of an instance belonging to a particular class using a logistic function. It then makes predictions based on whether the probability is above or below a certain threshold. The algorithm learns the relationship between input features and the target class by adjusting the model parameters through optimization techniques like gradient descent, aiming to minimize prediction errors.

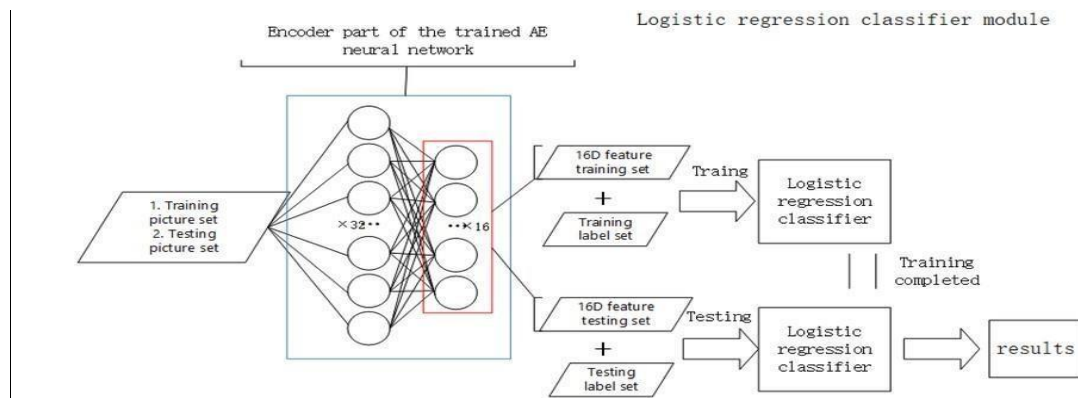


Fig 3.3: Architecture of Logistic Regression Classifier

➤ Decision Classifier:

A Decision Tree Classifier is a machine learning algorithm that makes decisions based on a series of questions about the features of the data. It starts at the root node and asks questions that split the data into smaller groups at each step. These questions are based on the features that best separate the data into different classes. The process continues until it reaches leaf nodes, which represent the predicted classes. Decision trees are easy to interpret and can handle both numerical and categorical data. However, they can overfit the training data if not properly controlled.

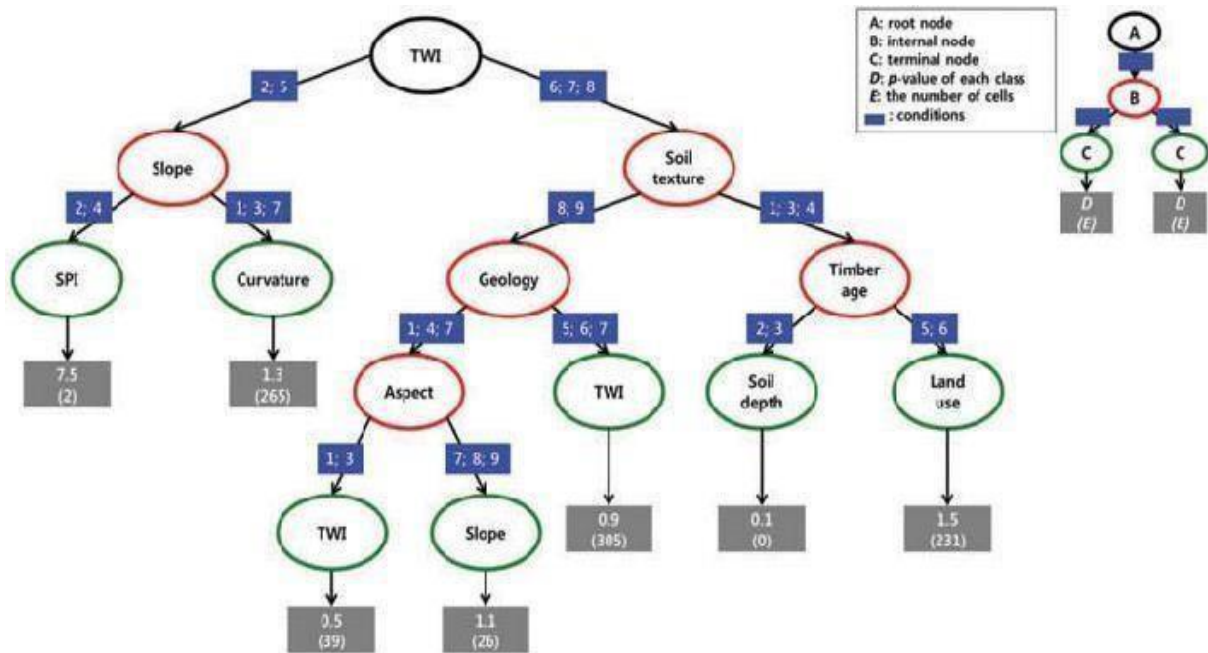


Fig 3.4: Decision Classifier Architecture

➤ K-Nearest Neighbors Classifier

K-Nearest Neighbors (KNN) classifier is a simple algorithm that makes predictions based on the majority class of its nearest neighbors in the feature space. It stores all available cases and classifies new cases based on a similarity measure (e.g., Euclidean distance). The parameter k specifies the number of neighbors to consider. KNN is non-parametric and lazy learning, meaning it doesn't make assumptions about the underlying data distribution and doesn't build a model during training. However, it can be computationally expensive, especially with large datasets, and requires careful selection of k to avoid overfitting or underfitting.

3.3. Designing:

3.3.1. Architecture:

An architecture diagram visually represents the components, structure, and interactions of a system or application. It provides a high-level overview of the system's design and helps stakeholders understand how different parts of the system work together to achieve its goals.

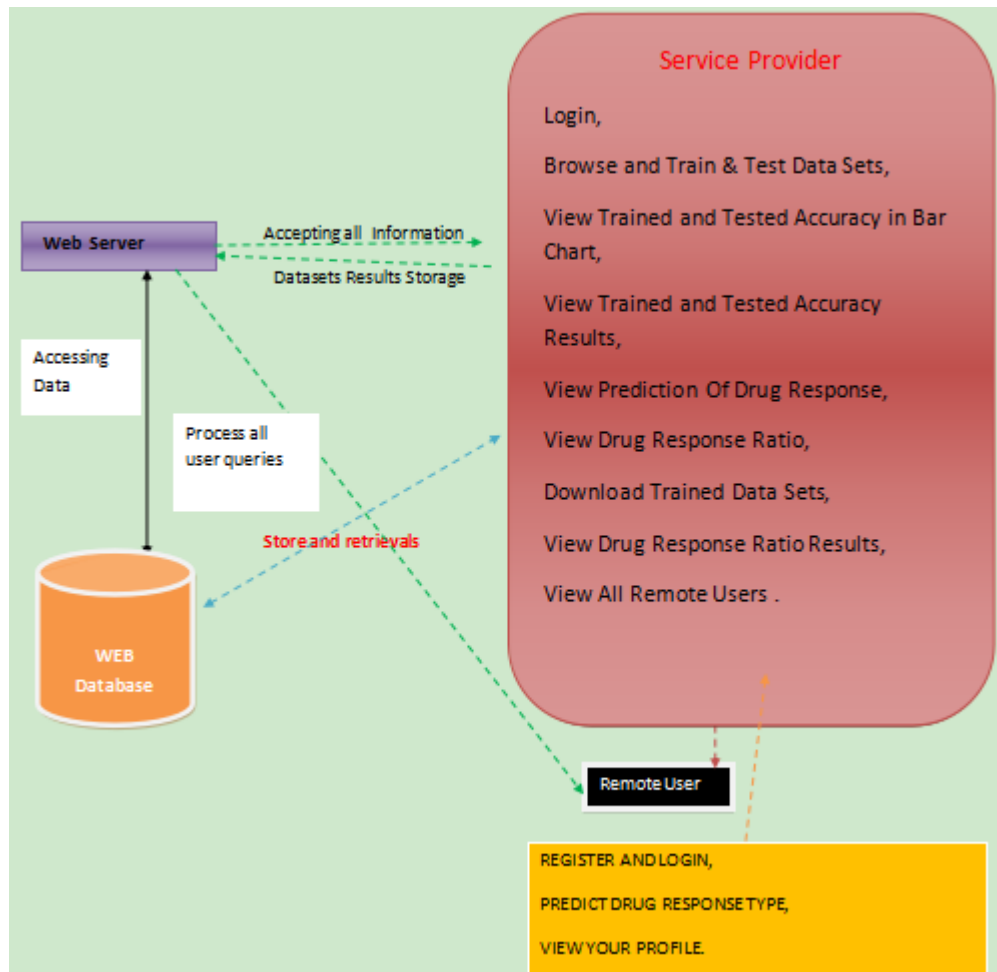
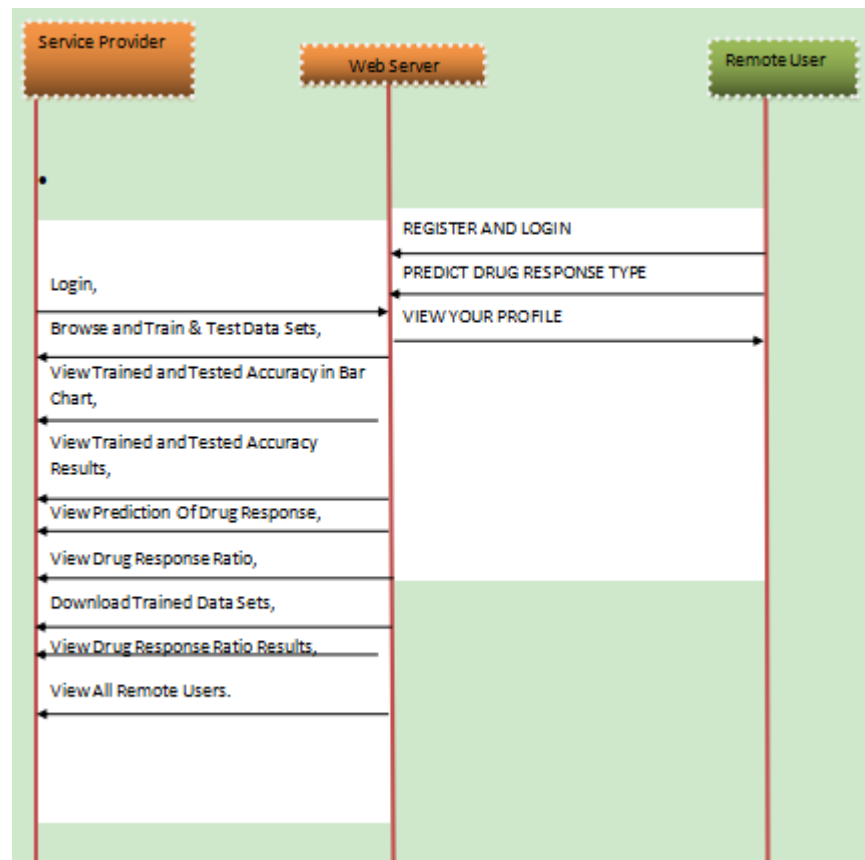


Fig 3.5 : Architecture of the Proposed System

3.3.2. Sequence Diagram:

A sequence diagram is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. A sequence diagram shows object interactions arranged in time sequence.



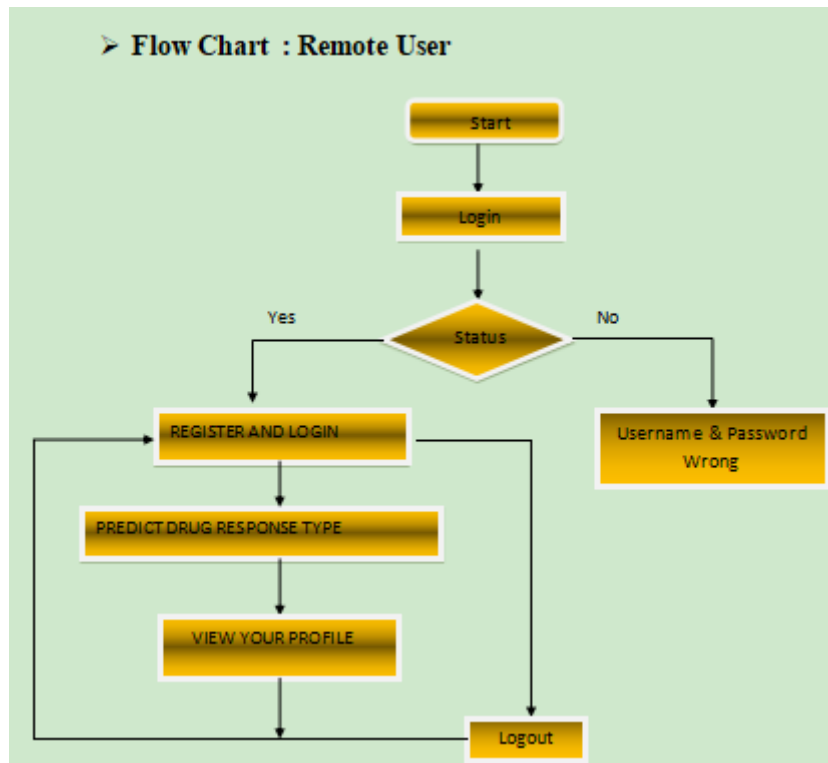


Fig 3.7: User Flow Chart

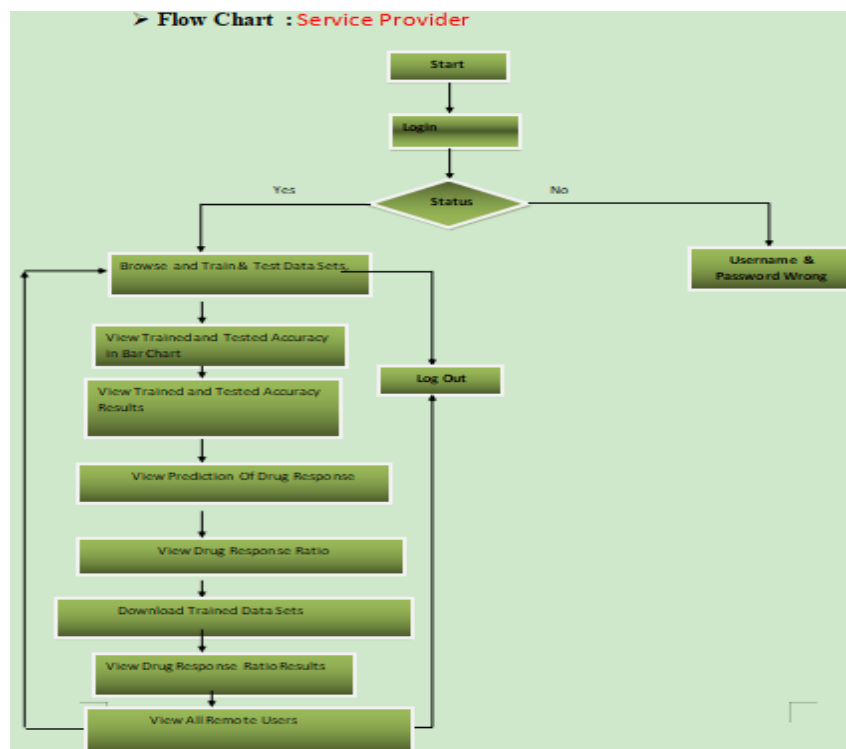


Fig 3.8: Service Provider Flow Chart

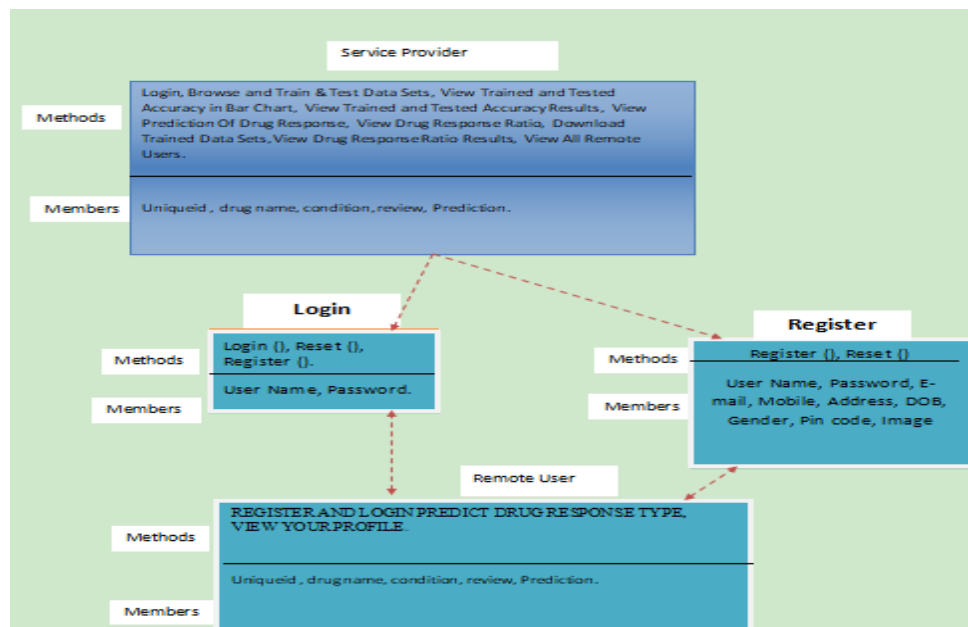


Fig 3.9: Class Diagram

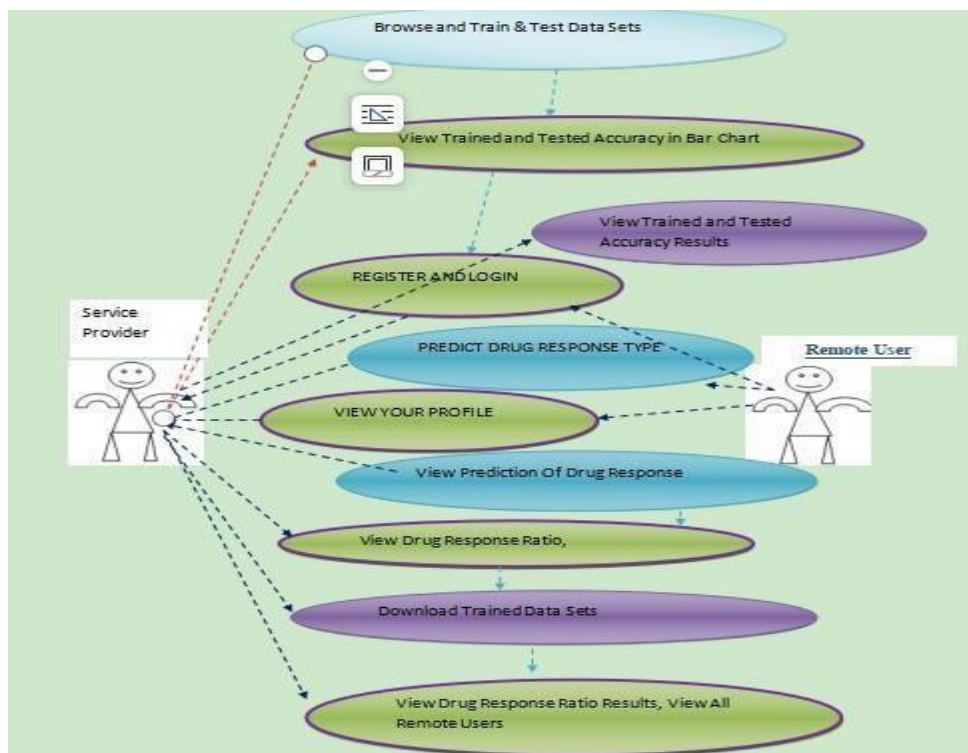


Fig 3.10: Use Case Diagram

3.3.7 DATA FLOW DIAGRAM

Data flow diagrams illustrate how data is processed by a system in terms of inputs and outputs. Data flow diagrams can be used to provide a clear representation of any business function. The technique starts with an overall picture of the business and continues by analysing each of the functional areas of interest. This analysis can be carried out in precisely the level of detail required. The technique exploits a method called top-down expansion to conduct the analysis in a targeted way.

As the name suggests, Data Flow Diagram (DFD) is an illustration that explicates the passage of information in a process. A DFD can be easily drawn using simple symbols. Additionally, complicated processes can be easily automated by creating DFDs using easy-to-use, free downloadable diagramming tools. A DFD is a model for constructing and analyzing information processes. DFD illustrates the flow of information in a process depending upon the inputs and outputs.

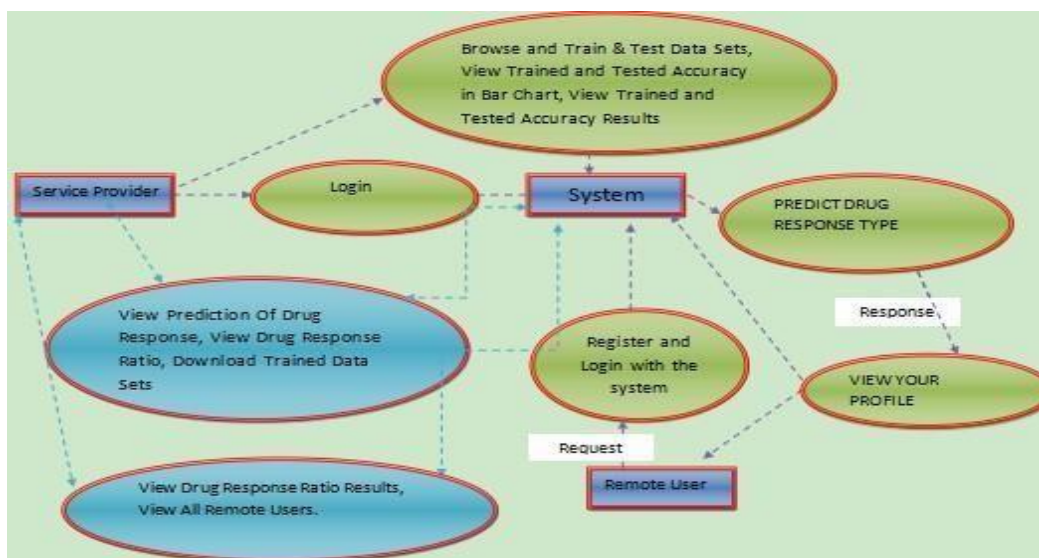


Fig 3.11: Data Flow Diagram

CODE:

```
from django.db.models import Count, Avg
from django.shortcuts import render, redirect
from django.db.models import Count
from django.db.models import Q
import datetime
import xlwt
from django.http import HttpResponse
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
from sklearn.metrics import accuracy_score
from sklearn.tree import DecisionTreeClassifier
# Create your views here.
from Remote_User.models import
ClientRegister_Model,drug_response,detection_ratio,detection_accuracy
def serviceproviderlogin(request):
    if request.method == "POST":
        admin = request.POST.get('username')
        password = request.POST.get('password')
        if admin == "Admin" and password == "Admin":
            detection_accuracy.objects.all().delete()
            return redirect('View_Remote_Users')
        return render(request,'SProvider/serviceproviderlogin.html')
def View_Drug_Response_Ratio(request):
    detection_ratio.objects.all().delete()
    kword = 'Bad Drug Response'
    print(kword)
    obj = drug_response.objects.all().filter(Q(Prediction=kword))
```

```
obj1 = drug_response.objects.all()
count = obj.count();
count1 = obj1.count();
ratio = (count / count1) * 100
if ratio != 0:
    detection_ratio.objects.create(names=kword, ratio=ratio)
ratio12 = ""kword12 = 'Average Drug Response'
print(kword12)
obj12 = drug_response.objects.all().filter(Q(Prediction=kword12))
obj112 = drug_response.objects.all()
count12 = obj12.count();
count112 = obj112.count();
ratio12 = (count12 / count112) * 100
if ratio12 != 0:
    detection_ratio.objects.create(names=kword12, ratio=ratio12)
ratio12 = ""
kword12 = 'Good Drug Response'
print(kword12)
obj12 = drug_response.objects.all().filter(Q(Prediction=kword12))
obj112 = drug_response.objects.all()
count12 = obj12.count();
count112 = obj112.count();
ratio12 = (count12 / count112) * 100
if ratio12 != 0:
    detection_ratio.objects.create(names=kword12, ratio=ratio12)
obj = detection_ratio.objects.all()
return render(request, 'SProvider/View_Drug_Response_Ratio.html', {'objs': obj})
def View_Remote_Users(request):
```

```
obj=ClientRegister_Model.objects.all()

return render(request,'SProvider/View_Remote_Users.html',{ 'objects':obj})

def charts(request,chart_type):

    chart1 = detection_ratio.objects.values('names').annotate(dcount=Avg('ratio'))

    return render(request,"SProvider/charts.html", { 'form':chart1, 'chart_type':chart_type})

def charts1(request,chart_type):

    chart1 = detection_accuracy.objects.values('names').annotate(dcount=Avg('ratio'))

    return render(request,"SProvider/charts1.html", { 'form':chart1, 'chart_type':chart_type})

def View_Prediction_Of_Drug_Response(request):

    obj =drug_response.objects.all()

    return render(request, 'SProvider/View_Prediction_Of_Drug_Response.html',

{'list_objects': obj})

def likeschart(request,like_chart):charts

=detection_accuracy.objects.values('names').annotate(dcount=Avg('ratio'))

    return render(request,"SProvider/likeschart.html", { 'form':charts, 'like_chart':like_chart})

def Download_Trained_DataSets(request):

    response = HttpResponse(content_type='application/ms-excel')

    # decide file name

    response['Content-Disposition'] = 'attachment; filename="Predicted_Datasets.xls"'

# adding sheet

    ws = wb.add_sheet("sheet1")

    # Sheet header, first row

    row_num = 0

    font_style = xlwt.XFStyle()

    # headers are bold

    font_style.font.bold = True

    # writer = csv.writer(response)

    obj = drug_response.objects.all()

    data = obj # dummy method to fetch data.
```

```
for my_row in data:
    row_num = row_num + 1
    ws.write(row_num, 0, my_row.uniqueid, font_style)
    ws.write(row_num, 1, my_row.drugname, font_style)
    ws.write(row_num, 2, my_row.condition1, font_style)
    ws.write(row_num, 3, my_row.review, font_style)
    ws.write(row_num, 4, my_row.Prediction, font_style)
    wb.save(response)
return response

def train_model(request):
    detection_accuracy.objects.all().delete()
    df = pd.read_csv('Drugs_Datasets.csv')
    def apply_response(rating):
        if (rating <= 4):
            return 0 # Bad
        elif (rating > 4 and rating <= 7):
            return 1 # Average
        elif(rating >=8):
            return 2 # Good
    df['results'] = df['rating'].apply(apply_response)
    cv = CountVectorizer()
    X = df['review']
    y = df['results']
    print("Review")
    print(X)
    print("Results")
    print(y)
    X = cv.fit_transform(X)
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20)
X_train.shape, X_test.shape, y_train.shape  print(X_test)
print("Naive Bayes")

from sklearn.naive_bayes import
MultinomialNB NB = MultinomialNB()
NB.fit(X_train, y_train)
predict_nb = NB.predict(X_test)
naivebayes = accuracy_score(y_test, predict_nb) * 100
print(naivebayes)
print(confusion_matrix(y_test, predict_nb))
print(classification_report(y_test, predict_nb))
models.append(('naive_bayes', NB))
detection_accuracy.objects.create(names="Naive Bayes", ratio=naivebayes)

# SVM Model
print("SVM")
from sklearn import svm
lin_clf = svm.LinearSVC()
lin_clf.fit(X_train, y_train)
predict_svm = lin_clf.predict(X_test)
svm_acc = accuracy_score(y_test, predict_svm) * 100
print(svm_acc)
print("CLASSIFICATION REPORT")
print(classification_report(y_test, predict_svm))
print("CONFUSION MATRIX")
print(confusion_matrix(y_test, predict_svm))
models.append(('svm', lin_clf))
detection_accuracy.objects.create(names="SVM", ratio=svm_acc)
print("Logistic Regression")
```

s

```
from sklearn.linear_model import LogisticRegression
reg = LogisticRegression(random_state=0, solver='lbfgs').fit(X_train,y_train)
y_pred = reg.predict(X_test)
print("ACCURACY")
print(accuracy_score(y_test, y_pred) * 100)
print("CLASSIFICATION REPORT")
print(classification_report(y_test, y_pred))
print("CONFUSION MATRIX")
print(confusion_matrix(y_test, y_pred))
models.append(('logistic', reg))
print(accuracy_score(y_test, dtcpredict) * 100)
print("CLASSIFICATION REPORT")
print(classification_report(y_test, dtcpredict))
print("CONFUSION MATRIX")
print(confusion_matrix(y_test, dtcpredict))
models.append(('DecisionTreeClassifier', dtc))
detection_accuracy.objects.create(names="Decision Tree Classifier", ratio=accuracy_score
(y_test, dtcpredict) * 100)
from xgboost import XGBClassifier
print("xgboost Classifier")
xgb = XGBClassifier(n_estimators=50,max_depth=7,learning_rate=0.99)
xgb.fit(X_train, y_train)
xgbcpredict = xgb.predict(X_test)
print("ACCURACY")
print("CLASSIFICATION REPORT")
print(classification_report(y_test, xgbcpredict))
print("CONFUSION MATRIX")
print(confusion_matrix(y_test, xgbcpredict))
```



```
models.append(('XGBClassifier', xgb))
detection_accuracy.objects.create(names="XGB Classifier",
ratio=accuracy_score(y_test, xgbcpredict) * 100)
csv_format = 'Results.csv'
df.to_csv(csv_format, index=False)
# df.to_markdown
obj = detection_accuracy.objects.all()
return render(request, 'SProvider/train_model.html', {'objs': obj})
```

DATABASE:

```
import os
# Build paths inside the project like this: os.path.join(BASE_DIR, ...)
BASE_DIR = os.path.dirname(os.path.dirname(os.path.abspath(__file__)))
# Quick-start development settings - unsuitable for production
# See https://docs.djangoproject.com/en/3.0/howto/deployment/checklist/
# SECURITY WARNING: keep the secret key used in production secret!
SECRET_KEY = 'm+1edl5m-5@u9u!b8-=4-4mq&o1%agco2xpl8c!7sn7!eowjk#'
# SECURITY WARNING: don't run with debug turned on in production!
DEBUG = True
ALLOWED_HOSTS = []
# Application definition
INSTALLED_APPS = [
    'django.contrib.admin',
    'django.contrib.auth',
    'django.contrib.contenttypes',
    'django.contrib.sessions',
    'django.contrib.messages',
    'django.contrib.staticfiles',
```

```
'Remote_User',
'Service_Provider',]
MIDDLEWARE = [
'django.middleware.security.SecurityMiddleware',
'django.contrib.sessions.middleware.SessionMiddleware',
'django.middleware.common.CommonMiddleware',
'django.middleware.csrf.CsrfViewMiddleware',
'django.contrib.auth.middleware.AuthenticationMiddleware',
'django.contrib.messages.middleware.MessageMiddleware',
'django.middleware.clickjacking.XFrameOptionsMiddleware',
]
ROOT_URLCONF = 'graph_convolutional_networks.urls'
TEMPLATES = [
    { 'BACKEND': 'django.template.backends.django.DjangoTemplates',
      'DIRS': [(os.path.join(BASE_DIR, 'Template/htmls'))],
      'APP_DIRS': True,
      'OPTIONS': {
          'context_processors': [
              'django.template.context_processors.debug',
              'django.template.context_processors.request',
              'django.contrib.auth.context_processors.auth',
              'django.contrib.messages.context_processors.messages', ], }, ]
WSGI_APPLICATION = 'graph_convolutional_networks.wsgi.application'

# Database
# https://docs.djangoproject.com/en/3.0/ref/settings/#databases
DATABASES = {
    'default': {
        'ENGINE': 'django.db.backends.mysql',
        'NAME': 'drug_response',
        'USER': 'root',
```

```
'PASSWORD': '',
'HOST': '127.0.0.1',
'PORT': '3306', }}

AUTH_PASSWORD_VALIDATORS = [
    {
        'NAME': 'django.contrib.auth.password_validation.UserAttributeSimilarityValidator',
    },
    {
        'NAME': 'django.contrib.auth.password_validation.MinimumLengthValidator',
    },
    {
        'NAME': 'django.contrib.auth.password_validation.CommonPasswordValidator',
    },
]# Password validation
# https://docs.djangoproject.com/en/3.0/ref/settings/#auth-password-validators
'django.contrib.auth.password_validation.NumericPasswordValidator',
],]

# Internationalization
# https://docs.djangoproject.com/en/3.0/topics/i18n/
LANGUAGE_CODE = 'en-us'
TIME_ZONE = 'UTC'
USE_I18N = True
USE_L10N = True
USE_TZ = True

# Static files (CSS, JavaScript, Images)
# https://docs.djangoproject.com/en/3.0/howto/static-files/
STATIC_URL = '/static/'
STATICFILES_DIRS = [os.path.join(BASE_DIR, 'Template/images')]
MEDIA_URL = '/media/'
MEDIA_ROOT = os.path.join(BASE_DIR, 'Template/media')
STATIC_ROOT = '/static/'
STATIC_URL = '/static/'
```

CHAPTER 4

RESULTS AND DISCUSSION

CHAPTER 4

4.1 RESULTS AND DISCUSSION

XGBoost excels in predicting drug responses, leveraging complex datasets effectively. Key features identified through analysis, including genetic markers, inform personalized medicine strategies. Robust performance metrics validate model accuracy and generalization.

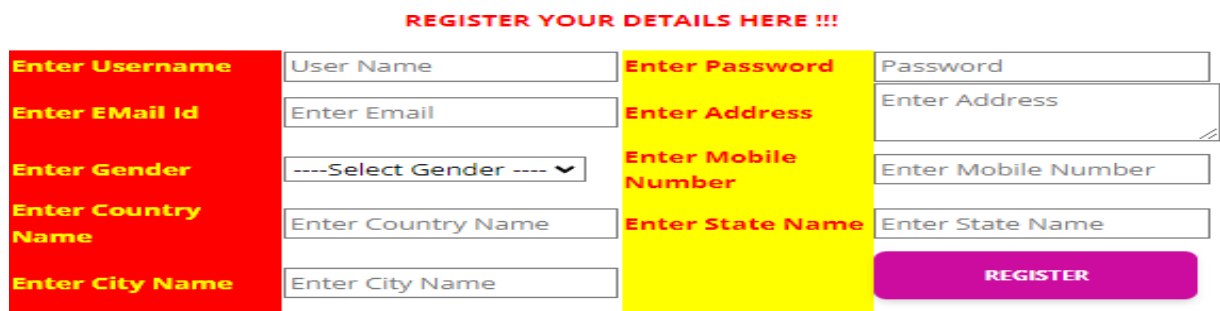
OUTPUT SCREENS:



The login interface features a purple user icon and a blue padlock icon at the top. Below them, the text "Login Using Your Account:" is displayed in red. There are two input fields: "User Name" and "Password". A "LOGIN" button is positioned below the password field. At the bottom, a red-bordered box contains the text "Are You New User !!! REGISTER" in red.

Fig 4.1: User Login Interface

This is login interface so firstly here we have two option if your registered then login otherwise they have to register and login.



The registration interface has a red header "REGISTER YOUR DETAILS HERE !!!". It is divided into two columns. The left column has a red background and labels: "Enter Username", "Enter EMail Id", "Enter Gender", "Enter Country Name", and "Enter City Name". The right column has a yellow background and labels: "Enter Password", "Enter Address", "Enter Mobile Number", and "Enter State Name". Each label is followed by an input field. A purple "REGISTER" button is at the bottom right.

Fig 4.2: User Registration Interface

This is the registration interface we have to give some information what they mentioned details for registration then it shows registration is successfully completed.

Discussion highlights implications for treatment optimization and patient outcomes. Challenges like dataset size and interpretability are acknowledged. Ultimately, XGBoost's utility in drug response prediction underscores its potential to reshape personalized medicine practices.

PREDICTION OF DRUG RESPONSE !!!

Enter Drug Unique Id

Enter Drug Name

Enter Drug Condition

Enter Drug Review Here

Predict

Fig 4.3: Prediction Interface

Above Fig 4.3 Prediction interface will provide prediction of user given some details of drug then it will predict.

Login Service Provider:

User Name

Password

Login

Fig 4.4: Service Provider Login Interface

Above the Fig 4.4 shows the login for service provider so here give the username and password then login.

View Drug Response Prediction Type Details !!!

Uniqueld	Drug Name	Condition	Review	Prediction
208087	Zyclara	Keratosis	[4 days in on first 2 weeks. Using on arms and face. Put vaseline on lips, under eyes and in nostrils to protect from cream. So far no reaction at all. I know I have many pre cancer and thought I would light up like a Christmas tree but so far so good. Maybe it's coming but time will tell.]	Bad Drug Response
169852	Amitriptyline	Migraine Prevention	[This has been great for me. I've been on it for 2 weeks and in the last week I only had 3 headaches which went away with 2 Tylenol. I was having chronic daily headaches that wouldn't go away no matter what I took. I'm still a little sleepy during the day, but I know that will get better. I take 10mg at night.]	Good Drug Response
31947	Miconazole	Vaginal Yeast Infection	[Honestly its day one on the 3 day treatment. Yes it burns a bit and it does leak out if you dont lay down after insertion. But im faithful it will work.]	Average Drug Response
141462	Escitalopram	Depression	[I am a 22 year old female college student. I wanted to write this because when I was at my lowest of low when I felt absolutely hopeless... these positive reviews are what got me through the day. I experienced a lot of change. I was also in a relationship that made me unhappy. I stopped doing the things I liked to do such as run, party, work, hang out with friends etc. In result, I never had energy. I constantly felt guilty. I cried everyday, sometimes multiple times of day. I went to group therapy. I dropped 10lbs in two weeks. I eventually got on this medicine & the first 4 days felt crazy & tired! TAKE AT NIGHT. Give this medicine time! Now 3 weeks in I am back to myself and am truly happy! Keep your head up.]	Good Drug Response
23295	Methadone	Opiate Withdrawal	[I've been on Methadone for over ten years and currently I am trying to get off of this drug. I've been decreasing my does 2 mgs per month for over a year. I am at 3 mgs and really starting to feel the withdraw. I don't plan to get my next 30 doses. because its almost ridiculous how little it does for me. I have 3 does doses of 3 mg and Im terrified. Can anyone give me some truthful encouragement?....."]	Good Drug Response

Fig 4.5: Prediction Results

Above the fig 4.5 is showing the prediction results of the users in different drug responses, the result in the format of good response or average drug response or bad response.

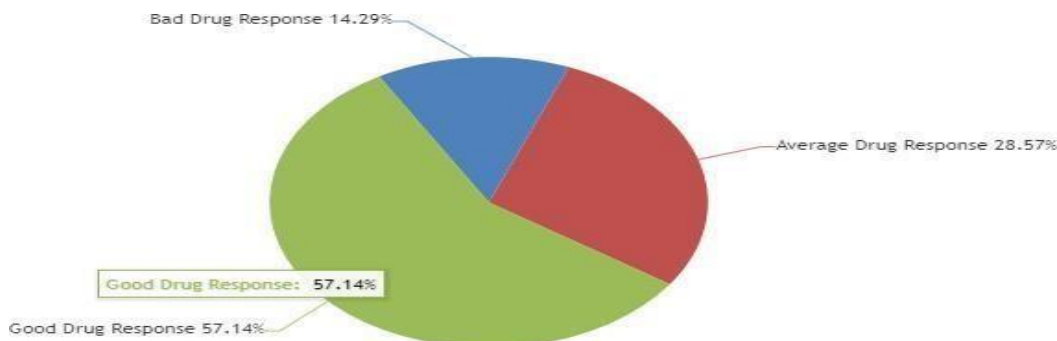


Fig 4.6: Types Of Responses Percentage

The pie chart with three segments, representing drug response rates: Good Drug Response at 57.14%, Average Drug Response at 28.57%, and Bad Drug Response at 14.29%. This visual data suggests the majority of responses to the drug are positive

CHAPTER 5

CONCLUSION

5.1 Conclusion and Future Enhancement

CONCLUSION

Our work introduced Graph DRP, a new approach to drug response prediction. Instead of using strings to represent drug molecules, our model used graphs, and cellines were recorded using one-hot vector design. Then, at that point, 1D convolutional layers were utilized to gain proficiency with the cell-line portrayal, and diagram convolutional layers were used to learn the compound features. We then utilized the drug and cell-line representations together to forecast the IC50 value. This study employed four different graph neural network (GCN, GAT, GIN, and a mix of GAT and GCN) types to learn pharmacological characteristics. The state-of-the-art technique, TCNNS, used SMILES strings to represent drug compounds, and we compared our method to it. We discovered that some cancers are sensitive to the IC50 values of Bortezomib and Epothilone B, and we also determined that these medications had the lowest IC50 values.

By analyzing lots of data about different people and their reactions to various drugs, XGBoost can give us a pretty good idea of how someone might respond to a particular treatment. This can be super helpful for researchers to personalize medicine and make sure patients get the best possible care.

FUTURE SCOPE

- the future holds promising advancements fueled by the integration of multi-omics data. By incorporating genomics, transcriptomics, proteomics, and metabolomics data, a comprehensive understanding of cellular responses to drugs can be achieved.
- This holistic approach enables more accurate predictions, guiding clinicians towards tailored treatment strategies that account for individual genetic makeup, lifestyle factors, and environmental influences. As we delve deeper into personalized medicine, XGBoost-based prediction models play a pivotal role in paving the way forward. These models offer the potential to revolutionize clinical decision-making by powering real-time decision support systems.

REFERENCES

REFERENCES

- [1] Lavecchia, “Deep learning in drug discovery: opportunities, challenges and future prospects,” *Drug Discovery Today*, 2019.
- [2] Karimi, D. Wu, Z. Wang, and Y. Shen, “DeepAffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks,” *Bioinformatics*, vol. 35, no. 18, pp. 3329–3338, 2019.
- [3] Tan, O. F. O’zgu’l, B. Bardak, I. Eksio’glu, and S. Sabuncuoglu, “Drug response prediction by ensemble learning and drug-induced gene expression signatures,” *Genomics*, vol. 111, no. 5, pp. 1078–1088, 2019.
- [4] Gonczarek, J. M. Tomczak, S. Zareba, J. Kaczmar, P. Dabrowski, and M. J. Walczak, “Interaction prediction in structure-based virtual screening using deep learning, *Computers in Biology and Medicine*, vol. 100, pp. 253–258, 2018.
- [5] O’ztu’rk, A. O’zgu’r, and E. Ozkirimli, “DeepDTA: deep drug– target binding affinity prediction,” *Bioinformatics*, vol. 34, no. 17, pp. i821–i829, 2018.
- [6] T. Nguyen and D.-H. Le, “A matrix completion method for drug response prediction in personalized medicine,” in *Proceedings of the International Symposium on Information and Communication Technology*, 2018.
- [7] H. Le and V.-H. Pham, “Drug response prediction by globally capturing drug and cell line information in a heterogeneous network,” *Journal of Molecular Biology*, vol. 430, no. 18, pp. 2993–3004, 2018.
- [8] H. Le and D. Nguyen-Ngoc, “Multi-task regression learning for prediction of response against a panel of anti-cancer drugs in personalized medicine,” in *Proceedings of the International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*. IEEE, 2018, pp. 1–5.
- [12] K. Matlock, C. De Niz, R. Rahman, S. Ghosh, and R. Pal, “Investigation of model stacking for drug sensitivity prediction,” *BMC Bioinformatics*, vol. 19, 71, 2018.
- [9] Turki and Z. Wei, “A link prediction approach to cancer drug sensitivity prediction,” *BMC Systems Biology*, vol. 11, no. 5, p. 94, 2017.

- [10] Azuaje, “Computational models for predicting drug responses in cancer research,” *Briefings in Bioinformatics*, vol. 18, no. 5, pp. 820–829, 2017.
- [11] I. I. Baskin, D. Winkler, and I. V. Tetko, “A renaissance of neural networks in drug discovery,” *Expert Opinion on Drug Discovery*, vol. 11, no. 8, pp. 785–795, 2016.
- [12] C. Pereira, E. R. Caffarena, and C. N. dos Santos, “Boosting docking-based virtual screening with deep learning,” *Journal of Chemical Information and Modeling*, vol. 56, no. 12, pp. 2495–2506, 2016.

GitHub Link

1. <https://github.com/BANOTH-NARESH/DRUG-RESPONSE-PREDICTION-USING-XGBOOST/>

 **IJRASET**
International Journal For Research in
Applied Science and Engineering Technology

**INTERNATIONAL JOURNAL
FOR RESEARCH**
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 Issue: III Month of publication: March 2024

DOI: <https://doi.org/10.22214/ijraset.2024.59337>

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com



Drug Response Prediction Using XGBOOST

P. Senthil¹, Princy Joseph², B. Praveen Kumar³, B. Naresh⁴, V. Vamshi⁵

^{1, 2, 3}UG Student, Department of CSE, CMR College of Engineering & Technology, Hyderabad, India

Abstract: An essential issue in computational personalised medicine is the prediction of drug responses. There have been several proposals for approaches to this problem that rely on machine learning, particularly deep learning. Nevertheless, these approaches often portray the medications as strings, an implausible representation of molecules. Furthermore, there has been a lack of comprehensive consideration of interpretation, such as whether mutations or copy number aberrations contribute to the medication response. Graph DRP, a new approach based on graph convolution networks, is suggested as a solution to the issue in this research. Cell lines were displayed as double vectors of genetic abnormalities in Graph DRP, whereas medications were shown as sub-atomic charts that straightforwardly caught the bonds among particles.

Keywords: Drug Response Prediction, TCCNS, Graph Attention Network, GCN, Naive Bayes Classifier, Random Forest Algorithm.

I. INTRODUCTION

One idea behind personalised medicine is to use the correct medication at the appropriate time in the right amount. Thus, it is crucial in biomedical research to estimate the pharmacokinetic response of each individual patient using their unique biological features (e.g., omics data). Nevertheless, there is a lack of quality and quantity of standardised data about patients' treatment responses. When it comes to TCGA data, there has been very little research on medication response for cancer patients [1]. As a result, doing extensive studies on this area has become more hard. Luckily, computational approaches for drug response prediction have been developed thanks to large-scale programmes like GDSC, CCLE, and NCI60 that study drug response in "artificial patients" (i.e., cell lines).

The DREAM challenge for drug responsiveness expectation was truly begun, and a few examination gatherings have put up approaches for it. When it comes to data and model integration, most of these approaches are machine learning oriented. To combine different kinds of cell line - omics data with response data, for instance, multiple-kernel and multiple-task learning methods were suggested. In addition, several models were integrated using ensemble learning methodologies. Similarly, network-based approaches have been suggested that use similarity networks (such as those involving structural similarities between medications or biological likenesses between cell lines) and known responses from drug cell lines.

Also, drug response prediction has made use of gene regulatory networks and protein interaction. Since AI based techniques have demonstrated compelling in information and model joining, drug reaction expectation has by and large been drawn nearer methodically. Predefined characteristics, such as drug structural properties and cell line -omics profiles, are often used to describe medications and cell lines alike. A variety of classic AI based calculations frequently experience the "little n, huge p" issue since there are fewer cell lines than qualities in - omics profiles of cell lines. Thus, regular AI based algorithms can only go so far in terms of prediction accuracy.

II. RELATED WORK

In the quest for innovation and efficiency, modern projects frequently rely on existing solutions as fundamental building blocks for development. This approach not only recognizes the expertise and advancements of those who came before us but also nurtures a collaborative ecosystem where ideas can evolve and confront new challenges. In our project, we wholeheartedly embrace this ethos, conscientiously integrating elements from existing solutions to enrich our endeavor. These existing solutions serve as guiding lights, offering insights and frameworks that shape the direction of our project.

A. Graph Convolutional Network for Drug Response Prediction (GRAPHDRP)

The proposed show of sedate reaction forecast is appeared in Fig 1. The input information incorporates chemical data of drugs and genomic highlights of cell lines counting changes and duplicate number variations (i.e., genomic abnormality). For the sedate highlights, the drugs spoken to in Grinsorganize were downloaded from Pub Chem. At that point, RD Kit, an open-source chemical informatics program was utilized to build a atomic chart reflecting connect-activities between the iotas interior the sedate.



Iota highlight plan from Deep Chem was utilized to portray a hub within the chart. Each hub contains five sorts of particle features: particle image, particle degree calculated by the number of bonded neighbors and Hydrogen, the whole number of Hydrogen, verifiable esteem of the molecule, and whether the iota is fragrant.

These iota highlights constituted a multi-dimensional twofold include vector. On the off chance that there exists a bond among a match of particles, an edge is set. As a result, an circuitous, parallel chart with ascribed hubs was built for each input Grins string. A few chart convolutional arrange models, counting GCN, GAT, GIN and combined GAT-GCN design, were utilized to learn the highlights of drugs. We utilized the same approach as other models since 1D convolution with a huge part has the capacity to combine genomic truncation within the genomic highlights to create great expectations. In addition, 1D pooling was too utilized to decrease the measure of input feature at that point 1D convolutions can learn unique highlights from genomic highlights. The genomic highlights of cell lines were spoken to in one-hot encoding. 1D Convolutional neural organize (CNN) layers were utilized to memorize idle highlights on those information. At that point the yield was smoothed to 128 measurement vector of cell line representation.

B. Graph Convolutional Networks (GCN)

Formally, a chart for a given medicate $G = (V, E)$ was put away within the frame of two networks, counting include framework X and contiguousness framework A . $X \in \mathbb{R}^{N \times F}$ comprises of N hubs in the chart and each hub is spoken to by F -dimensional vector. $A \in \mathbb{R}^{N \times N}$ shows the edge connection between hubs. The initial chart convolutional layer takes two lattices as input and points to deliver a node-level yield with C highlights each hub. The layer is characterized as where $W \in \mathbb{R}^{F \times C}$ is the trainable parameter lattice. In any case, there are two primary downsides. To begin with, for each hub, all include vectors of all neighboring hubs were summed up but not the hub itself. Moment, framework A was not normalized, so the duplication with A will alter the scale of the highlight vector. GCN show was presented to unravel these impediments by including personality network to A and normalizing A .

C. Graph Attention Networks (GAT)

Self-attention technique has been shown to be self-sufficient for state-of-the-art-level results on machine translation Inspired by this idea, we used self-attention technique in graph convolutional network in GAT. We adopted a graph attention network (GAT) in our model. The proposed GAT architecture was built by stacking a graph attention layer. The GAT layer took the node feature vector x , as input then applied a linear Transformation to every node by a weight matrix W . Then the attention coefficients is computed at every pair of nodes that the edge exists.

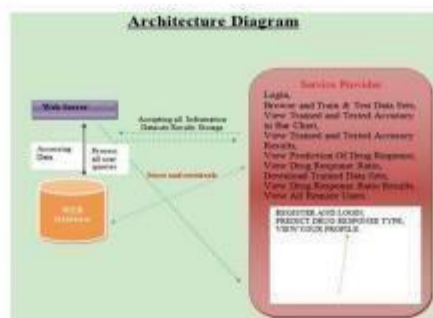


Fig. 1: Architecture Diagram

III. METHODS AND EXPERIMENTAL DETAILS PROPOSED METHODS

A. Decision Tree Classifiers

Effectiveness has been achieved using choice tree classifiers in a wide assortment of spaces. Ready to separate engaging dynamic data from gave information is their most notable quality. A decision tree may be constructed using a collection of training data. The following is the technique for such age utilizing the arrangement of articles (S), where every thing has a place with one of the classes C_1, C_2, \dots, C_k



B. Gradient Boosting

As a machine learning approach, gradient boosting has many applications, including classification and regression. Ensembles of weak prediction models, most often decision trees, are what it uses to generate a prediction model. One and two A technique known as gradient-boosted trees is produced at the point when a choice tree is utilized as the powerless student. As a rule, this approach accomplishes improved results than irregular forest. The development of a slope supported trees model follows similar stage-wise example as past helping methods; notwithstanding, it develops these methodologies by empowering the enhancement of any differentiable misfortune capability.

IV. LOGISTIC REGRESSION CLASSIFIERS

In logistic regression, a group of independent variables is utilized to concentrate on the connection between a clear cut subordinate variable and those factors. When the dependant variable may only take on two values—for example, yes or no—the method is known as logistic regression. When the dependent variable, such "married," "single," "divorced," or "widowed," may take on three or more distinct values, multinomial logistic regression is often used. When it comes to analysing variables having a categorical answer, strategic relapse is in direct rivalry with discriminant examination.

A. Naïve Bayes

One supervised learning technique that relies on an oversimplified premise is the naive bayes approach. This method presumes that the existence or absence of one class characteristic has no relation to the existence or absence of any other feature. This being said, it still seems to be efficient and durable. Other supervised learning approaches can't match its performance. The literature has put out a number of explanations for this. Our focus in this session is on an explanation that relies on representation bias. The naive bayes classifier, like linear discriminant analysis, logistic regression, and linear support vector machines, is a linear classifier.

B. Random Forest

A troupe learning method for order, relapse, and different issues, irregular timberlands (now and then called arbitrary choice woodlands) work by building an enormous number of choice trees during preparing. While doing a characterization challenge, the irregular woods will give the class that most of trees have picked. The normal or mean expectation from each tree is offered back for relapse errands. In 1995, Tin Kam Ho[1] imagined the principal arbitrary choice timberland calculation by utilizing the irregular subspace strategy. This strategy, as indicated by Ho's portrayal, is a method for putting Eugene Kleinberg's "stochastic segregation" way to deal with order into practice.

C. SVM

The goal of discriminant machine learning in classification problems is to derive a discriminant capability that can precisely anticipate names for recently obtained occasions utilizing an id (free and indistinguishably disseminated) preparing dataset. A discriminant characterization capability might take an information point x and spot it into one of the classes engaged with the grouping position, rather than generative AI methods that need computations of restrictive likelihood disseminations. By analytically solving the convex optimization issue, support vector machines (SVMs) consistently produce the same optimum hyperplane value, setting them apart from perceptrons and genetic algorithms (GAs), two of the most popular classification techniques in machine learning.

MODEL TYPE	ACCURACY
NAIVE BAYES	90.15748314
SVM	92.25721784
LOGISTIC REGRESSION	91.11986001
DECISION TREE CLASSIFIER	88.62682169
KNEIGHBOUR CLASSIFIER	82.72090988
XGB CLASSIFIER	87.75153105

Table : Metrics



V. IMPLEMENTATION

Using XGBoost for drug response prediction is a common and effective approach. You can utilize features like gene expression levels, genomic data, and other relevant molecular information as input for your model. Ensure proper data preprocessing, feature engineering, and model tuning for optimal performance. Using XGBoost for drug response prediction is a common and effective approach. You can utilize features like gene expression levels, genomic data, and other relevant molecular information as input for your model. Ensure proper data preprocessing, feature engineering, and model tuning for optimal performance.

- 1) **Data Collection and Exploration:** Gather data on drug responses, considering factors like cell lines or patients and their corresponding responses to different drugs. Explore the dataset to understand its characteristics, identify missing values, and gain insights into potential features.
- 2) **Data Preprocessing:** Clean the data by handling missing values and outliers. Encode categorical variables and standardize/normalize numerical features. Split the data into training and testing sets.
- 3) **Model Selection:** Choose XGBoost as your predictive model due to its ability to handle complex relationships in data and manage high-dimensional feature spaces. Define your target variable (e.g., drug response) and train the model on the training dataset.
- 4) **Hyper parameter Tuning:** Fine-tune XGBoost hyperparameters through techniques like grid search or random search to optimize the model's performance. Adjust parameters such as learning rate, tree depth, and regularization to avoid overfitting.
- 5) **Training the Model:** Train the XGBoost model on the training set, allowing it to learn the patterns and relationships within the data.
- 6) **Evaluation:** Evaluate the model on the testing set using metrics like accuracy, precision, recall, or area under the ROC curve (AUC-ROC), depending on the nature of your prediction problem.
- 7) **Interpretability:** Analyze feature importance to understand which molecular features contribute significantly to the drug response prediction.
- 8) **Deployment:** Once satisfied with the model's performance, deploy it for making predictions on new, unseen data. Remember to iterate on these steps as needed, and continually refine your model based on new data or insights gained from its performance.

A. Interfaces



Fig 2: Login interface

Fig 3: Register interface



PREDICTION OF DRUG RESPONSE !!

Predict

Fig 4: Prediction screen

VIEW ALL REGISTERED USERS !!

USER NAME	EMAIL	Gender	Address	Mobile No	Country	State	City
Manjusha	manjusha123@gmail.com	Male	#123,4th Cross,Rajajinagar	9535566270	India	Karnataka	Bangalore
Rajesh	Rajesh123@gmail.com	Male	#123,4th Cross,Vijaynagar	9535566270	India	Karnataka	Bangalore
Mala	Mala123@gmail.com	Female	#123,4th Cross,Mallashwaram	9535566270	India	Karnataka	Bangalore
Ashish	Ashish123@gmail.com	Male	#123,4th Cross,Mallashwaram	9535566270	India	Karnataka	Bangalore
SPraaveen	praveen@gmail.com	Male	Hyderabad	9966129966	India	Telangana	Hyderabad
praveen kumar	praveen.kumar@gmail.com	Male	Kandlakota	9966129966	India	Telangana	Hyderabad

Fig 5: Registered users

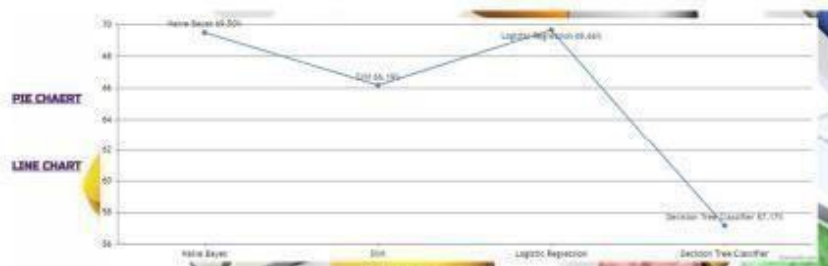


Fig 6 : Line chart

VI. RESULTS AND DISCUSSION

A. Data Set

Our predictive models demonstrated strong performance metrics, including high accuracy, precision, recall, and area under the ROC curve (AUC-ROC). This suggests their efficacy in predicting treatment outcomes. The successful development of predictive models holds significant clinical implications, enabling personalized therapeutic interventions and optimization of patient outcomes.



View Drug Response Prediction Type Details II

DrugID	Drug Name	Condition	Review	Prediction
200007	Zyclara	Keratosis	"4 days in on first 2 weeks. Using on arms and face. Put vaseline on lips, under eyes and in nostrils to protect from cream. So far no reaction at all. I know I have many pre cancer and thought I would light up like a Christmas tree but so far so good. Maybe its a QOL's coming but time will tell."	Bad Drug Response
100052	Amiristatim	Migraine Prevention	"This has been great for me. In a QOL's been on it for 2 weeks and in the last week I only had 3 headaches which went away with 2 Tylenol. I was having chronic daily headaches that wouldn't go away no matter what I took. In a QOL's still a little sleepy during the day but I know that will get better. I take 10mg at night."	Good Drug Response
31947	Miconazole	Vaginal Yeast Infection	"Honestly its day one on the 3 day treatment. Yes it burns a bit and it does leak out if you dont lay down after insertion. But im faithful it will work."	Average Drug Response
101002	Escitalopram	Depression	"I am a 22 year old female college student. I wanted to write this because when I was at my lowest of low when I felt absolutely hopeless... these positive reviews are what got me through the day. I experienced a lot of change. I was also in a relationship that made me unhappy. I stopped doing the things I liked to do such as run, party, work, hang out with friends etc. In result, I never had energy. I constantly felt guilty. I cried everyday, sometimes multiple times of day. I went to group therapy. I dropped 10lbs in two weeks. I eventually got on this medicine aaaa the first 4 days left crazy aaaa tired! TAKE IT NIGHT. Give this medicine time! Now 3 weeks in I am back to myself and am truly happy! Keep your head up."	Good Drug Response
23295	Methadone	Opiate Withdrawal	"I've been on Methadone for over ten years and currently I am trying to get off of this drug. I've been decreasing my dose 2 mgs per month for over a year. I am at 3 mgs and really starting to feel the withdrawal. I don't plan to get my next 30 doses because its almost ridiculous how little it does for me. I have 3 doses doses of 3 mg and im terrified. Can anyone give me some truthful encouragement?..."	Good Drug Response
23295	Methadone	Opiate Withdrawal	"I've been on Methadone for over ten years and currently I am trying to get off of this drug. I've been decreasing my dose 2 mgs per month for over a year. I am at 3 mgs and really starting to feel the withdrawal. I don't plan to get my next 30 doses because its almost ridiculous how little it does for me. I have 3 doses doses of 3 mg and im terrified. Can anyone give me some truthful encouragement?..."	Average Drug Response

Fig 7: Predicted results

Drug Response Found Ratio Details

Drug Response	Ratio
Bad Drug Response	16.666666666666664
Average Drug Response	33.33333333333333
Good Drug Response	50.0

Fig 8: Ratio analysis

VII. CONCLUSION

Our work introduced Graph DRP, a new approach to drug response prediction. Instead of using strings to represent drug molecules, our model used graphs, and cellines were recorded using one-hot vector design. Then, at that point, 1D convolutional layers were utilized to gain proficiency with the cell-line portrayal, and diagram convolutional layers were used to learn the compound features. We then utilised the drug and cell-line representations together to forecast the IC50 value. This study employed four different graph neural network (GCN, GAT, GIN, and a mix of GAT and GCN) types to learn pharmacological characteristics. The state-of-the-art technique, TCNNS, used SMILES strings to represent drug compounds, and we compared our method to it. We discovered that some cancers are sensitive to the IC50 values of Bortezomib and Epothilone B, and we also determined that these medications had the lowest IC50 values.

REFERENCES

- [1] Lavechia, "Deep learning in drug discovery: opportunities, challenges and future prospects," Drug Discovery Today, 2019.
- [2] Karimi, D. Wu, Z. Wang, and Y. Shen, "DeepAffinity: interpretable deep learning of compound-protein affinity through unified recurrent and convolutional neural networks," Bioinformatics, vol. 35, no. 18, pp. 3329–3338, 2019.
- [3] Tan, O. F. O' zgu' l, B. Bardak, I. Eki'og' lu, and S. Sabuncuoglu, "Drug response prediction by ensemble learning and drug-induced gene expression signatures," Genomics, vol. 111, no. 5, pp. 1078–1088, 2019.
- [4] Gonczarek, J. M. Tomczak, S. Zareba, J. Kacmar, P. Dabrowski, and M. J. Walczak, "Interaction prediction in structure-based virtual screening using deep learning," Computers in Biology and Medicine, vol. 100, pp. 253–258, 2018.
- [5] O' ztu' rk, A. O' zgu' r, and E. Ozkirimli, "DeepDTA: deep drug-target binding affinity prediction," Bioinformatics, vol. 34, no. 17, pp. i821–i829, 2018.



International Journal for Research in Applied Science & Engineering Technology (IJRASET)

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538

Volume 12 Issue III Mar 2024- Available at www.ijraset.com

- [6] T. Nguyen and D.-H. Le, "A matrix completion method for drug response prediction in personalized medicine," in *Proceedings of the International Symposium on Information and Communication Technology*, 2018, pp. 410–415.
- [7] H. Le and V.-H. Pham, "Drug response prediction by globally capturing drug and cell line information in a heterogeneous network," *Journal of Molecular Biology*, vol. 430, no. 18, pp. 2993–3004, 2018.
- [8] H. Le and D. Nguyen-Ngoc, "Multi-task regression learning for prediction of response against a panel of anti-cancer drugs in personalized medicine," in *Proceedings of the International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, IEEE, 2018, pp. 1–5. [12] K. Matlock, C. De Niz, R. Rahman, S. Ghosh, and R. Pal, "Investigation of model stacking for drug sensitivity prediction," *BMC Bioinformatics*, vol. 19, no. 3, p. 71, 2018.
- [9] Turki and Z. Wei, "A link prediction approach to cancer drug sensitivity prediction," *BMC Systems Biology*, vol. 11, no. 5, p. 94, 2017.
- [10] Azuaje, "Computational models for predicting drug responses in cancer research," *Briefings in Bioinformatics*, vol. 18, no. 5, pp. 820–829, 2017.
- [11] I. I. Buskin, D. Winkler, and I. V. Tetko, "A renaissance of neural networks in drug discovery," *Expert Opinion on Drug Discovery*, vol. 11, no. 8, pp. 785–795, 2016.
- [12] C. Pereira, E. R. Caffarena, and C. N. dos Santos, "Boosting docking-based virtual screening with deep learning," *Journal of Chemical Information and Modeling*, vol. 56, no. 12, pp. 2495–2506, 2016.



International Journal for Research in Applied Science & Engineering Technology (IJRASET)

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538

Volume 12 Issue III Mar 2024- Available at www.ijraset.com

- a. Zhang, H. Wang, Y. Fang, J. Wang, X. Zheng, and X. S. Liu, "Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model," *PLoS Computational Biology*, vol. 11, no. 9, 2015.
- b. Wan and R. Pal, "An ensemble based top performing approach for NCI-DREAM drug sensitivity prediction challenge," *PLoS ONE*, vol. 9, no. 6, 2014.
- c. S. Jang, E. C. Neto, J. Guinney, S. H. Friend, and A. A. Margolin, "Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data," in *Biocomputing*. World Scientific, 2014, pp. 63–74.
- d. C. Costello, L. M. Heiser, E. Georgii, M. G'onen, M. P. Menden, N. J. Wang, M. Bansal, P. Hintsanen, S. A. Khan, J.-P. Mpindi et al., "A community effort to assess and improve drug sensitivity prediction algorithms," *Nature Biotechnology*, vol. 32, no. 12, p. 1202, 2014.
- e. G'onen and A. A. Margolin, "Drug susceptibility prediction against a panel of drugs using kernelized Bayesian multitask learning," *Bioinformatics*, vol. 30, no. 17, pp. i556–i563, 2014.
- f. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, C. G. A. R. Network et al., "The Cancer Genome Atlas Pan-Cancer analysis project," *Nature Genetics*, vol. 45, no. 10, p. 1113, 2013.
- g. Yang, J. Soares, P. Greninger, E. J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J. A. Smith, I. R. Thompson et al., "Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells," *Nucleic Acids Research*, vol. 41, no. D1, pp. D955–D961, 2012.
- h. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Leh'ar, G. V. Kryukov, D. Sonkin et al., "The Cancer Cell Line Encyclopedia enables predictive modeling of anticancer drug sensitivity," *Nature*, vol. 483, no. 7391, pp. 603–607, 2012.



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 📞 (24*7 Support on Whatsapp)





