

APPLICATION DEVELOPMENT FOR PHARMACOGENOMICS

Submitted in partial fulfillment of the requirements for the award of

Master of Science in Bioinformatics and Data Science

By

HUMAIRA THUSNEEM. S. (41739008)



**DEPARTMENT OF BIOINFORMATICS
SCHOOL OF BIO AND CHEMICAL ENGINEERING**

SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

**Accredited with Grade "A" by NAAC | 12B Status by UGC | Approved by AICTE
Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai - 600 119**

APRIL – 2023



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited with "A" grade by NAAC | 12B Status by UGC | Approved by AICTE

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai – 600 119

www.sathyabama.ac.in

DEPARTMENT OF BIOINFORMATICS

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **HUMAIRA THUSNEEM. S. (41739008)**, who carried out the final year project entitled "**APPLICATION DEVELOPMENT FOR PHARMACOGENOMICS**" under our supervision from January 2023 to April 2023.

Dr. DANIEL ALEX ANAND
Internal Guide

Dr. H. JEMMY CHRISTY
Head of the Department

Submitted for Viva Voce Examination held on

Internal Examiner

External Examiner

DECLARATION

I, **HUMAIRA THUSNEEM. S. (41739008)**, hereby declare that the Project Report entitled “**APPLICATION DEVELOPMENT FOR PHARMACOGENOMICS**”, was done by me under the guidance of **Dr. ALEX ANAND DANIEL** at Sathyabama Institute of Science and Technology, Jeppiaar Nagar, Rajiv Gandhi Road, Chennai-600119, Tamil Nadu, India, is submitted in partial fulfillment of the requirements for the award of **Master of Science in Bioinformatics and Data Science**.

DATE:
PLACE:



SIGNATURE OF THE STUDENT

CHAPTER-1

INTRODUCTION

1.1 OUTLINE OF THE PROJECT

Pharmacogenomics has emerged as one of the most useful types of research in the field of health care and medicine. And my study focuses on pharmacogenomics online applications, where physicians may quickly use the application for drug dose and prescription efficient drugs to patients. It is critical to improve the usage and necessity of medication and gene pharmacogenomics, and this application would assist doctors in reducing the adverse effects of drugs in patients. In this project, the dataset from PharmGKB is trained and then deployed as a web application .

1.1.1 PHARMACOGENOMICS

Pharmacogenomics is the study of how a person's genes affect how they react to medications. To find genetic differences that can impact how a person metabolises, absorbs, or reacts to a medication, it merges the sciences of pharmacology (the study of medicines) with genomics (the study of genes and their roles).

Pharmacogenomics can assist medical professionals in customising treatment strategies and maximising pharmaceutical usage by revealing how a person's genetic make-up affects their reaction to medications. Additionally, it can help with the creation of novel medications that are matched to certain genetic profiles, providing patients with more efficient and secure medical care.

The study of how pharmaceuticals act in the body is known as pharmacology, whereas genomics is the study of the structure, function, and evolution of genes. Pharmacogenomics combines these two topics by investigating how a person's genetic composition influences their medication response.

Pharmacogenomics is the study of genetic differences that can influence how a person metabolises, absorbs, or reacts to a medicine. These differences can affect how medications are digested in the body, as well as their efficacy and toxicity. Understanding a person's genetic composition allows healthcare practitioners to personalise pharmacological therapies to their specific needs, resulting in more effective and safer treatments.

Pharmacogenomics can also aid in the identification of possible drug interactions or adverse effects, allowing healthcare practitioners to make better educated prescription selections. Overall, pharmacogenomics provides a personalised approach to medicine that has the potential to enhance patient outcomes while lowering the risk of adverse drug responses. (Hahn & Roll, 2021)

1.1.2 RELATIONSHIP BETWEEN GENE AND DRUG

The interaction between genes and medications is intricate and complicated. Genes influence how an individual responds to medications, and drugs can also influence gene expression and regulation.

Drug efficacy and safety can be affected by genetic differences in how medications are metabolised, absorbed, and distributed in the body. By analysing a patient's genetic composition, doctors may adapt pharmacological therapies to their specific needs, maximising therapy advantages while minimising unwanted effects.

The development of tailored medicines is another manner in which genes and pharmaceuticals are linked. As previously said, targeted treatments are medications that are precisely developed to target genes or proteins involved in the disease process. These medications can be beneficial and have fewer negative effects than traditional treatments that impact many different sections of the body because they target particular genes or proteins.

Drugs can also influence gene expression and regulation. Some medications, for example, can switch off or on specific genes, affecting the creation of specific proteins in the body. This technique can be utilised to treat disorders caused by an excess or deficiency of a certain protein. (Lee et al., 2023)

1.2 PharmGKB

PharmGKB (Pharmacogenomics Knowledgebase) is a database that gathers, curates, and disseminates knowledge on how genetic differences impact pharmacological reactions. PharmGKB was created to aid research in the subject of pharmacogenomics, which is concerned with understanding how genetic diversity

impacts medication efficacy and safety. The database contains data on the connection between genetic variants, drug metabolism, medication effectiveness, and adverse drug responses. To offer researchers with a full knowledge of the interaction between genetics and pharmacological response, this material is organised into gene summaries, drug summaries, and pathway summaries. The PharmGKB database is open to academics, doctors, and the general public, with the goal of improving patient care and advancing the science of pharmacogenomics.

The Pharmacogenomics Knowledgebase (PharmGKB) is a publicly accessible, online knowledge base that is responsible for the accumulation, curation, integration, and distribution of knowledge about the influence of human genetic variation on medication response.[1] It is supported by the National Institutes of Health (NIH) National Institute of General Medical Sciences (NIGMS) and is a member of the NIH Pharmacogenomics Research Network (PGRN). It has been handled by Stanford University since its founding in 2000.

1.2.1 PURPOSE OF PharmGKB

PharmGKB's major objective is to help researchers better understand pharmacogenomics, also known as pharmacogenetics (PGx), which is the study of how genetic diversity impacts how an individual responds to a medicine. PharmGKB manually collects PGx information from the source literature in order to accomplish this aim and then saves it in the knowledge base. PharmGKB can uncover recurrent genetic variant-drug response relationships by aggregating this information. The possibility of implementing variant-drug interactions in the clinic can then be examined when there is a wealth of supporting data.[3] PharmGKB cofounded the Clinical Pharmacogenetic Implementation Consortium (CPIC), an organisation responsible for developing and disseminating peer-reviewed, openly accessible genotype-based drug-dosing recommendations for situations where there is strong PGx literature evidence for clinicians.

1.3 MACHINE LEARNING USING PYTHON

Machine learning is an artificial intelligence subject that focuses on creating algorithms and models that allow computers to learn from data and make predictions or judgements without being explicitly programmed. In other words, machine learning algorithms employ statistical approaches to find patterns and correlations in data and then utilise this information to generate predictions or perform actions in response to fresh input.

Machine learning algorithms are classified into three types: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning algorithms learn from labelled examples, which include the right answers as well as the input data. Unsupervised learning algorithms learn from unlabeled data and attempt to detect patterns and correlations in the data without explicit instruction. Reinforcement learning algorithms learn through trial and error, obtaining feedback in the form of rewards or penalties based on their behaviours in response to various stimuli.

From picture and speech recognition to fraud detection, predictive maintenance, and personalised recommendations, machine learning has a wide range of applications. It is a fast expanding field with the potential to transform many parts of our life.

1.3.1 MACHINE LEARNING USES IN MEDICINE

- Machine learning algorithms can be used to analyse medical imagery and patient data to aid in illness diagnosis. This has resulted in the creation of programmes such as computer-aided diagnostics systems, which can assist to enhance accuracy and minimise diagnostic time.
- To uncover prospective medication candidates, machine learning algorithms may be used to analyse enormous databases of chemical molecules. This can assist to speed up drug discovery and lower the cost of creating new medications.
- Machine learning algorithms may be used to analyse patient data and produce personalised treatment regimens. By personalising therapy to individual patients based on their unique traits, this can assist to improve patient outcomes.

- Machine learning algorithms may be used to analyse patient data in order to forecast the possibility of specific diseases emerging. This can aid in identifying high-risk individuals and allowing for early intervention.
- Data from clinical trials may be analysed using machine learning algorithms to enhance the design and execution of future studies. This has the potential to minimise the cost and time necessary for medication development while also improving patient outcomes.

1.3.2 ALGORITHMS IN MACHINE LEARNING

In machine learning, an algorithm is a collection of instructions or methods that a computer programme uses to discover patterns or rules from data. These algorithms let machines to learn from data and improve their performance on a certain activity automatically.

An algorithm in the context of machine learning may be thought of as a script that a computer follows to analyse and interpret data. These algorithms often use complicated mathematical models and statistical approaches that aid in the identification of patterns in data and the prediction of future events based on those patterns.

Some common types of machine learning algorithms are as follows,

1. Linear Regression Algorithm
2. Logistic Regression Algorithm
3. Decision Tree
4. SVM
5. KNN
6. K-Means Clustering
7. Random Forest

1.3.2.1 LINEAR REGRESSION

Linear regression is one of the most popular and simple machine learning algorithms that is used for predictive analysis. Here, predictive analysis defines prediction of something, and linear regression makes predictions for continuous numbers

1.4 WEB APPLICATION

A web application is a computer programme that can be accessed through the internet using a web browser. Web apps are built to operate on web servers and may be accessed from anywhere in the world with an internet connection.

Web-based applications may be used for a variety of functions, ranging from simple tasks like filling out a form or looking for information to more complicated operations like online banking, e-commerce, and social networking.

Web applications can run on a number of platforms, including on-premises servers, cloud-based servers, and third-party hosting services. They are also available on a range of platforms, including desktop computers, laptop computers, tablets, and smartphones.

Overall, web applications are a strong and adaptable method of delivering software solutions via the internet, and they are becoming increasingly significant in our daily lives.

CHAPTER-2

LITERATURE SURVEY

CHAPTER-2

LITERATURE SURVEY

2.1 PHARMACOGENOMICS

The article by Whirl-Carrillo et al. (2021) proposes a methodology for evaluating pharmacogenomics information and its application to personalised medicine. To assure the safety and efficacy of pharmacogenomic testing and personalised prescribing, the authors emphasise the significance of evidence-based evaluation. The authors' framework is made up of three parts: 1) the strength and consistency of the evidence linking genetic variation to drug response, 2) the clinical validity of pharmacogenomic tests in predicting drug response, and 3) the clinical utility of pharmacogenomic testing in improving patient outcomes. The authors provide a score system to evaluate the evidence for each of these components. The paper covers the difficulties in assessing pharmacogenomic knowledge and emphasises the importance of standardised rules and criteria for assessing pharmacogenomic tests. The authors also stress the need of including patient preferences and values in the review process to ensure that personalised prescribing corresponds with patient goals and preferences. Overall, this paper presents a useful paradigm for assessing pharmacogenomic knowledge and its application to personalised treatment. To assure the safety and efficacy of pharmacogenomic testing and personalised prescribing, the authors emphasise the significance of evidence-based evaluation. The proposed methodology can be used to evaluate the evidence supporting pharmacogenomic tests and guide therapeutic decision-making.

Another study presents an outline of the current state of pharmacogenomics and its future prospects. The study begins by discussing pharmacogenomics and its significance in personalised medicine. The author then goes over the present state of pharmacogenomics research, including how genetic variations can be used to predict drug responsiveness and adverse drug reactions. The author also discusses

some of the limitations of pharmacogenomics, such as the necessity for large-scale research to uncover genetic variants related with drug response and the need for improved methods for analysing and interpreting genetic data. The study then goes over some of the most recent advances in pharmacogenomics, such as the use of machine learning and artificial intelligence to create more accurate and trustworthy predictive models. The author also highlights the possible application of predictive models. The author also highlights the possible application of pharmacogenomics in the creation of novel pharmaceuticals as well as the improvement of drug safety and efficacy. Finally, the author discusses the future of pharmacogenomics and how this area has the potential to revolutionise the way medications are developed, prescribed, and used. To realise the full promise of this subject, the author emphasises the significance of continuous investment in pharmacogenomics research as well as cooperation between academia, industry, and regulators. Overall, the paper gives a detailed assessment of the current state of pharmacogenomics and its future prospects, stressing both the obstacles and potential connected with this discipline. The study also discusses current advances in pharmacogenomics and how this discipline has the potential to change the way medications are manufactured and used in the future (Pirmohamed, 2023).

Another study offers a review of the role of pharmacogenomics in COVID-19 treatments. The authors begin by explaining the notion of pharmacogenomics and its usefulness in personalised medicine, emphasising pharmacogenomics' potential to improve the safety and efficacy of COVID-19 medicines. The authors then go over the pharmacogenomics of other COVID-19 treatments, such as remdesivir, hydroxychloroquine, and tocilizumab. The authors examine the genetic variations that have been found as potentially affecting drug response for each therapy and emphasise the need for additional research to confirm these connections. The authors also examine the potential for pharmacogenomics to guide COVID-19 drug selection, emphasising the need for additional research to construct prediction models for COVID-19 therapy response based on genetic data. Finally, the authors discuss the challenges and limitations of pharmacogenomics in the context of COVID-19 therapies, such as the need for large-scale studies to identify genetic variants associated with drug response and improved methods for analysing and interpreting genetic data. Overall, the paper presents a thorough discussion of the

significance of pharmacogenomics in the context of COVID-19 treatments, emphasising both the potential benefits and problems of this discipline. The report also discusses the specific pharmacogenomics of COVID-19 medicines, as well as the necessity for additional research to fully realise the potential of pharmacogenomics in the fight against COVID-19 (Takahashi et al., 2020).

2.2 MACHINE LEARNING

The article by Sarker's (2021) presents an overview of machine learning algorithms, real-world applications, and future research prospects. The author explores many types of machine learning algorithms, such as supervised learning, unsupervised learning, and reinforcement learning, as well as instances of their applications in domains such as healthcare, finance, and social media. The article examines typical strategies for feature selection, dimensionality reduction, and outlier detection, as well as the importance of data quality and preprocessing in machine learning. The author also discusses the significance of model selection and evaluation, as well as the relevance of interpretability and transparency in machine learning models, particularly in applications involving human beings. The article examines current machine learning challenges and future research objectives, such as the need for more robust and efficient algorithms. Overall, this article gives a thorough examination of machine learning algorithms and their practical applications. In machine learning, the author emphasises the relevance of data quality and pre-processing, model selection and evaluation, and interpretability and transparency. The page contains useful information for machine learning researchers, practitioners, and policymakers.

Another study provides a practical introduction to machine learning in the context of medical research and practise. The authors present an overview of various types of machine learning algorithms, such as supervised and unsupervised learning, as well as their possible applications in medicine. The paper emphasises the relevance of data quality and pre-processing in machine learning applications, such as feature selection and dimensionality reduction. In addition, the authors examine model selection and evaluation, emphasising the relevance of cross-validation and hyperparameter adjustment in achieving optimal model performance. Several examples of machine learning applications in medicine are provided in the article,

including disease diagnosis and prognosis, drug discovery, and personalised treatment planning. The authors also examine the possible benefits and drawbacks of machine learning in medicine, as well as the importance of carefully considering ethical and legal consequences. Overall, this essay is an excellent introduction to machine learning in medicine. The authors offer practical recommendations on data quality and pre-processing, model selection, and model evaluation, as well as highlighting the possible applications and limitations of machine learning in medicine. The article is useful for researchers and practitioners in the field of medical research and practise (Sidey-Gibbons & Sidey-Gibbons, 2019).

Another article compared the performance of numerous supervised machine learning algorithms for disease prediction. The authors begin by outlining the concept of disease prediction using machine learning algorithms, as well as the possible benefits of adopting such methods for disease detection and prevention. The authors then go into the datasets that were used in the study, which included data from the UCI Machine Learning Repository and the National Institute of Diabetes and Digestive and Kidney Diseases. Following that, the authors compare the performance of various supervised machine learning algorithms, such as decision trees, random forests, support vector machines, naive Bayes, k-nearest neighbours, and logistic regression. The authors analyse the performance of each method using many performance criteria, including accuracy, sensitivity, specificity, and area under the curve (AUC). In terms of accuracy, sensitivity, and specificity, the results demonstrate that random forests and support vector machines beat the other techniques. The authors do caution, however, that the efficiency of each algorithm differs based on the dataset and disease being predicted. The authors also conduct a feature selection study to determine which features are most significant for each disease prediction task. The findings indicate that different traits are significant for different diseases, emphasising the necessity of personalised feature selection in disease prediction. Overall, the research compares multiple supervised machine learning methods for disease prediction and emphasises the necessity of personalised feature selection for accurate and effective disease prediction. The findings of this study may help in the future development of machine learning-based disease prediction tools (Uddin et al., 2019).

2.3 PHARMACOGENOMICS & ML

The work by Mehmood et al.'s (2023) discusses how machine learning and pharmacogenomics are used to discover effective pharmaceuticals and biomarkers for breast cancer treatment. The authors developed a machine learning model to predict the efficacy of different medications for breast cancer treatment using publically available gene expression data from breast cancer patients. The researchers also used pharmacogenomic data to discover potential breast cancer indicators. The study's findings imply that the machine learning model can effectively predict the efficacy of various medications for breast cancer treatment. Several potential biomarkers for breast cancer treatment were also identified by the authors, including genes involved in DNA repair and the immune system. The essay emphasises the importance of machine learning and pharmacogenomics in personalised medicine, notably in oncology. Overall, this paper describes a novel use of machine learning and pharmacogenomics to find beneficial medications and biomarkers for breast cancer treatment. The authors provide important insights on the possibilities of personalised

Another paper describes how to create machine learning models for drug sensitivity prediction based on cell line-based pharmacogenomics data. The study opens by exploring the challenges and prospects of drug sensitivity prediction, as well as the usefulness of machine learning in resolving these challenges. The authors then go over the steps for creating machine learning models for drug sensitivity prediction, including data preparation, feature selection, model selection, and model evaluation. The recommendations were created after a thorough examination of publicly accessible cell line-based pharmacogenomics datasets and the effectiveness of multiple machine learning models for drug sensitivity prediction. The authors also offered a full overview of the numerous machine learning algorithms employed, as well as the significance of proper data preparation and feature selection for model performance. The guidelines' utility was proved by the authors' development and evaluation of different machine learning models for drug sensitivity prediction utilising cell line-based pharmacogenomics data. The guidelines provided a framework for constructing strong and accurate machine learning models for medication sensitivity prediction, according to the findings. Overall, the research is a well-designed and well-executed work on providing guidelines for developing

machine learning models for drug sensitivity prediction utilising cell line-based pharmacogenomics data. The study also includes a full discussion of the machine learning techniques employed, as well as the relevance of proper data preparation and feature selection for model performance, which can be useful for academics interested in constructing comparable models (Sharifi-Noghabi et al., 2021).

2.4 WEB APPLICATION

The study report outlines the creation of SMART, an interactive web tool for analysing and visualising DNA methylation data. The paper opens by explaining the significance of DNA methylation in gene regulation, as well as the numerous approaches used to study it. The authors next describe the SMART web programme, which includes a comprehensive array of capabilities for DNA methylation study, such as data preprocessing, quality control, statistical analysis, and visualisation. The authors used publicly available DNA methylation data to conduct tests that demonstrated the utility of the SMART web tool for various types of DNA methylation analyses. The application was also compared to other available tools for DNA methylation analysis, and the results revealed that the SMART online application provided more extensive and user-friendly data analysis and visualisation features. The R Shiny framework was used to create the SMART online application, which can be accessed via a web browser. The application has an easy-to-use interface for entering data, selecting analytic settings, and seeing findings. The report also includes a full discussion of the system architecture and the numerous data analysis algorithms. Overall, the work presents a well-planned and executed investigation into the development of an interactive web application for DNA methylation research and visualisation. The SMART web programme includes a full range of tools for DNA methylation analysis and visualisation, which might be valuable for epigenetics academics and practitioners. The study also includes a full description of the system architecture and the various methods utilised, which will be useful for academics working on similar applications (Li et al., 2019).

Another study report presents the creation of CemrgApp, an interactive medical imaging application for cardiovascular research. The study begins by emphasising the importance of medical imaging in cardiovascular research as well as the numerous image analysis methodologies. The authors then present the CemrgApp,

a comprehensive suite of tools for medical image analysis that includes image processing, computer vision, and machine learning toolkits. The CemrgApp is written in MATLAB and may be accessible via a graphical user interface. The application has an easy-to-use interface for importing medical photos, segmenting and processing them, and analysing the resulting data. The application also includes a number of machine learning toolkits for automatic picture analysis. The authors conducted trials with several medical imaging datasets to demonstrate the CemrgApp's utility in various sorts of cardiovascular research. The CemrgApp was also compared to other existing medical imaging programmes, with the results indicating that the CemrgApp provided more extensive and user-friendly capabilities for medical picture processing and visualisation. Overall, the study on the development of an interactive medical imaging application for cardiovascular research is well-designed and well-executed. The CemrgApp provides a comprehensive range of tools for medical image analysis and visualisation that can be valuable for cardiovascular researchers and practitioners. The study also includes a full description of the system architecture and the various data analysis techniques, which can be useful for academics interested in building comparable applications (Razeghi et al., 2020).

CHAPTER-3

AIM AND SCOPE

3.1 AIM

To create a web application based on pharmacogenomic studies and databases that will help clinicians with dosage, drug prescription, and tailored medicine.

3.2 SCOPE

- **Clinical Decision Support:** A pharmacogenomics online application might deliver real-time pharmacogenomic information to doctors to aid in clinical decision-making. This might contain warnings about medication interactions, dosing recommendations, and therapy recommendations based on a patient's unique genetic profile.
- **Personalised Medicine:** A pharmacogenomics online application might be created to give patients with personalised treatment recommendations based on genetic information. This information might include medication effectiveness, adverse effects, and possible interactions.
- **Research and Development:** By giving access to a comprehensive pharmacogenomics database, a pharmacogenomics online application might aid in the research and development of novel medications. This might contain information about pharmacological targets, interactions, and genetic variants that influence drug response.
- **Collaboration:** A pharmacogenomics online application might allow researchers, physicians, and pharmaceutical corporations to work together. This might include data sharing features, cooperation on research initiatives, and access to resources like clinical trial data and genetic databases.
- **Data Management:** A pharmacogenomics online application might be built to manage massive volumes of genetic, clinical, and drug-related data. This might contain elements like data storage, data analysis, and data visualisation.
- **Education and Training:** A pharmacogenomics online application might be built to deliver pharmacogenomics education and training to healthcare

professionals, researchers, and patients. Online classes, webinars, and instructional materials are examples of such resources.

- **Regulatory Compliance:** A pharmacogenomics online application might be created to fulfil HIPAA (Health Insurance Portability and Accountability Act) rules for data protection and security.
- **Integration with Electronic Health Records (EHRs):** To give doctors with simple access to pharmacogenomic information during patient consultations, a pharmacogenomics online application might be connected with electronic health record systems.
- **Patient Engagement:** A pharmacogenomics online application might be built to include patients in their own care by giving them access to genetic data and personalised medicine recommendations.
- **Mobile Access:** A pharmacogenomics web application might be built to allow patients and physicians to access genetic data and treatment recommendations while on the go.

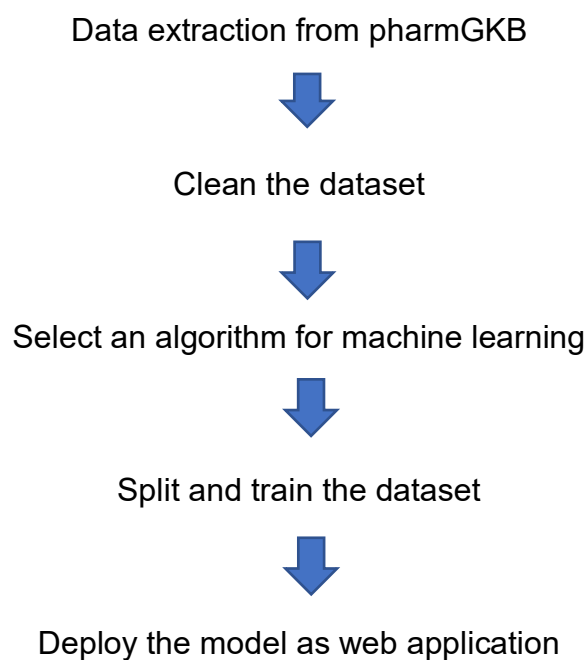
CHAPTER-4

MATERIAL AND METHODS

4.1 METHODOLOGY

The dataset is acquired from pharmgkb, trained using machine learning methods like linear regression, and then the resulting model is made available as a web application using libraries like streamlit and flask.

4.1.1 WORKFLOW OF PROJECT



4.2 DATASET

4.2.1 PHARMGKB

Data on clinical variations is extracted from the csv file PharmGKB. This file includes a list of all the clinical annotations, together with the degree of evidence for each

variant-drug pair. The table includes elements like the level of evidence, the drug, the kind (either efficient or hazardous), the gene, and variations.

The dataset contains information about each entry that corresponds to a single genetic variant, such as the gene name, the position of the variant within the gene, the type of variant, the reference and alternative alleles, and the clinical significance of the variant.

	A	B	C	D	E	F	G	H	I
1	variant	gene	type	level of evidence	chemicals	phenotypes			
2	CYP2C9*1	CYP2C9	Metabolism/PK	1A	meloxicam				
3	CYP2C9*1	CYP2C9	Metabolism/PK	1A	lornoxicam				
4	CYP2C9*1	CYP2C9	Metabolism/PK	1A	siponimod				
5	rs1737684	DPYD	Toxicity	1A	capecitabine	Neoplasms			
6	rs2297595	DPYD	Toxicity	1A	capecitabine	Neoplasms			
7	rs1801265	DPYD	Toxicity	1A	capecitabine	Neoplasms			
8	CYP2C19*	CYP2C19	Metabolism/PK	1A	dexlansoprazole				
9	CYP2C9*1	CYP2C9	Toxicity	1A	warfarin	over-anticoagulation			
10	rs1801160	DPYD	Toxicity	1A	capecitabine	Neoplasms			
11	rs1801159	DPYD	Toxicity	1A	capecitabine	Neoplasms			
12	UGT1A1*1	UGT1A1	Toxicity	1A	atazanavir	HIV Infections,Hyperbilirubinemia			
13	TPMT*1	TPMT	Dosage	1A	azathioprine	dose reduction			
14	rs1737684	DPYD	Toxicity	1A	fluorouracil	Neoplasms			
15	rs1801158	DPYD	Toxicity	1A	fluorouracil	Neoplasms			
16	rs1801159	DPYD	Toxicity	1A	fluorouracil	Neoplasms			
17	rs1801158	DPYD	Toxicity	1A	capecitabine	Neoplasms			
18	rs1801265	DPYD	Toxicity	1A	fluorouracil	Neoplasms			
19	rs1801160	DPYD	Toxicity	1A	fluorouracil	Neoplasms			
20	rs2108622	CYP4F2	Dosage	1A	warfarin				
21	rs1214345	EGFR	Efficacy	1A	gefitinib	Carcinoma, Non-Small-Cell Lung			
22	rs3745274	CYP2B6	Metabolism/PK	1A	efavirenz	HIV Infections			
23	rs9923231	VKORC1	Dosage	1A	phenprocoumon				
24	rs1168552	NUDT15	Dosage	1A	mercaptopurine				
25	CYP2D6*1	CYP2D6	Metabolism/PK	1A	hydrocodone				
26	CYP2D6*1	CYP2D6	Efficacy	1A	tropisetron	Vomiting			
27	CYP2C9*1	CYP2C9	Metabolism/PK	1A	ibuprofen				
28	CYP2D6*1	CYP2D6	Metabolism/PK	1A	desipramine	Depressive Disorder, Major,Mental Disorders			
29	rs7450333	CFTR	Efficacy	1A	ivacaftor	Cystic Fibrosis			

Fig 4.1 DATASET

4.3 PROGRAMMING LANGUAGE -PYTHON

Python is the ultimate platform for this project since it is easily used for both web applications and machine learning. A high-level, interpreted programming language called Python is used for a variety of tasks, including creating websites and analysing data as well as for scientific computers and artificial intelligence. Python's popularity can be attributed in part to its readability and simplicity, which make it simple to understand and apply. Along with several third-

party libraries that are readily installable and usable, it features a sizable standard library that offers a wide range of tools and capabilities out of the box.

Additionally, Python is very adaptable, enabling programmers to use it for a variety of tasks. It can be used for web development using frameworks like Django and Flask, these frameworks offer a methodical approach to developing online applications, with integrated assistance for dealing with web requests, controlling databases, and displaying HTML templates. And for data analysis using libraries like Pandas and NumPy, scientific computing using SciPy, and machine learning using libraries like TensorFlow and PyTorch. These libraries include a wide range of instruments and methods for creating, refining, and testing machine learning models.

It is a strong and adaptable programming language that works well for a variety of tasks, from straightforward scripts to intricate, enterprise-level systems.

Building intelligent web applications using Python machine learning and web development may be quite effective. For instance, you might employ machine learning to provide consumers personalised suggestions based on what they have searched for. Python gives you the freedom to create a broad variety of sophisticated web apps that may enhance user experience and support your business objectives.

4.4 DATA CLEANING (EXCEL-FILTER FEATURE)

In Excel, cells are organised in a grid with rows and columns. Text, numbers, formulae, and other data can all be stored in a cell. Users can do computations on the data in the cells using formulas. Along with data validation, conditional formatting, and data sorting and filtering, Excel also offers these features.

The filter function in Excel is an effective tool for cleaning up data. It makes it simpler to find and eliminate mistakes, inconsistencies, and duplication since it enables you to quickly isolate and analyse groups of data based on certain criteria.

The following are some techniques for cleaning up data using Excel's filter feature:

Eliminating duplicates: To make sure that your data is correct and consistent, you may use the filter tool to find and delete duplicate values within a single column or across many columns.

5016	rs2273697 ABCC2	Toxicity	4 methotrex Arthritis, Rheumatoid
5017	rs1128503 ABCB1	Toxicity	4 methotrex Arthritis, Rheumatoid
5018	rs1045642 ABCB1	Efficacy	4 methotrex Arthritis, Juvenile Rheumatoid,Arthritis, Rheumatoid
5019	rs1805087 MTR	Toxicity	4 methotrex Brain Diseases,Drug Toxicity,Osteosarcoma,Precursor Cell Lymphoblastic Leukemia-Lymphoma
5020	rs1801394 FASTKD3,A	Toxicity	4 methotrex Arthritis, Juvenile Rheumatoid,Arthritis, Rheumatoid
5021	rs2231142 ABCG2	Toxicity	4 methotrex Burkitt Lymphoma,Drug Toxicity,Lymphoma, T-Cell,Osteosarcoma,Precursor Cell Lymphoblastic Leukemia-Lymphoma
5022	rs1104587 SLC01B1	Toxicity	4 methotrex Lymphoma,Osteosarcoma,Precursor Cell Lymphoblastic Leukemia-Lymphoma
5023	rs1801131 MTHFR	Toxicity	4 methotrex Burkitt Lymphoma,Lymphoma, Non-Hodgkin,Lymphoma, T-Cell,Precursor Cell Lymphoblastic Leukemia-Lymphoma
5024	HLA-B*58: HLA-B	Toxicity	4 carbamaz drug reaction with eosinophilia and systemic symptoms,Epidermal Necrolysis, Toxic,Maculopapular Exanthema,severe cutaneous adverse reactions,Stevens-Johnson Syndrome
5025	rs1128503 ABCB1	Dosage	4 methadon Heroin Dependence,Opioid-Related Disorders
5026	rs1045642 ABCB1	Dosage	4 methadon Heroin Dependence,Opioid-Related Disorders
5027	CYP2D6*1 CYP2D6	Efficacy	4 oxycodone Pain
5028	rs1799971 OPRM1	Toxicity	4 oxycodone adverse events,Constipation,Delirium,Nausea,Pruritus,somnolence,Urinary Retention
5029	rs4680 COMT	Toxicity	4 morphine adverse events,Constipation,Delirium,Nausea,Pruritus,somnolence,Urinary Retention
5030	rs4680 COMT	Toxicity	4 oxycodone adverse events,Constipation,Delirium,Nausea,Pruritus,somnolence,Urinary Retention
5031	rs4680 COMT	Efficacy	4 methylphe Attention Deficit Disorder with Hyperactivity
5032	CYP2C19* CYP2C19	Efficacy	4 citalopram Depressive Disorder, Major,Obsessive-Compulsive Disorder
5033	rs1045642 ABCB1	Efficacy	4 paclitaxel Breast Neoplasms,Neoplasms
5034	rs1045642 ABCB1	Efficacy	4 docetaxel Breast Neoplasms,Neoplasms
5035	rs1045642 ABCB1	Toxicity	4 paclitaxel Drug Toxicity,Neoplasms,Neutropenia,Peripheral Nervous System Diseases
5036	CYP2C8*1, CYP2C8	Toxicity	4 paclitaxel Drug Toxicity,Neutropenia,Peripheral Nervous System Diseases,Toxic liver disease
5037	rs1801253 ADRB1	Efficacy	4 Beta Block Cardiomyopathy, Dilated,Heart Failure
5038	rs2230345 GRK5	Efficacy	4 Beta Block Heart Failure
5039	rs1042713 ADRB2	Efficacy	4 Beta Block Cardiomyopathy, Dilated,Heart Failure
5040	HLA-A*02: HLA-A	Toxicity	4 carbamazepine Drug Hypersensitivity,drug reaction with eosinophilia and systemic symptoms,Epidermal Necrolysis, Toxic,Maculopapular Exanthema,severe cutaneous adverse reactions,Stevens-Johnson Syndrome
5041	CYP3A4*1 CYP3A4	Dosage	4 fentanyl Pain, Postoperative

4.4.1 FEATURE SELECTION

4.5 FEATURE EXPLANATION

Gene: The name of the gene containing the under investigation genetic mutation is specified in this column. This knowledge is crucial because it aids in the understanding of the biological environment in which the variation is found and the potential processes by which it may impact medication response by researchers and medical professionals.

Variant: The information in this column describes the particular genetic variation under investigation, including its position within the gene and its nature (such as single nucleotide polymorphism or insertion/deletion). Given that different types of variations may have varying impacts on treatment response, this information is crucial for identifying and characterising the variant.

Evidence: Information on the research or studies that documented a link between a variation and medication response or toxicity is provided in this column. This information includes the study's design, participant count, and statistical significance of the connection. This data is crucial for assessing the reliability and validity of the association.

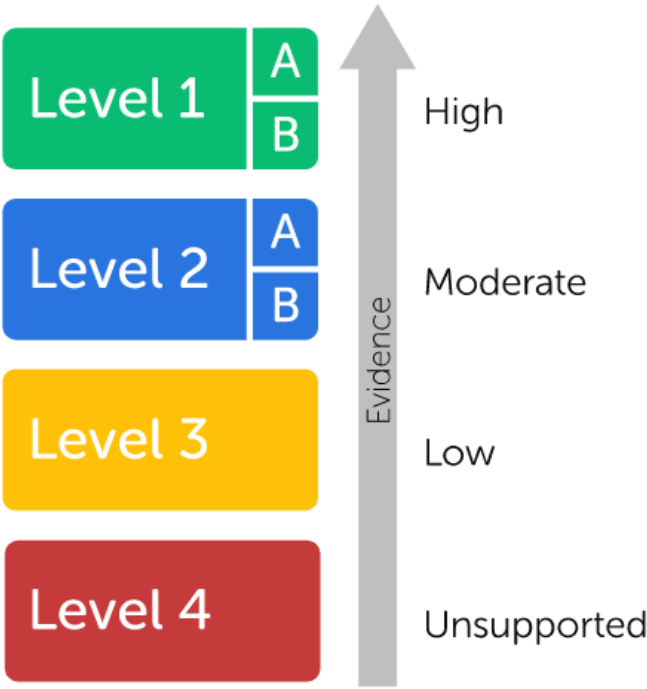


FIG 4.2

Drug(s): The drug or medicines that were investigated in connection to the variation are listed in this column. This knowledge is crucial for determining how

the mutation may alter a particular drug's reaction and for directing personalised medicine strategies.

4.6 EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is a method for analysing and comprehending data sets by highlighting key aspects of them via the use of statistical and visual methods. EDA is frequently the initial stage of data analysis and may aid in finding patterns, connections, and outliers in the data.

Understanding the data and its underlying structure is EDA's main objective. It entails looking for patterns and trends in the data, spotting outliers and anomalies locating missing or inconsistent data, and assessing the data for any potential biases or inaccuracies.

```
▶ RawData.info()
```

```
↳ <class 'pandas.core.frame.DataFrame'>  
RangeIndex: 3836 entries, 0 to 3835  
Data columns (total 6 columns):  
#   Column                Non-Null Count  Dtype  
---  ---  
0   variant                3836 non-null   object  
1   gene                   3836 non-null   object  
2   type                   3836 non-null   object  
3   level of evidence      3836 non-null   object  
4   chemicals              3836 non-null   object  
5   phenotypes             3836 non-null   object  
dtypes: object(6)  
memory usage: 179.9+ KB
```

```
[ ] RawData.shape
```

```
(3836, 6)
```

Fig Raw data shape and size

After cleaning,

Total number of rows:3836

And total number of columns :6

```
▶ for cols in df:
    print(f'{cols}:{df[cols].unique()}')
    print("-----\n")

variant:['rs17376848' 'rs2297595' 'rs1801265' ...
        'CYP2D6*1, CYP2D6*2, CYP2D6*2xN, CYP2D6*3, CYP2D6*4, CYP2D6*6'
        'CYP2D6*1, CYP2D6*3, CYP2D6*4, CYP2D6*5, CYP2D6*10, CYP2D6*17, CYP2D6*41'
        'CYP3A4*1, CYP3A4*18']
-----

gene:['DPYD' 'CYP2C9' 'UGT1A1' ... 'PTPRM' 'AMPD1' 'ATM,C11orf65']
-----

type:['Toxicity' 'Dosage' 'Efficacy' 'Metabolism/PK' 'Efficacy,Toxicity'
      'Other' 'Dosage,Toxicity' 'Efficacy,Metabolism/PK'
      'Toxicity,Metabolism/PK' 'Dosage,Efficacy' 'Dosage,Metabolism/PK'
      'Efficacy,Toxicity,Metabolism/PK']
-----

level of evidence:['1A' '1B' '2A' '2B' '3' '4']
-----

chemicals:['capecitabine' 'warfarin' 'atazanavir,atazanavir / ritonavir'
           'azathioprine' 'fluorouracil' 'gefitinib' 'efavirenz' 'tropisetron'
           'desipramine' 'ivacaftor' 'rasburicase' 'trimipramine' 'clopidogrel'
           'ivacaftor / lumacaftor' 'ivacaftor / tezacaftor' 'flucloxacillin'
           'hormonal contraceptives for systemic use' 'phenytoin' 'desflurane'
           'amitriptyline'
           'desflurane,enflurane,halothane,isoflurane,methoxyflurane,sevoflurane,succinylcholine']
```

This image allows us to examine all of the feature categories.

```
#null
df.isnull().sum()
```

```
variant      0
gene          0
type          0
level of evidence  0
chemicals     0
phenotypes    0
dtype: int64
```

```
[ ] #duplicate
df.duplicated().sum()
```

```
14
```

```
for feature in categorical:
    print('The feature is {} and number of categories are {}'.format(feature, len(df[feature].unique())))
```

```
The feature is variant and number of categories are 2439
The feature is gene and number of categories are 1008
The feature is type and number of categories are 12
The feature is level of evidence and number of categories are 6
The feature is chemicals and number of categories are 696
The feature is phenotypes and number of categories are 662
```

Analyses are performed to see whether any characteristics have duplicate or missing data, and those are subsequently eliminated. Additionally, each characteristic is examined to determine the overall categories.

4.7 SPLITTING AND TRAINING DATA

We split the dataset into training and testing sets in order to assess how well our machine learning model performs when applied to new data. The model may perform well on the training data but poorly on fresh, untested data if we train and test it on the same dataset. Overfitting occurs when a model memorises training data rather than learning broad patterns and connections that can be applied to fresh data.

We can train the model on the training set and assess its performance on the testing set by dividing the dataset into training and testing sets. This enables us to predict how effectively our model will function with brand-new data that it has never encountered before. Additionally, occasionally our data may be unbalanced, with one type being significantly more frequent than the other. In these circumstances, dividing the data at random may lead to an uneven representation of the classes in the training and testing sets. Stratified sampling, which divides

the data while retaining the proportion of each class in the training and testing sets, can be used to address this.

We can reliably assess the performance of our machine learning model and ensure that it generalises successfully to new, unknown data by dividing the dataset into training and testing sets.

```
[ ] #splitting the data based on independent and dependent
    x=df.iloc[:, :-1].values
    y=df.iloc[:, 1].values
```

```
[ ] #dividing the data based on testing and training dataset
    x_train,x_test,y_train,y_test=train_test_split(x,y,train_size=2/3,random_state=0)
```

Using the linear regression algorithm, the trained dataset is then turned into a model, which is then stored in pickle format and given the name "model.pkl."

4.8 LIBRARIES USED

4.8.1 PIP

Python has a package installer called pip. The first and main part that is most helpful is pip since it makes it simple to install multiple packages. Python packages, which are collections of code that may be used in other projects, are installed, upgraded, and uninstalled using it. You must have Python set up on your computer in order to utilise pip. The Python Package Index (PyPI), a repository containing over 300,000 packages, may be accessed and installed via Pip.

4.8.2 NUMPY

NumPy, Numerical Python is the name of a Python library. On arrays and matrices, it is utilised to carry out mathematical and scientific procedures. Using NumPy, you can effectively store and manage huge arrays of numerical data.

Multidimensional arrays can be used to represent matrices, vectors, or other data structures. NumPy supports the creation of multidimensional arrays.

Mathematical operations: A variety of mathematical operations, including as addition, subtraction, multiplication, and division, may be carried out on arrays using NumPy.

By automatically distributing the smaller array to match the shape of the bigger array, NumPy enables you to execute arithmetic operations on arrays of various sizes and forms.

4.8.3 PANDAS

Pandas is a well-known open-source toolkit for Python that facilitates data manipulation and analysis. It offers simple data structures and tools for data analysis so that users may handle and manipulate data quickly and effectively.

Among the essential characteristics of pandas are:

Data structures: Pandas has two primary data structures, Series and DataFrame, which provide you broad storage and data manipulation options.

Data transformation and fusion are made possible by a variety of methods provided by pandas, including as filtering, sorting, grouping, and aggregation.

Data visualisation: To produce graphs and charts that aid in a better understanding of the data, pandas may be used with other data visualisation tools like as Matplotlib and Seaborn.

4.8.4 MATPLOTLIB

Python users frequently use Matplotlib to build static, animated, and interactive visualisations. It offers a variety of 2D and 3D graphs for visualising data. The NumPy-based library is frequently used with other scientific computing libraries like Pandas, SciPy, and scikit-learn.

4.8.5 SCIKIT-LEARN

A well-known Python framework for machine learning called Scikit-learn offers a number of tools for modelling and data analysis. Numerous machine learning methods, including classification, regression, clustering, and dimensionality reduction, are included in Scikit-learn. Additionally, it offers tools for assessment, data preparation, and model selection.

Some of the popular algorithms provided by scikit-learn are:

- Linear Regression

- Logistic Regression
- Decision Trees
- Random Forests
- Support Vector Machines (SVM)
- K-Nearest Neighbors (KNN)
- K-Means Clustering

Numerous academic and professional applications, including text classification, picture recognition, and financial modelling, use Scikit-learn extensively. It is an open-source library that is frequently updated with fresh features and computational methods.

4.8.6 PICKLE

A Python package called Pickle is used to serialise and de-serialize items. De-serialization is the act of recovering the original item from the serialised form. Serialisation is the process of transforming an object into a format that can be stored or transferred.

Pickle may be used to transfer Python objects over a network connection or save and load them to and from disc. This can be helpful for a variety of purposes, including caching pricey calculations, transferring data across programmes, and saving model parameters for later use.

```
with open('model.pkl', 'wb') as f:
    pickle.dump(model, f)
```

```
from google.colab import files
files.download('model.pkl')
```

Fig

The linear regression model for this project is recorded in pickle format, and its filename is 'model.pkl'.

4.8.7 STREAMLIT

A Python module called Streamlit makes it simple for programmers to build interactive web apps. Web application development is intended to be as straightforward as writing Python scripts.

You can easily develop web apps using Streamlit that let users interact with data, visualise data, and carry out different activities. It has several widgets that you can use to build sliders, drop-down menus, buttons, and other user interface elements.

It also works nicely with well-known Python data science tools like Scikit-learn, Matplotlib, and Pandas.

4.8.8 FLASK

Python programmers may create web apps using the popular Flask web framework. It is a wonderful option for both small and large-scale projects since it is a lightweight, modular framework that is simple to learn and use.

Making online apps that can manage user authentication, data storage and retrieval, and integration with other web services is simple using Flask. For designing routes, managing HTTP requests and answers, and generating templates, it offers a straightforward and clean syntax.

The extensibility and adaptability of Flask are two of its main benefits. It supports a large variety of extensions and plugins that let you give your application extra features like support for databases, authentication, and testing.

In addition to having a vibrant developer community, Flask is renowned for having a wealth of online resources accessible to aid in learning and using the framework. Building web apps in Python using Flask is an excellent option whether you're a novice or seasoned developer.

4.8.9 TENSORFLOW

Google created the open-source TensorFlow software library for numerical calculation utilising data flow graphs. Although it was primarily created for deep neural networks in machine learning applications, it may be applied to a wide range of numerical tasks.

TensorFlow is a Python library that offers an API for creating and developing machine learning models. Through its API, it also supports more programming languages, including C++, Java, and Go.

TensorFlow must be installed before it can be used in Python. Pip, a Python package installer, may be used for this.

4.9 ALGORITHM USED

LINEAR REGRESSION

The relationship between a dependent variable and one or more independent variables may be modelled statistically using linear regression. Finding the optimum linear equation to explain the connection between the variables is the goal of linear regression.

In predictive modelling, machine learning, and data analysis, linear regression is frequently used to estimate the value of the dependent variable based on the values of the independent variables. It is also employed to test hypotheses and establish the direction and strength of the association between the variables. And the model is saved in pickle format .

4.10 GOOGLE COLAB

Users may create and execute code in a free Jupyter notebook environment using Google Colab, a cloud-based platform. It is constructed on top of Google Drive, which gives users the option to save their journals and data files in the cloud.

Users may create and run Python code in a web browser using Google Colab without having to set up a local development environment. Additionally, Colab offers free usage of potent hardware, like GPUs and TPUs, which may be applied to projects like deep learning.

Colab notebooks are a helpful tool for collaboration and education since they can be quickly shared with others. Google Sheets, Google BigQuery, and Google Cloud Storage are among the various Google services that it interfaces with.

CHAPTER-5

RESULT AND DISCUSSION

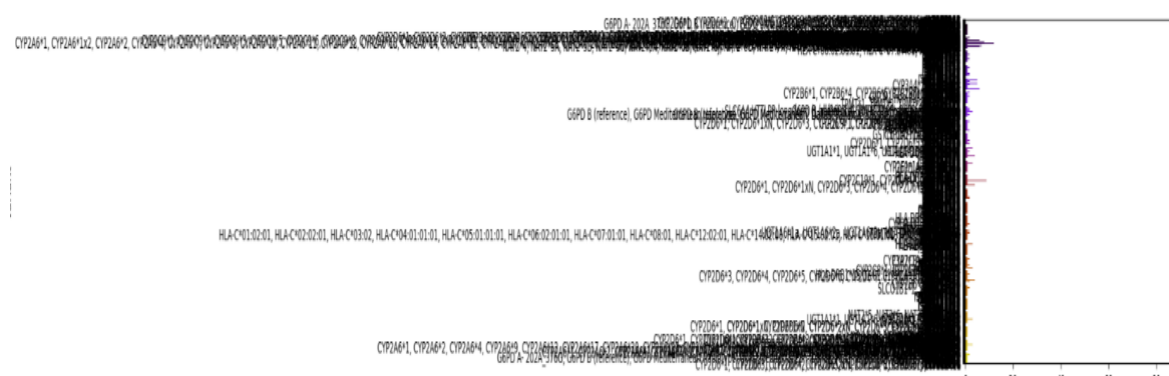
5.1 PRELIMINARY DATA ANALYSIS

Clinical variant data is chosen by analysing the features of all the datasets, from sets such as Variant and Clinical Annotations Data, Variant, Gene and Drug Relationship Data, Clinical Guideline Annotations Data, Drug Label Annotations Data, Pathways Data, Literature Occurrence Data, and Automated Annotations. This online application is best suited to the features of clinical variations.

5.2 GRAPHICAL REPRESENTATION OF THE FEATURE CATEGORIES(COUNT FROM THE DATASET AFTER CLEANING)

5.2.1 VARIANTS

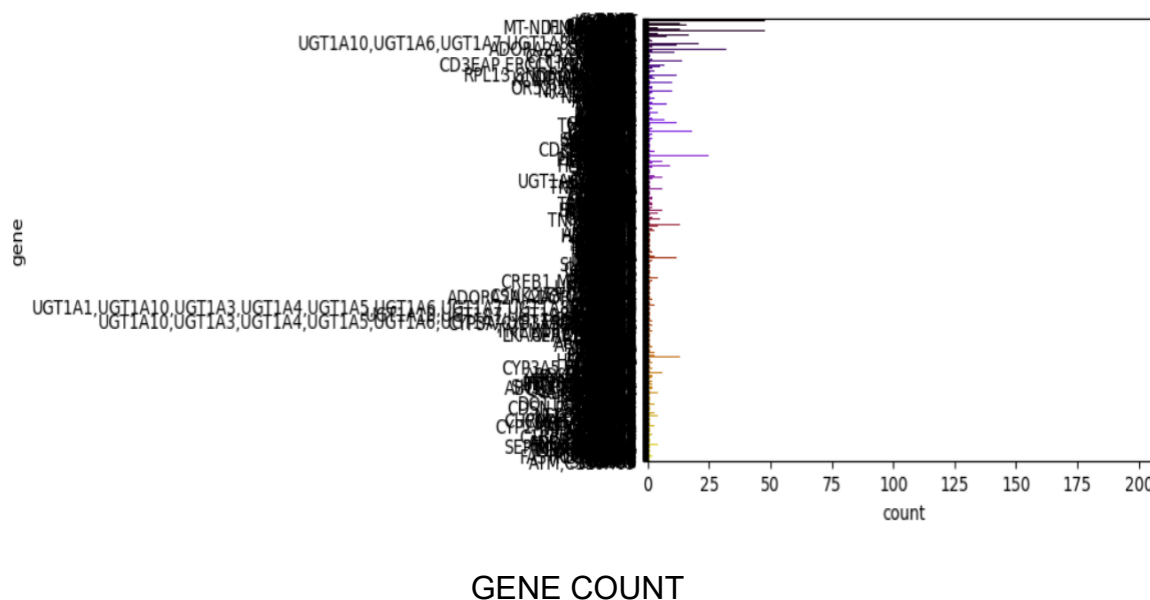
A variant feature is a distinctive trait of a genetic variation that has been discovered via scientific study and is connected to a specific impact on medication response or illness risk in the context of clinical variants from PharmGKB. This information on the variation's position within the genome, the sort of genetic mutation that took place (such as a deletion or single nucleotide substitution), and the functional effects of the variant on protein structure and function are some examples of variant characteristics. Variant characteristics are a crucial part of clinical variant annotation and may be used to guide therapeutic decisions, such as choosing the right medicine or dose based on a patient's genetic profile.



Variant count

5.2.2 GENE

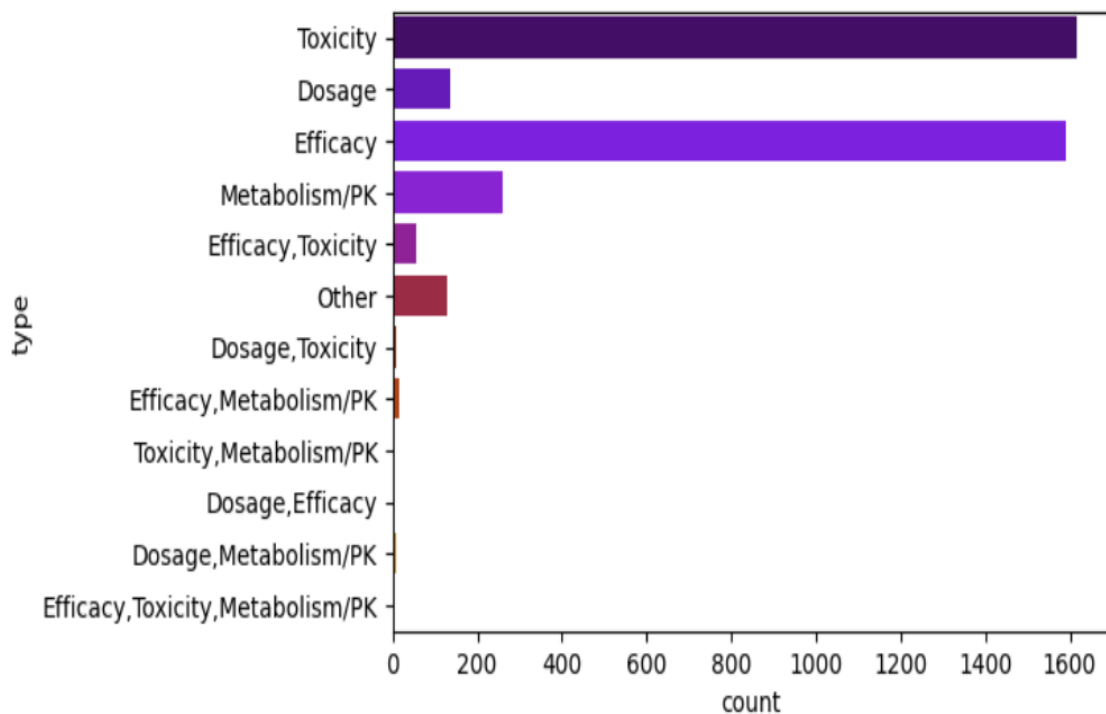
A gene feature is a distinctive trait of a gene that has been discovered via scientific study and is connected to a particular influence on medication responsiveness or illness risk in the context of clinical variations from PharmGKB. These gene attributes may include details on the gene's location within the genome, the purpose of the gene and the proteins it encodes, and the clinical importance of any genetic variations in the gene that are known or believed to exist. Gene features are a crucial part of clinical variant annotation and may be used to guide therapeutic decisions, such as choosing the best drugs or estimating the likelihood of adverse drug responses based on a person's genetic profile. Gene characteristics can also assist researchers in finding fresh medication targets and creating more specialised illness preventative and therapy strategies.



5.2.3 TYPE

A type feature is the unique kind of genetic variation that has been discovered in a particular genomic area or gene when discussing clinical variants from PharmGKB. These characteristics offer crucial details on the kind of genetic variation and its possible influence on illness risk or therapeutic response. Type characteristics are a crucial part of clinical variant annotation and may be used to help clinicians make clinical decisions, such as choosing the best drugs or estimating the likelihood of negative drug responses based on a patient's genetic profile. Type

traits can also assist researchers in finding fresh medication targets and creating more specialised illness preventative and treatment strategies.



TYPE COUNT

5.2.4 LEVEL OF EVIDENCE

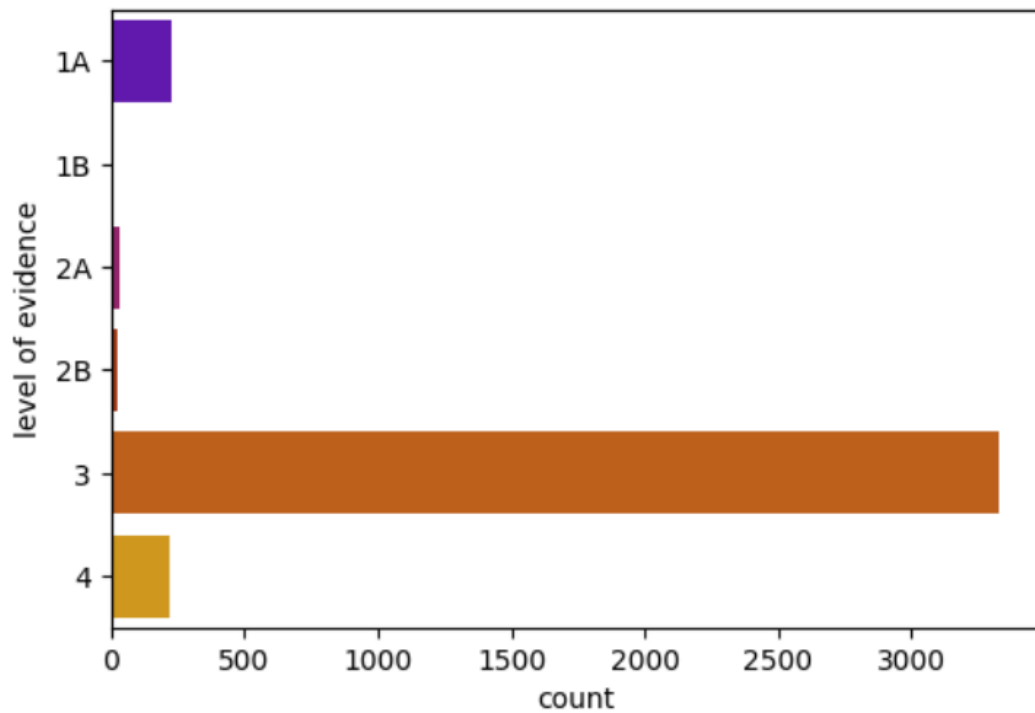
The level of evidence feature is an essential component of clinical variant annotation because it informs doctors and researchers about the quality of the evidence that links a genetic variant to its impact on illness risk or medication response. To better understand the clinical consequences of genetic variation, this information may be utilised to direct clinical decision-making and to pinpoint areas in need of further study.

Level 1: Validated: The link between the genetic variation and treatment response or disease risk is well supported by data from several studies

Level 2: Evidence: The link between the genetic variation and treatment response or disease risk is supported by some evidence from one or more studies.

Level 3: minimal: The link between the genetic variation and treatment response or disease risk has received minimal support from one or more studies.

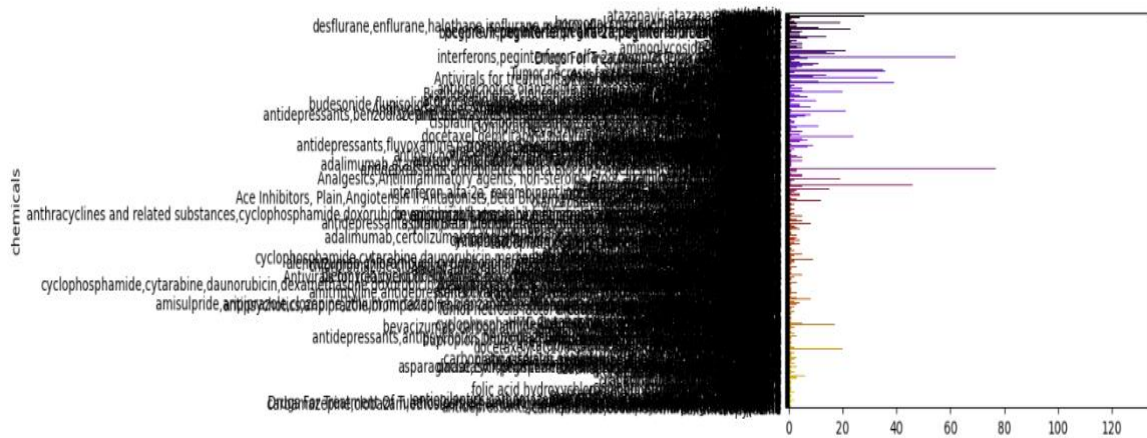
Level 4: No evidence The link between the genetic variation and treatment response or illness risk has not been shown by published research.



Level of evidence count

5.2.5 CHEMICALS

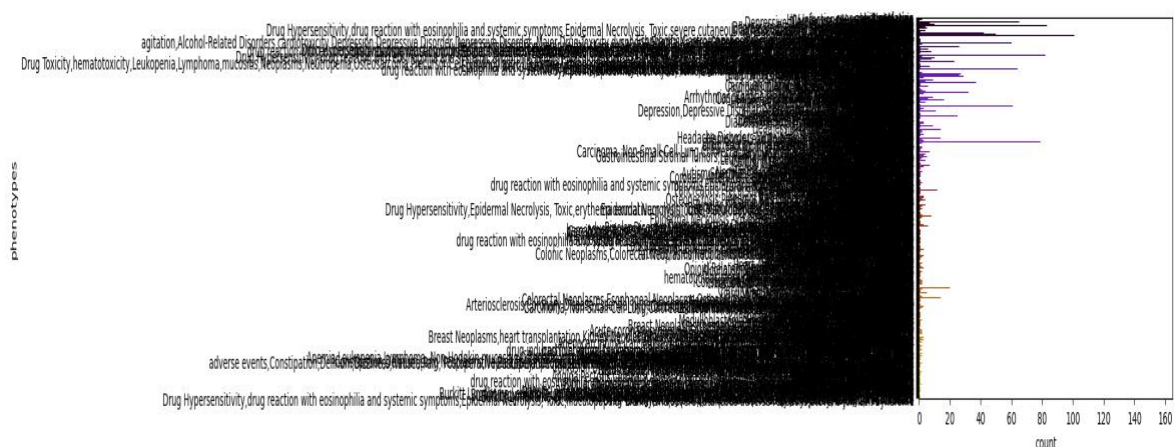
An individual trait of a medicine or other chemical component is referred to as a chemical feature. This trait may interact with a genetic variation to impact a drug's response or illness risk. Pharmacogenomics, the study of how genetic variation impacts an individual's reaction to medications, includes chemical characteristics as a key component. Chemical characteristics can be used to discover medications or substances that may be more or less effective for people with specific genetic variations and to assist forecast the possible influence of a genetic variant on drug response. For instance, a genetic variation may impact a drug's metabolism, slowing down how quickly it leaves the body and perhaps causing harmful adverse effects. Alternately, a genetic variation may have an impact on a drug's target protein, reducing its ability to effectively treat a particular condition.



CHEMICAL COUNT

5.2.6 PHENOTYPE

The visible traits or clinical manifestations of a genetic mutation in an individual are referred to as phenotypic features. Phenotype characteristics are a crucial part of clinical variant annotation because they reveal how a person's health or reaction to particular medications or treatments may be affected by genetic variation. Phenotype characteristics can be used to inform clinical judgement and to spot people who could be more susceptible to specific illnesses or negative medication responses. For instance, a genetic variation may be linked to a higher risk of contracting a certain form of cancer or a larger chance of having a particular kind of medication toxicity.



PHENOTYPE COUNT

5.3 PREPROCESSED DATA

For the purpose of doing a mathematical or statistical analysis on the data, we transform type objects to numeric objects. Values or variables with a numerical or quantitative representation, such as integers, floats, or decimals, are known as numerical objects. Calculations of many kinds, including addition, subtraction, multiplication, and division, as well as more involved statistical analysis, may be performed using these kinds of items.

	level of evidence_1A	level of evidence_1B	level of evidence_2A	level of evidence_2B	level of evidence_3	level of evidence_4	phenotypes_Abortion, Spontaneous	phenotypes_Ac Immunodeficiency
level of evidence_1A	1.000000	-0.012822	-0.024066	-0.019048	-0.640465	-0.061262	-0.004050	-0.000000
level of evidence_1B	-0.012822	1.000000	-0.004906	-0.003883	-0.130558	-0.012488	-0.000826	-0.000000
level of evidence_2A	-0.024066	-0.004906	1.000000	-0.007288	-0.245054	-0.023440	-0.001550	-0.000000
level of evidence_2B	-0.019048	-0.003883	-0.007288	1.000000	-0.193953	-0.018552	-0.001226	-0.000000
level of evidence_3	-0.640465	-0.130558	-0.245054	-0.193953	1.000000	-0.623805	0.006323	0.000000
...	...	...	...	...	...	...	...	...
variant_rs9973653	-0.004050	-0.000826	-0.001550	-0.001226	0.006323	-0.003944	-0.000261	-0.000000
variant_rs9977268	-0.004050	-0.000826	-0.001550	-0.001226	0.006323	-0.003944	-0.000261	-0.000000
variant_rs997917	-0.005728	-0.001168	-0.002192	-0.001735	0.008944	-0.005579	-0.000369	-0.000000
variant_rs9981861	-0.004050	-0.000826	-0.001550	-0.001226	0.006323	-0.003944	-0.000261	-0.000000

5.4 MODEL

5.4.1 SPLITTING THE DATA BASED ON INDEPENDENT AND DEPENDENT

Modelling the connection between the independent variables (predictors) and the dependent variable is the aim of a linear regression model. The independent variables serve as the model's inputs and as predictors of the dependent variable's value, which is the model's output. You must separate the independent factors from the dependent variables in your dataset in order to divide the data based on independent and dependent variables. The dependent variable is often kept in a

separate vector or array, frequently indicated by y, while the independent variables are typically kept in a matrix or dataframe, frequently designated by X. This makes it simple to feed the data into a linear regression model for prediction and training.

```
#splitting the data based on independent and dependent  
x=df.iloc[:, :-1].values  
y=df.iloc[:, 1].values
```

5.4.2 TEST AND TRAIN DATA

Data analysis and machine learning both include the division of the data into training and testing sets. The main objective of this is to assess how effectively the model generalises to fresh, untested data.

A machine learning model learns to suit the patterns in the training data when it is trained. As the model has already seen this data and has learned to fit it well, the performance may be overly optimistic if it is evaluated using the same data that it was trained on. However, because the model hasn't yet learned to fit the patterns in this data, it might not perform as well when given fresh, previously unexplored data.

We can predict how well the model will perform on fresh data by measuring its performance on the testing set. It is likely that the model will do well on fresh, untested data if it does well on the testing set. The model may have overfit the training data if it performs badly on the testing set and may not generalise effectively to new data.

```
#dividing the data based on testing and training dataset  
x_train,x_test,y_train,y_test=train_test_split(x,y,train_size=2/3,random_state=0)
```

5.4.3 LINEAR REGRESSION

A well-liked machine learning approach for tackling regression issues is linear regression. Modelling the connection between a dependent variable and one or more independent variables is the main goal of linear regression.

Finding the best-fit line or hyperplane that illustrates the connection between the independent factors and the dependent variable is the objective of linear regression. Then, predictions for fresh, unforeseen data can be made using this line or hyperplane.

An easy-to-understand procedure called linear regression can give important insights into how different variables relate to one another. It is a helpful tool for data exploration and visualisation and is frequently used as a starting point technique for more complicated machine learning models.

```
▶ #Implement the classifier on simple linear regression
simpleLinearRegression=LinearRegression()
simpleLinearRegression.fit(x_train,y_train)
```

```
↳ ▾ LinearRegression
   LinearRegression()
```

```
[32] y_predict=simpleLinearRegression.predict(x_test)
      y_predict

array([-5.72458747e-17, -3.24746914e-05, -5.47391233e-05, ...,
        5.33817067e-04,  2.23762234e-04,  2.82632405e-03])
```

5.4.4 ACCURACY AND ERROR

One popular statistic for assessing how well a machine learning model is doing is accuracy. It calculates the percentage of the model's total predictions that were accurate forecasts. It is essential to remember that accuracy might not always be the optimum criterion for assessing a machine learning model's performance.

```
▶ #calculate the accuracy
accuracy=simpleLinearRegression.score(x,y)
accuracy
```

0.9483078

Error in machine learning refers to the discrepancy between a model's expected and actual output. In machine learning, a variety of errors can happen, and each type of error offers a unique perspective on how well the model is working.

```
[34] #lets find out the error
      error_in_model=1-accuracy
      error_in_model
```

0.0516922

5.4.5 MSE AND MAE

The mean squared error (MSE), which calculates the average squared difference between the expected and actual outputs, is one prevalent form of error. In regression issues when the objective is to predict a continuous output variable, the MSE is frequently utilised. The model's predictions perform better the lower the MSE. The mean squared error indicates how close you are to discovering the line of best fit. Depending on your data, it might not be feasible to obtain a mean squared error number that is extremely minimal.

```
# Calculate root mean squared error
rmse = np.sqrt(mse)
```

```
print('Mean squared error:', mse)
```

Mean squared error: 0.0008153179054338989

The average absolute difference between the projected and actual outputs is measured by the mean absolute error (MAE), which is another sort of error. The MAE offers a measurement of the errors' size and is frequently utilised in regression issues. The model is more accurate the closer MAE is near 0. There isn't a set standard for what a good score is because MAE is returned on the same scale as the target you are projecting for.

```
▶ # Calculate mean squared error
mse = mean_squared_error(y_test, y_predict)
```

```
▶ print('Root mean squared error:', rmse)
```

Root mean squared error: 0.028553772175211788

5.4.6 STREAMLIT

Import the dataset and the model and then the predict response is defined and the url is developed.

```
[6] df = pd.read_csv('/content/clinicalVariants_csv.csv')
```

```
[7] import pickle

with open("/content/model.pkl", 'rb') as f:
    model = pickle.load(f)
```

```
[8] def predict_response(gene_variants, drug):
    features = [[gene_variants, drug]]
    response = model.predict(features)
    return response[0]
```

```
[9] if __name__ == '__predict response__':
    try:
        predict_response()

    except Exception as e:
        st.write("Error:", e)
```

Streamlit has developed url for both local system and for external networks.

```
!streamlit run --server.port 80 DRUG_GENE_RESPONSE_app.py &
```



Collecting usage statistics. To deactivate, set browser.gatherUsageStats to False

You can now view your Streamlit app in your browser.

Network URL: <http://172.28.0.12:80>

External URL: <http://34.141.200.18:80>

Define the configuration of the tunnel

```
# Define the configuration for the tunnel
config = {
    "name": "http-tunnel",
    "addr": "http://localhost:8000",
    "proto": "http",
    "port": 80
}
```

Open tunnel and keep it open

```
from pyngrok import ngrok

# Open a HTTP tunnel on port 80
ngrok_tunnel = ngrok.connect(80, "http")
```

WARNING:pyngrok.process.ngrok:t=2023-05-03T09:40:28+0000 lvl=warn msg="ngrok config file found at legacy location"

```
# Keep the tunnel open until the script is interrupted
ngrok_process = ngrok.get_ngrok_process()
try:
    ngrok_process.proc.wait()
except KeyboardInterrupt:
    pass
```

Public url for the tunnel

```
# Print the public URL of the tunnel
print("Public URL:", ngrok_tunnel.public_url)
```

Public URL: <https://c552-34-141-200-18.ngrok.io>

```
print("Public URL:", public_url)
```

Public URL: NgrokTunnel: "<https://90b1-34-141-200-18.ngrok.io>" -> "<http://localhost:80>"

5.5 WEB APPLICATION

Finally, a web application is created with a gene, a variant, and potentially a phenotype as inputs, and a projected medication type as the final output.

```
<input type="text" name="variant" placeholder="variant" required="required" /><br><br>
<input type="text" name="gene" placeholder="gene" required="required" /><br><br>
<input type="text" name="phenotype" placeholder="phenotype" required="required" /><br><br>

<button type="submit" class="btn btn-primary btn-block btn-large">Predict</button>
```

HTML code for inputs



CHAPTER-6

SUMMARY AND CONCLUSION

Pharmacogenomics is the study of how a person's genetic make-up influences how they react to medications. By adjusting pharmacological therapy to a person's genetic profile, personalised medicine has been made possible thanks to this branch of study. Now that next-generation sequencing technologies have been developed, it is possible to sequence a person's complete genome or just a few interesting locations in order to find mutations that might affect how well they respond to medication.

A database called PharmGKB (Pharmacogenomics Knowledgebase) collects data on genetic variations that are known to affect medication response. The medications that these genes affect are included in a database along with information on their pharmacokinetic and pharmacodynamic effects. A group of subject-matter specialists selected the data for PharmGKB, and it is often updated as new data becomes available.

A web application that makes use of the clinical variations data set from PharmGKB might be quite helpful in this situation. Researchers and physicians might swiftly search for information on certain variations and the medications they affect with the use of such an application. Additionally, based on a person's genetic profile, it might offer tailored recommendations for drug therapy.

The user would need to come first when designing the web application, and it would need to have an easy-to-use interface that made data searching and filtering simple. Additionally, it ought to include concise, clear information on each variant, the medications they affect, and any pertinent clinical recommendations.

Developing a pharmacogenomics online application may face difficulties in assuring the data's accuracy and completeness. To guarantee that users have access to the most recent information, the database must be updated when new information becomes available.

Making sure the programme is usable by a variety of users, including those with little technological expertise, is another problem. To solve this, the programme should be made with a user-friendly interface that doesn't require a lot of technical

knowledge. In conclusion, an online application that makes use of the clinical variations data set from PharmGKB has the potential to be a useful resource for academics and clinicians working in the field of pharmacogenomics. The application may assist to increase the effectiveness and safety of pharmacological therapy for specific individuals by giving access to the most recent data on genetic variations that affect drug response. The user should, however, be considered while designing the application, and the data should be accurate and current.

REFERENCE

1. Ahmed, N., Ahammed, R., Islam, M. M., Uddin, M. A., Akhter, A., Talukder, M. A., & Paul, B. K. Machine learning based diabetes prediction and development of smart web application. *International Journal of Cognitive Computing in Engineering*, 2, (2021, June). 229–241. <https://doi.org/10.1016/j.ijcce.2021.12.001>
2. Cacabelos, R., Cacabelos, N., & Carril, J. C. The role of pharmacogenomics in adverse drug reactions. *Expert Review of Clinical Pharmacology*, 12(5), (2019, April 24). 407–442. <https://doi.org/10.1080/17512433.2019.1597706>
3. Gong, L., Whirl-Carrillo, M., & Klein, T. E. PharmGKB, an Integrated Resource of Pharmacogenomic Knowledge. *Current Protocols*, 1(8). (2021, August). <https://doi.org/10.1002/cpz1.226>
4. Goodman, C. W., & Brett, A. S. Race and Pharmacogenomics—Personalized Medicine or Misguided Practice? *JAMA*, 325(7), (2021, February 16). 625. <https://doi.org/10.1001/jama.2020.25473>
5. Hakim, H., Bettinger, J. A., Chambers, C. T., Driedger, S. M., Dubé, E., Gavaruzzi, T., Giguere, A. M. C., Kavanagh, R., Leask, J., MacDonald, S. E., Orji, R., Parent, E., Paquette, J. S., Roberge, J., Sander, B., Scherer, A. M., Tremblay-Breault, M., Wilson, K., Reinhartz, D., & Witteman, H. O. A Web Application About Herd Immunity Using Personalized Avatars: Development Study. *Journal of Medical Internet Research*, 22(10), (2020, October 30). e20113. <https://doi.org/10.2196/20113>
6. Jain, V., Kavitha, H., & Mohana Kumar, S. Credit Card Fraud Detection Web Application using Streamlit and Machine Learning. 2022 IEEE International Conference on Data Science and Information System (ICDSIS). (2022, July 29). <https://doi.org/10.1109/icdsis55133.2022.9915901>
7. Jiang, T., Gradus, J. L., & Rosellini, A. J. Supervised Machine Learning: A Brief Primer. *Behavior Therapy*, 51(5), (2020, September). 675–687. <https://doi.org/10.1016/j.beth.2020.05.002>
8. Lauschke, V. M., & Ingelman-Sundberg, M. Emerging strategies to bridge the gap between pharmacogenomic research and its clinical implementation. *Npj Genomic Medicine*, 5(1). (2020, March 5). <https://doi.org/10.1038/s41525-020-0119-2>

9. Li, Y., Ge, D., & Lu, C. The SMART App: an interactive web application for comprehensive DNA methylation analysis and visualization. *Epigenetics & Chromatin*, 12(1). (2019, Dec). <https://doi.org/10.1186/s13072-019-0316-3>
10. Lin, E., Lin, C. H., & Lane, H. Y. Machine Learning and Deep Learning for the Pharmacogenomics of Antidepressant Treatments. *Clinical Psychopharmacology and Neuroscience*, 19(4), (2021, November 30). 577–588. <https://doi.org/10.9758/cpn.2021.19.4.577>
11. Mehmood, A., Nawab, S., Jin, Y., Hassan, H., Kaushik, A. C., & Wei, D. Q. Ranking Breast Cancer Drugs and Biomarkers Identification Using Machine Learning and Pharmacogenomics. *ACS Pharmacology & Translational Science*, 6(3), (2023, February 24). 399–409. <https://doi.org/10.1021/acsptsci.2c00212>
12. Morganti, S., Tarantino, P., Ferraro, E., D'Amico, P., Duso, B. A., & Curigliano, G. Next Generation Sequencing (NGS): A Revolutionary Technology in Pharmacogenomics and Personalized Medicine in Cancer. *Translational Research and Onco-Omics Applications in the Era of Cancer Personal Genomics*, (2019). 9–30. https://doi.org/10.1007/978-3-030-24100-1_2
13. Mukhiya, S. K., & Ahmed, U. *Hands-On Exploratory Data Analysis with Python: Perform EDA Techniques to Understand, Summarize, and Investigate Your Data*. (2020, March 27).
14. Pandi, M. T., Koromina, M., Tsafaridis, I., Patsilnakos, S., Christoforou, E., van der Spek, P. J., & Patrinos, G. P. A novel machine learning-based approach for the computational functional assessment of pharmacogenomic variants. *Human Genomics*, 15(1). (2021, August 9). <https://doi.org/10.1186/s40246-021-00352-1>
15. Pirmohamed, M. Pharmacogenomics: current status and future perspectives. *Nature Reviews Genetics*. (2023, January 27). <https://doi.org/10.1038/s41576-022-00572-8>
16. Qin, W., Lu, X., Shu, Q., Duan, H., & Li, H. Building an information system to facilitate pharmacogenomics clinical translation with clinical decision support. *Pharmacogenomics*, 23(1), (2022, January). 35–48. <https://doi.org/10.2217/pgs-2021-0110>
17. Razeghi, O., Solís-Lemus, J. A., Lee, A. W., Karim, R., Corrado, C., Roney, C. H., de Vecchi, A., & Niederer, S. A. CemrgApp: An interactive medical imaging application with image processing, computer vision, and machine learning

- toolkits for cardiovascular research. *SoftwareX*, 12, (2020, July). 100570. <https://doi.org/10.1016/j.softx.2020.100570>
18. Saia, S. M., Nelson, N. G., Young, S. N., Parham, S., & Vandegrift, M. Ten simple rules for researchers who want to develop web apps. *PLOS Computational Biology*, 18(1), (2022, January 6). e1009663. <https://doi.org/10.1371/journal.pcbi.1009663>
 19. Sarker, I. H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, 2(3). (2021, March 22). <https://doi.org/10.1007/s42979-021-00592-x>
 20. Sarmah, R. J., Ding, Y., Wang, D., Lee, C. Y. P., Li, T. J. J., & Chen, X. A. Geno. *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. (2020, October 20). <https://doi.org/10.1145/3379337.3415848>
 21. Schreeck, F., Ahne, G., Tremmel, R., Schaeffeler, E., & Schwab, M. Pharmacogenomics in pediatric medicine and drug development. *Pharmacogenomics*, 23(13), (2022, August). 709–712. <https://doi.org/10.2217/pgs-2022-0105>
 22. Shah, D., Patel, S., & Bharti, S. K. Heart Disease Prediction using Machine Learning Techniques. *SN Computer Science*, 1(6). (2020, October 16). <https://doi.org/10.1007/s42979-020-00365-y>
 23. Sharifi-Noghabi, H., Jahangiri-Tazehkand, S., Smirnov, P., Hon, C., Mammoliti, A., Nair, S. K., Mer, A. S., Ester, M., & Haibe-Kains, B. Drug sensitivity prediction from cell line-based pharmacogenomics data: guidelines for developing machine learning models. *Briefings in Bioinformatics*, 22(6). (2021, August 11). <https://doi.org/10.1093/bib/bbab294>
 24. Sidey-Gibbons, J. A. M., & Sidey-Gibbons, C. J. Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*, 19(1). (2019, March 19). <https://doi.org/10.1186/s12874-019-0681-4>
 25. Takahashi, T., Luzum, J. A., Nicol, M. R., & Jacobson, P. A. Pharmacogenomics of COVID-19 therapies. *Npj Genomic Medicine*, 5(1). (2020, August 18). <https://doi.org/10.1038/s41525-020-00143-y>
 26. Taylor, C., Crosby, I., Yip, V., Maguire, P., Pirmohamed, M., & Turner, R. M. A Review of the Important Role of CYP2D6 in Pharmacogenomics. *Genes*, 11(11), (2020, October 30). 1295. <https://doi.org/10.3390/genes11111295>

27. Truong, T. M., Lipschultz, E., Schierer, E., Danahey, K., Ratain, M. J., & O'Donnell, P. H. Patient insights on features of an effective pharmacogenomics patient portal. *Pharmacogenetics and Genomics*, 30(9), (2020, July 7). 191–200. <https://doi.org/10.1097/fpc.0000000000000413>
28. Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, 19(1). (2019, December). <https://doi.org/10.1186/s12911-019-1004-8>
29. Whirl-Carrillo, M., Huddart, R., Gong, L., Sangkuhl, K., Thorn, C. F., Whaley, R., & Klein, T. E. An Evidence-Based Framework for Evaluating Pharmacogenomics Knowledge for Personalized Medicine. *Clinical Pharmacology & Therapeutics*, 110(3), (2021, July 22). 563–572. <https://doi.org/10.1002/cpt.2350>
30. Yu, M. K., Ma, J., Fisher, J., Kreisberg, J. F., Raphael, B. J., & Ideker, T. Visible Machine Learning for Biomedicine. *Cell*, 173(7), (2018, June). 1562–1565. <https://doi.org/10.1016/j.cell.2018.05.056>
31. Zhou, Y., Tremmel, R., Schaeffeler, E., Schwab, M., & Lauschke, V. M. Challenges and opportunities associated with rare-variant pharmacogenomics. *Trends in Pharmacological Sciences*, 43(10), (2022, October). 852–865. <https://doi.org/10.1016/j.tips.2022.07.002>
32. Zhu, Q., Tao, C., Shen, F., & Chute, C. G. Exploring The Pharmacogenomics Knowledge Base (Pharmgkb) For Repositioning Breast Cancer Drugs By Leveraging Web Ontology Language (Owl) And Cheminformatics Approaches. *Biocomputing* 2014. (2013, November). https://doi.org/10.1142/9789814583220_0017
33. Zhu, W., Xie, L., Han, J., & Guo, X. The Application of Deep Learning in Cancer Prognosis Prediction. *Cancers*, 12(3), (2020, March 5). 603. <https://doi.org/10.3390/cancers12030603>