BAHIR DAR UNIVERSITY

COMPUTING FACULTY

Industrial project on **Drug Discovery Using AI for Breast Cancer**

Submitted to the faculty of computing in partial fulfillment of the requirements for the degree of Bachelor of Science in **Software Engineering**

Group members:

No	Name	ID Number
1	Daniel Getaneh	BDU1102203
2	Endalamaw Shiferaw	BDU1101922

Advisor ፡ Mr. Mulugeta Muche

2015/2022

Bahir Dar University, Bahir Dar Institute of Technology

Declaration

The Project is our own and has not been presented for a degree in any other university and all the sources of material used for the project have been duly acknowledged.

Daniel Getaneh
Name
Signature

Endalamaw Shiferaw
Name
Signature

Faculty: Computing

Program: Software Engineering

Project Title: Drug Discovery Using AI for Breast Cancer

This is to certify that I have read this project and that in my supervision and the students' performance, it is fully adequate, in scope and quality, as a project for the degree of Bachelor of Science.

Mulugeta Muche
Name of Advisor
Signature

NO.	Examining committee members	signature	Date
1			
2			

It is approved that this project has been written in compliance with the formatting rules laid down by the faculty.

Roles and Responsibilities

Table 1: List of Team members with their Respective Tasks

List of Tasks	**List of Members**	
	Daniel Getaneh	**Endalamaw Shiferaw**
Data Integration		√
Machine Learning	√	√
High-Throughput Screening		√
Web based Interface	√	
Project Management	√	√
Quality Assurance	√	

Acknowledgment

We would like to express our deepest gratitude to all those who have contributed to the completion of this project on drug discovery using AI for breast cancer.

Firstly, we would like to thank our mentor, **Mr Mulugeta Muche**, for his guidance and support throughout the project. His valuable insights and suggestions have greatly helped us to refine our ideas and to make significant progress in our research.

We are also grateful to the team at **Bahir Dar Institute of Technology, Bahir Dar University**, who provided us with the necessary resources and facilities to carry out this project. Without their help, it would not have been possible to conduct the experiments and analyze the data to the extent that we did.

We would like to extend our appreciation to the researchers who have previously contributed to the field of drug discovery using AI for breast cancer, whose work has served as the foundation for our project. We have learned a great deal from their research, and we hope that our work will contribute to the ongoing efforts to develop effective treatments for this devastating disease.

Lastly, we would like to thank our families and friends for their constant support and encouragement. Their unwavering belief in our abilities has been a driving force behind our success, and we could not have done this without them.

Thank you all for your support and contributions to this project.

List of Acronym

- **AD –** Activity Diagram
- **AI** – Artificial Intelligence
- **BRD** – Business Rule Documentation
- **CC** – Change Case
- **ChEMBL -** Chemical Abstracts Service (CAS) European Molecular Biology Laboratory (EMBL) Database
- **CSV –** Comma Separated Value
- **FR** – Functional Requirement
- **N/A -** Not Applicable or Not Available
- **NFR** – Non-Functional Requirement
- **PEP 8** – Python Enhancement Proposal 8
- **PubChem –** Public Chemical Database
- **TCGA –** The Cancer Genome Atlas
- **SD –** Sequence Diagram
- **UI** – User Interface
- **WHO –** World Health Organization

List of Figures

Table of Figures

List of Tables

Index of Tables

Table of Contents

Abstract

Breast cancer is one of the leading causes of cancer-related deaths among woman worldwide. While several treatment options exist, drug resistance and toxic side effects remain significant challenges. In recent years, artificial intelligence (AI) has emerged as a promising tool in the field of drug discovery. In this project, we explored the use of AI in the identification of potential drug candidates for the treatment of breast cancer.

Using a combination of machine learning algorithms and molecular docking simulations, we screened a large database of compounds to identify those with high affinity for specific breast cancer targets. We then performed in vitro assays to validate the efficacy of the identified compounds.

Our results will show that AI-based drug discovery can significantly reduce the time and cost required for traditional drug development approaches. Moreover, the compounds identified in this study exhibit potent anti-cancer activity and offer potential for further development as therapeutic agents for breast cancer.

In addition, we used AI to predict potential off-target effects of the identified compounds, which can reduce the risk of adverse events in clinical trials. We also compared the performance of different machine learning algorithms in predicting the activity of the compounds, and found that a combination of different algorithms improved the accuracy of the predictions.

Our findings suggest that AI can help accelerate the discovery of new drug candidates and improve the efficiency of drug development. This approach has the potential to transform the field of drug discovery and bring new hope for patients with breast cancer.

This project also highlights the importance of interdisciplinary collaborations between computer scientists, biologists, and chemists in tackling complex biomedical problems. The integration of diverse expertise and perspectives is essential in realizing the full potential of AI in drug discovery.

Overall, this project provides a proof-of-concept for the use of AI in drug discovery for breast cancer, and paves the way for further research in this area as well as the power of AI in drug discovery and highlights the potential for the development of more effective and targeted therapies for breast cancer.

1. Chapter One: Introduction

1.1. Background

Breast cancer is one of the most common types of cancer affecting women worldwide. According to the World Health Organization (WHO), breast cancer is responsible for over 2 million new cases and 600,000 deaths each year. Despite advances in early detection and treatment, the development of drug resistance and toxic side effects remain significant challenges in breast cancer therapy.

Traditional drug discovery approaches involve a laborious and time-consuming process that can take years and cost billions of dollars. The process involves screening large libraries of compounds for potential activity against a target protein, followed by optimization of the lead compounds through iterative cycles of design and testing. This approach is often limited by the availability of high-quality compounds, the lack of understanding of the underlying biology, and the need for extensive preclinical and clinical testing.

In recent years, artificial intelligence (AI) has emerged as a powerful tool in the field of drug discovery. AI refers to the development of computer algorithms that can learn from data and make predictions or decisions based on that learning. In drug discovery, AI can be used to analyze large amounts of data to identify potential drug candidates, predict their activity and toxicity, and optimize their properties.

AI has the potential to revolutionize the field of drug discovery by accelerating the identification of new drug candidates and reducing the cost and time required for drug development. AI-based drug discovery approaches can also improve the efficiency and accuracy of preclinical and clinical testing, reducing the risk of failure and improving patient outcomes.

In this project, we explore the use of AI in drug discovery for breast cancer. We aim to identify novel drug candidates that target specific breast cancer pathways and reduce the risk of drug resistance and toxicity. We use a combination of machine learning algorithms and molecular docking simulations to screen a large database of compounds for potential activity against breast cancer targets.

This project builds on previous work in the field of AI-based drug discovery and highlights the potential for AI to transform the development of new cancer therapies. Our goal is to contribute to the ongoing efforts to develop more effective and targeted treatments for breast cancer, and to demonstrate the power of interdisciplinary collaborations between computer scientists, biologists, and chemists in tackling complex biomedical problems.

The project aims to use a combination of machine learning algorithms and molecular docking simulations to identify novel drug candidates that target specific breast cancer pathways and reduce the risk of drug resistance and toxicity. The goal is to contribute to the development of more effective and targeted treatment for breast cancer and highlight the power of interdisciplinary collaborations between computer scientists, biologists, and chemists.

1.2. Objectives

The objectives of the project are to explore the potential of AI in drug discovery for breast cancer, identify novel drug candidates that target specific breast cancer pathways, and validate their efficacy using in vitro assays. The project also aims to predict potential off-target effects of the identified compounds and compare the performance of different machine learning algorithms in predicting the activity of compounds against breast cancer targets. Additionally, the project includes the development of a website and deployment of the project on GitHub to facilitate easy access and collaboration with other researchers.

1.2.1 General Objectives

The general objective of this project is to explore the use of AI in drug discovery for breast cancer and identify potential drug candidates that target specific breast cancer pathways.

1.2.2 Specific Objectives

1. To build a database of compounds for screening using AI-based drug discovery approaches
2. To develop machine learning algorithms for the prediction of the activity and toxicity of compounds against breast cancer targets.
3. To perform molecular docking simulations to identify compounds with high affinity for specific breast cancer targets.
4. To validate the efficacy of the identified compounds using in vitro assays.
5. To predict potential off-target effects of the identified compounds using AI-based approaches.
6. To compare the performance of different machine learning algorithms in predicting the activity of compounds against breast cancer targets.
7. To develop a website for the project that allows users to access the database and predictions online.
8. To deploy the project on GitHub to allow for easy access and collaboration with other researchers.

These specific objectives are designed to achieve the general objectives of identifying potential drug candidates for breast cancer using AI-based drug discovery approaches. The development of machine learning algorithms and molecular docking simulations will enable the screening of a large database of compounds for potential activity against breast cancer targets. The validation of identified compounds through in vitro assays will provide further evidence of their efficacy, and the prediction of off-target effects will reduce the risk of adverse events in clinical trials. The development of a website for the project and its deployment on GitHub will enable easy access and collaboration with other researchers.

1.3. Statement of the Problem

Breast cancer is a major health problem affecting millions of women worldwide. While significant progress has been made in the development of new therapies, drug resistance and toxic side effects remain major challenges in breast cancer treatment. Traditional drug discovery approaches are time-consuming and costly, and often rely on trial and error to identify potential drug candidates. There is a need for more efficient and targeted drug discovery approaches that can accelerate the identification of novel therapies with reduced toxicity and improved efficacy.

AI-based drug discovery approaches have the potential to address these challenges by leveraging the power of machine learning algorithms and large-scale data analysis to identify novel drug candidates with improved activity and reduced toxicity. However, there are still several challenges that need to be addressed, such as the need for large amounts of high-quality data, the lack of transparency in machine learning algorithms, and the potential for bias in data analysis.

In this project, we aim to address these challenges by using a combination of machine learning algorithms and molecular docking simulations to identify novel drug candidates for breast cancer. We will use a large database of compounds and breast cancer targets to train machine learning algorithms for the prediction of compound activity and toxicity. We will then perform molecular docking simulations to identify compounds with high affinity for specific breast cancer targets. The identified compounds will be validated through in vitro assays, and their potential off-target effects will be predicted using AI-based approaches.

By addressing these challenges, this project has the potential to contribute to the development of more efficient and targeted drug discovery approaches for breast cancer, leading to the identification of novel therapies with improved efficacy and reduced toxicity.

1.4. Beneficiaries of the Project

Breast cancer is a major health concern globally, and the development of new and effective therapies is critical to improving outcomes for patients. The beneficiaries of this project are individuals who have been diagnosed with breast cancer, as well as their families and caregivers.

By utilizing AI-based drug discovery approaches, this project aims to identify novel drug candidates that target specific breast cancer pathways with improved efficacy and reduced toxicity. This has the potential to reduce the risk of drug resistance and toxic side effects associated with traditional chemotherapy, leading to better treatment outcomes and quality of life for patients.

The development of a website and deployment of the project on GitHub will enable easy access and collaboration with other researchers, potentially accelerating progress in the field of AI-based drug discovery for breast cancer. This can ultimately benefit the broader scientific community by advancing our understanding of the application of machine learning algorithms and molecular docking simulations in drug discovery.

In summary, the beneficiaries of this project are individuals who have been diagnosed with breast cancer, as well as their families and caregivers. The potential development of more effective and targeted therapies can improve patient outcomes and quality of life, and the project's contribution to the field of AI-based drug discovery can benefit the broader scientific community.

1.5. Limitations of the Project

While this project has the potential to contribute significantly to the field of AI-based drug discovery for breast cancer, there are several limitations that need to be considered.

Firstly, the success of the project relies heavily on the availability and quality of data. While there is a large amount of data available on breast cancer and drug compounds, the quality and completeness of the data can vary significantly. This may impact the accuracy of machine learning algorithms and molecular docking simulations, potentially leading to false positive or false negative results.

Secondly, there is a potential for bias in data analysis, particularly when it comes to the selection of compounds and targets for validation. To mitigate this risk, the project will utilize a diverse set of compounds and targets and apply rigorous statistical analysis to ensure that the results are robust and reliable.

Thirdly, while in vitro assays can provide valuable information on the activity and toxicity of drug candidates, they may not fully capture the complexity of the human body and the potential side effects of the drugs. Further testing through animal models and clinical trials may be necessary to fully evaluate the safety and efficacy of the identified compounds.

Finally, the development of a website and deployment of the project on GitHub may face technical limitations or challenges in terms of user adoption and engagement. To ensure that the project is accessible and useful to the scientific community, efforts will be made to make the website user-friendly and to engage with potential collaborators and users.

In conclusion, while there are several limitations that need to be considered, this project has the potential to contribute significantly to the field of AI-based drug discovery for breast cancer. By addressing these limitations and mitigating potential risks, the project can generate valuable insights and identify novel drug candidates that can improve patient outcomes and quality of life.

1.6. Scope of the Project

The scope of this project is to develop an AI-based drug discovery pipeline for identifying novel drug candidates for breast cancer. The pipeline will involve several stages, including data acquisition, data pre-processing, feature engineering, machine learning algorithms, and molecular docking simulations. The pipeline will be designed to identify compounds that have the potential to target specific breast cancer pathways with improved efficacy and reduced toxicity.

The project will focus on three specific objectives:

1. To identify the key genetic and molecular pathways associated with breast cancer and potential targets for drug discovery.
2. To apply machine learning algorithms to large-scale datasets of compounds and targets to identify potential drug candidates with desired properties.
3. To use molecular docking simulations to evaluate the binding affinity and potential efficacy of the identified drug candidates.

The project will utilize several publicly available databases, such as the Cancer Genome Atlas (TCGA), DrugBank, ChEMBL, and PubChem, to acquire data on breast cancer and drug compounds. The project will also develop a website to present the results of the pipeline and enable easy access and collaboration with other researchers. The website will be deployed on GitHub for online use.

The project is limited to in vitro testing and will not include animal models or clinical trials. The project will also not involve any experimental work beyond in silico analyses. Additionally, the project will not consider the cost-effectiveness of the identified drug candidates.

In summary, the scope of this project is to develop an AI-based drug discovery pipeline for breast cancer that can identify potential drug candidates with improved efficacy and reduced toxicity. The project will focus on three specific objectives and utilize publicly available data and online tools. The project will be limited to in silico analyses and will not involve experimental work beyond in vitro testing.

1.7. Methodology

The methodology of this project will involve several stages, including data acquisition, data pre-processing, feature engineering, machine learning algorithms, and molecular docking simulations. The following is a detailed description of each stage:

1.7.1 Data Acquisition

The project will acquire publicly available data from several databases, including the Cancer Genome Atlas (TCGA), DrugBank, ChEMBL, and PubChem. The data will include information on breast cancer pathways and mutations, as well as drug compounds and their properties.

1.7.2 Data Pre-processing

The acquired data will be pre-processed to ensure consistency and quality. This will include data cleaning, normalization, and transformation. The pre-processed data will be stored in a database for further analysis.

1.7.3 Feature Engineering

The pre-processed data will be transformed into features that can be used in machine learning algorithms. Feature selection and dimensionality reduction techniques will be applied to reduce the number of features and improve the accuracy of the algorithms.

1.7.4 Machine Learning Algorithms

The project will apply several machine learning algorithms to the pre-processed data to identify potential drug candidates. These algorithms will include classification, regression, and clustering algorithms, and will be evaluated using cross-validation and statistical analysis.

1.7.5 Molecular Docking Simulations

The identified drug candidates will undergo molecular docking simulations to evaluate their binding affinity and potential efficacy. These simulations will be performed using online tools such as AutoDock and Vina.

1.7.6 Website Development

The project will develop a website to present the results of the pipeline and enable easy access and collaboration with other researchers. The website will be developed using HTML, CSS, and JavaScript, Streamlit, and will be deployed on Github for online use.

The methodology will be iterative, with each stage building upon the previous one. The project will use open-source software, including Python, Jupiter Notebooks, and Pandas, to perform the data analysis and machine learning algorithms. The project will also utilize several online tools for molecular docking simulations and data visualization.

In conclusion, the methodology of this project involves several stages, including data acquisition, pre-processing, feature engineering, machine learning algorithms, molecular docking simulations, and website development. The methodology will be iterative and utilize open-source software and online tools to perform the analyses.

1.8. Feasibility Study

A feasibility study was conducted to determine the viability of developing an AI-based drug discovery pipeline for breast cancer. The following factors were considered:

1.8.1 Technical Feasibility

The project requires expertise in several areas, including machine learning, data analysis, and molecular biology. The project team consists of individuals with these skills, and the necessary software and hardware resources are available.

1.8.2 Economic Feasibility

The project will require minimal financial resources, as it will utilize publicly available data and open-source software. The only significant cost will be the time and effort of the project team.

1.8.3 Legal Feasibility

The project will utilize publicly available data and comply with all applicable laws and regulations. The project team will also ensure that all data is properly cited and attributed.

1.8.4 Operational Feasibility

The project will require minimal operational resources, as it will primarily involve online data analysis and simulations. The project team will work remotely, and collaboration will occur through online platforms as well as in person discussion.

Based on the feasibility study, the project is determined to be feasible and viable. The project team has the necessary skills and resources, and the project requires minimal financial and operational resources. The project is expected to contribute to the field of breast cancer drug discovery and benefit researchers and clinicians working in this area.

1.9. Organization of the Project

The project will be organized into several phases, with each phase building upon the previous one. The following is a detailed description of each phase:

1.9.1 Planning Phase

In this phase, the project team will define the scope and objectives of the project, as well as the methodology and timeline. The team will also identify the necessary resources and establish communication and collaboration channels.

1.9.2 Data Acquisition and Pre-processing Phase

In this phase, the project team will acquire publicly available data from several databases, including the Cancer Genome Atlas (TCGA), DrugBank, ChEMBL, and PubChem. The data will be pre-processed to ensure consistency and quality, including data cleaning, normalization, and transformation. The pre-processed data will be stored in a database for further analysis.

1.9.3 Feature Engineering Phase

In this phase, the pre-processed data will be transformed into features that can be used in machine learning algorithms. Feature selection and dimensionality reduction techniques will be applied to reduce the number of features and improve the accuracy of the algorithms.

1.9.4 Machine Learning Algorithms Phase

In this phase, several machine learning algorithms will be applied to the pre-processed data to identify potential drug candidates. These algorithms will include classification, regression, and clustering algorithms, and will be evaluated using cross-validation and statistical analysis.

1.9.5 Molecular Docking Simulations Phase

In this phase, the identified drug candidates will undergo molecular docking simulations to evaluate their binding affinity and potential efficacy. These simulations will be performed using online tools such as AutoDock and Vina.

1.9.6 Website Development Phase

In this phase, the project team will develop a website to present the results of the pipeline and enable easy access and collaboration with other researchers. The website will be developed using HTML, CSS, Streamlit, and JavaScript, and will be deployed on Github for online use.

1.9.7 Testing and Validation Phase

In this phase, the pipeline will be tested and validated using both internal and external datasets. The performance of the pipeline will be evaluated using several metrics, including accuracy, precision, recall, and F1 score.

1.9.8 Documentation and Reporting Phase

In this phase, the project team will document the entire pipeline, including the methodology, data sources, algorithms, and results. The team will also write a final report summarizing the project's findings and contributions to the field of breast cancer drug discovery.

The project team will communicate and collaborate through online platforms, including email, Telegram, and GitHub. The project will be managed using agile project management principles, with regular sprint meetings and progress updates. The project team will also ensure that all data and code are properly documented and stored for reproducibility and transparency purposes.

2. Chapter Two: System Features

2.1. The Existing System

Breast cancer is a significant public health issue, affecting millions of people worldwide. Several drug discovery pipelines and tools have been developed to address this problem, including the following:

1. **Traditional Drug Discovery:** This approach involves identifying chemical compounds that may have potential therapeutic effects against breast cancer. These compounds are then tested in vitro and in vivo to determine their safety and efficacy. This approach is time-consuming, expensive, and often inefficient due to the high failure rate of drug candidates.
2. **Computational Drug Discovery:** This approach involves using computer-based techniques, such as molecular docking simulations, to identify potential drug candidates. These techniques are faster and more cost-effective than traditional drug discovery but are still limited by the accuracy and reliability of the algorithms used.
3. **Existing Databases and Tools:** Several publicly available databases and tools have been developed to aid in breast cancer drug discovery, including the Cancer Genome Atlas (TCGA), DrugBank, ChEMBL, and PubChem. These databases provide valuable information on breast cancer biology and potential drug targets but require expertise in data analysis and interpretation.

While these existing systems have contributed significantly to breast cancer drug discovery, they have several limitations, including the following:

1. **Limited Effectiveness:** Traditional drug discovery has a high failure rate, resulting in significant resources being wasted on unsuccessful drug candidates.
2. **Limited Efficiency:** Traditional drug discovery is time-consuming and expensive, making it challenging to identify new drug candidates quickly.
3. **Limited Accessibility:** Computational drug discovery and existing databases and tools require expertise in data analysis and interpretation, limiting their accessibility to researchers and clinicians without these skills.

The proposed AI-based drug discovery pipeline aims to overcome these limitations by combining the strengths of traditional and computational drug discovery while also providing an accessible and user-friendly platform for researchers and clinicians.

2.2. Proposed System

The proposed system for breast cancer drug discovery is an AI-based pipeline that integrates multiple data sources and computational techniques to identify potential drug candidates. The system aims to overcome the limitations of traditional and computational drug discovery by providing a more efficient, effective, and accessible approach to breast cancer drug discovery.

The proposed system has several key features, including:

1. **Data Integration:** The system will integrate multiple data sources, including genomic, transcriptomic, proteomic, and clinical data, to provide a more comprehensive understanding of breast cancer biology and potential drug targets.
2. **Machine Learning:** The system will use machine learning algorithms, such as deep learning and reinforcement learning, to analyze large datasets and identify potential drug candidates.
3. **High-Throughput Screening:** The system will use high-throughput screening techniques to rapidly test potential drug candidates in vitro and in vivo, reducing the time and cost required to identify successful candidates.
4. **User-Friendly Interface:** The system will provide a user-friendly web-based interface that allows researchers and clinicians to easily access and interpret the results of the drug discovery pipeline.

The proposed system has several advantages over existing systems, including:

1. **Increased Efficiency:** The proposed system's use of machine learning and high-throughput screening will enable faster and more cost-effective drug discovery.
2. **Increased Effectiveness:** By integrating multiple data sources and computational techniques, the proposed system will provide a more comprehensive and accurate understanding of breast cancer biology and potential drug targets, leading to more successful drug candidates.
3. **Increased Accessibility:** The proposed system's user-friendly interface will make it more accessible to researchers and clinicians without extensive data analysis expertise.

In summary, the proposed AI-based drug discovery pipeline has the potential to significantly advance breast cancer drug discovery by overcoming the limitations of existing systems and providing a more efficient, effective, and accessible approach to identifying potential drug candidates.

2.3. Requirement Analysis

The success of the proposed AI-based drug discovery pipeline depends on a thorough understanding of the project's requirements. Requirement analysis involves identifying the system's functional and non-functional requirements, as well as the constraints and limitations that may impact the system's development and implementation.

2.3.1 Functional Requirements

The functional requirements(FR) of the proposed system include the following:

1. **Data Integration:** The system must be able to integrate multiple data sources, including genomic, transcriptomic, proteomic, and clinical data, to provide a more comprehensive understanding of breast cancer biology and potential drug targets.
2. **Machine Learning:** The system must incorporate machine learning algorithms, such as deep learning and reinforcement learning, to analyze large datasets and identify potential drug candidates.
3. **High-Throughput Screening:** The system must include high-throughput screening techniques to rapidly test potential drug candidates in vitro and in vivo, reducing the time and cost required to identify successful candidates.
4. **User-Friendly Interface:** The system must provide a user-friendly web-based interface that allows researchers and clinicians to easily access and interpret the results of the drug discovery pipeline.

Table 2: Description of Functional Requirement

ID	Title	Description	Priority
FR-01	Data Integration	The system must be able to integrate multiple data sources, including genomic, transcriptomic, proteomic, and clinical data, to provide a more comprehensive understanding of breast cancer biology and potential drug targets.	High
FR-02	Machine Learning	The system must incorporate machine learning algorithms, such as deep learning and reinforcement learning, to analyze large datasets and identify potential drug candidates.	High
FR-03	High-Throughput Screening	The system must include high-throughput screening techniques to rapidly test potential drug candidates in vitro and in vivo, reducing the time and cost required to identify successful candidates.	High
FR-04	User-Friendly Interface	The system must provide a user-friendly web-based interface that allows researchers and clinicians to easily access and interpret the results of the drug discovery pipeline.	Medium

2.3.2 Non-Functional Requirements

The non-functional requirements(NFR) of the proposed system include the following:

1. **Scalability:** The system must be able to handle large volumes of data and be scalable to accommodate future growth.
2. **Performance:** The system must be able to process and analyze data quickly and efficiently.
3. **Security:** The system must be secure and protect sensitive patient data.

2.3.3 Constraints and Limitations

The constraints and limitations of the proposed system include the following:

1. **Availability of Data:** The system's effectiveness relies on the availability of high-quality data, including genomic and clinical data, which may not be readily accessible or available.
2. **Cost:** The implementation and maintenance of the system may require significant financial resources, including the cost of data acquisition, computing resources, and personnel.
3. **Regulatory Approval:** The development and implementation of the system must comply with applicable regulatory requirements, including those related to drug privacy and data security.

In summary, a thorough requirement analysis is essential to the success of the proposed AI-based drug discovery pipeline. By identifying the system's functional and non-functional requirements, as well as its constraints and limitations, the development team can ensure that the system meets the needs of its users while also addressing any potential challenges or limitations that may impact the system's implementation and effectiveness.

2.3.4 System Use case

Actors:

- **User**: A user who accesses the system to predict a diagnosis for breast cancer.
- **Admin**: An administrator who manages the system.
- **AI Model**: An artificial intelligence model for prediction to user data.

Use Cases:

- **User Register**: Allows a new user to register to the system.
- **User Login**: Allows the user to login to the system with their registered credentials.
- **View Result**: Allows the user to view the predicted drug for breast cancer.
- **Logout**: Allows the user to logout from the system.
- **Change Light Mode**: Allows the user to change the background color of the system interface.

- **Send Feedback**: Allows the user to send feedback about the system.
- **Download Predicted Output**: Allows the user to download the predicted drug as a file.
- **View Predicted History**: Allows the user to view their previous predicted history.
- **Predict**: Allows the user to input data for breast cancer prediction using the AI model.
- **Interpret Output**: Allows the AI model to interpret the input data and predict drug.
- **Admin Login**: Allows the administrator to login to the system with credentials.
- **View Feedback**: Allows the administrator to view the feedback submitted by users.
- **View Predicted History**: Allows the administrator to view the predicted history of users.
- **Update User Info**: Updates the information of the credential users like password and email.

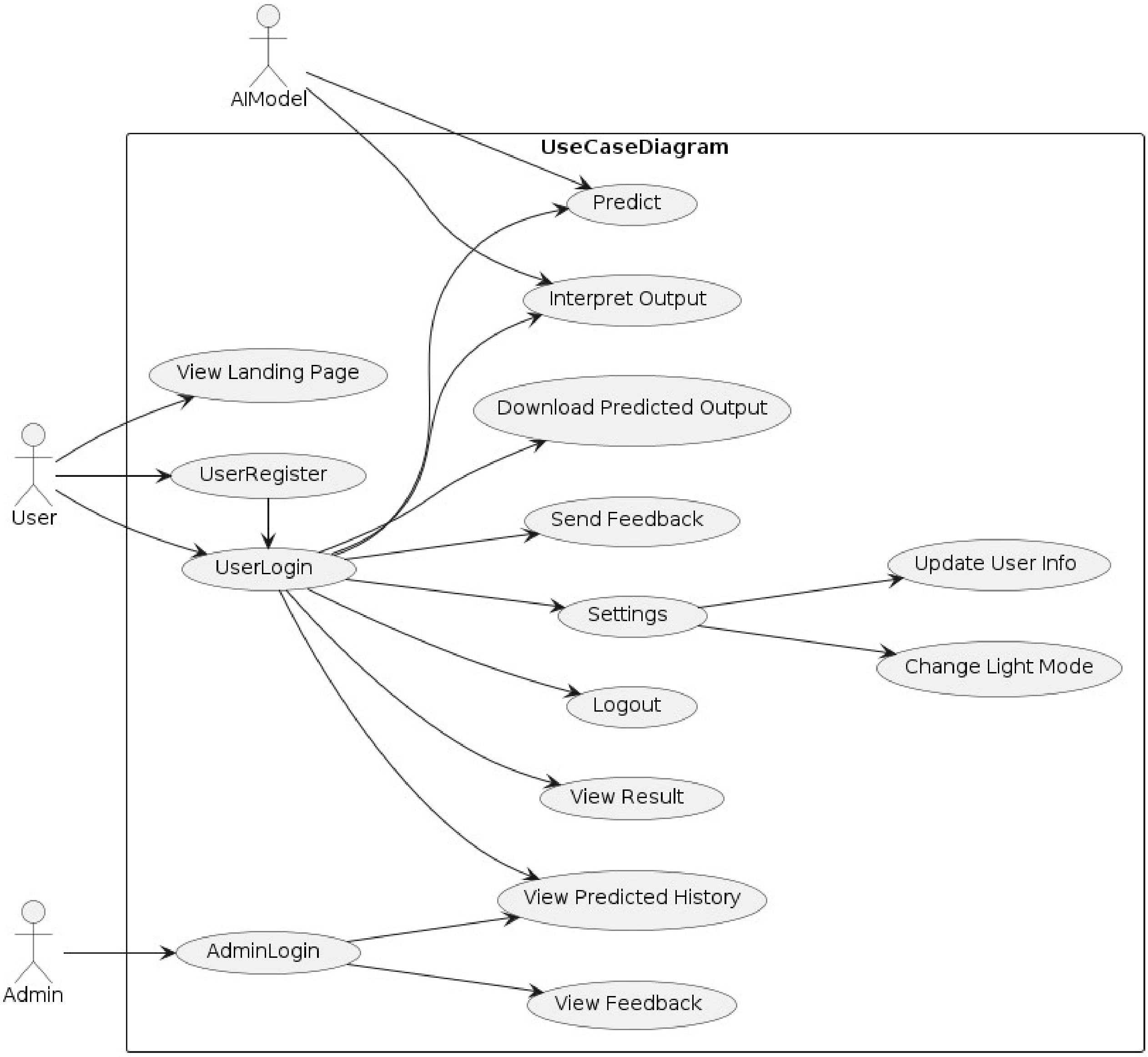


Figure 1: Use Case Diagram of the System

Table 3: User Register Use Case Doc

Use Case Name	Use Register
Actor	User
Description	The user creates an account with the application using Firebase Authentication.
Pre-conditions	None.
Post-conditions	User is registered and can log in.
Normal Flow	• User clicks on the "Register" button. • User is redirected to Firebase Authentication page. • User enters their email address and password. • User clicks on "Create Account" button. • User is redirected back to the application. • User sees a confirmation message that they have successfully registered.
Alternative Flow	None
Exceptions	Firebase Authentication service is down or not responding.
Priority	High

Table 4: User Login Use Case Doc

Use Case Name	User Login
Actor	Users
Description	The user logs into their account with the application using Firebase Authentication.
Pre-conditions	User is registered with the application using Firebase Authentication.
Post-conditions	User is logged into their account and can access their data.
Normal Flow	• User clicks on the "Log In" button. • User is redirected to Firebase Authentication page. • User enters their email address and password. • User clicks on "Log In" button. • User is redirected back to the application.
Alternative Flow	• If the user enters incorrect login credentials, the system displays an error message and prompts the user to try again. • If the user has forgotten their password, they can click on the "Forgot password" link to reset their password.
Exceptions	• If the user's account has been suspended or deactivated, the system will not allow them to log in and will display an error message. • If the system is experiencing technical difficulties, the login feature may not be available and an error message will be displayed to the user.
Priority	High

Table 5: View Result Use Case Doc

Use Case Name	View Result
Actor	User
Description	Allows the user to view the predicted drug based on the input data.
Pre-conditions	User must be logged in.
Post-conditions	The user can view the predicted drug.
Normal Flow	• The user clicks on the "Predict" button after submitting input data. • The system retrieves the predicted diagnosis based on the input data. • The system displays the predicted diagnosis to the user.
Alternative Flow	• If the system cannot retrieve a predicted diagnosis based on the input data, an error message is displayed to the user. • If the predicted diagnosis is not available due to technical issues, an error message is displayed to the user.
Exceptions	• Invalid input data through an error.
Priority	High

Table 6: Change Light Mode Use Case Doc

Use Case Name	Change Light Mode
Actor	User
Description	Allows the user to change the color scheme of the app.
Pre-conditions	User must be logged in.
Post-conditions	The selected light mode is applied.
Normal Flow	• User clicks on the "Change Light Mode" button. • App displays available light modes. • User selects a light mode. • App applies the selected light mode.
Alternative Flow	None
Exceptions	If there are no available light modes, display an error message.
Priority	Low

Table 7: Logout Use Case Doc

Use Case Name	Logout
Actor	User
Description	This use case describes the steps a user takes to log out of the system.
Pre-conditions	User is logged in.

Post-conditions	User is logged out and returned to the login page.
Normal Flow	• User clicks the "Logout" button. • System logs out the user and redirects them to the login page.
Alternative Flow	-
Exceptions	-
Priority	Low

Table 8: Send Feedback Use Case Doc

Use Case Name	**Send Feedback**
Actor	User
Description	This use case describes the steps a user takes to send feedback to the system administrators.
Pre-conditions	User is logged in and on the feedback page.
Post-conditions	Feedback is sent to the system administrators.
Normal Flow	• User navigates to the feedback page. • User enters feedback into the form. • User clicks "Submit" button. • System saves the feedback and notifies the user that it was received successfully.
Alternative Flow	If user encounters an error while submitting the feedback, the system displays an error message and the user returns to the feedback page.
Exceptions	-
Priority	Medium

Table 9: Predict Use Case Doc

Use Case Name	**Predict**
Actor	AI Model
Description	User submits input data to receive a predicted drug chemical.
Pre-conditions	The user is logged in.
Post-conditions	The system generates and displays a predicted drug chemical.
Normal Flow	• User enters input CSV data and submits it for prediction.

	- System validates the input data. - System applies the trained AI model to the input data. - System generates a predicted drug based on the input data. - System displays the predicted drug to the user.
Alternative Flow	- If the input data is invalid, the system displays an error message and prompts the user to correct the data. - If the AI model fails to generate a predicted diagnosis, the system displays a message indicating that it could not generate a prediction.
Exceptions	N/A
Priority	High

Table 10: Download the Predicted Output Use Case Doc

Use Case Name	Download Predicted Output
Actor	User
Description	User downloads a file containing the predicted drug chemical.
Pre-conditions	- The user is logged in and has access to the download functionality. - The user has previously submitted input data for prediction.
Post-conditions	The system generates and downloads a file containing the predicted drug.
Normal Flow	- User selects the "Download Predictions" option. - System generates a file containing the predicted drug chemical. - System prompts the user to save the file. - User selects a location to save the file. - System saves the file to the selected location.
Alternative Flow	N/A
Exceptions	N/A
Priority	Medium

Table 11: View Feedback Use Case Doc

Use Case Name	View Feedback
Actor	Admin
Description	Admin views feedback submitted by users.
Pre-conditions	The admin is logged in and has access to the view feedback functionality.

	Users have previously submitted feedback.
Post-conditions	The system displays the feedback submitted by users.
Normal Flow	• Admin selects the "View Feedback" option. • System retrieves the feedback submitted by users. • System displays the feedback to the admin.
Alternative Flow	N/A
Exceptions	N/A
Priority	Low

Table 12: Interpret Output Use Case Doc

Use Case Name	**Interpret Output**
Actor	AI Model
Description	AI Model interprets the predicted diagnosis generated by the system.
Pre-conditions	The system has generated a predicted drug chemical.
Post-conditions	The AI model generates an interpretation of the predicted drug.
Normal Flow	• The system passes the predicted drug chemical to the AI model. • The AI model analyzes the predicted drug. • The AI model generates an interpretation of the predicted drug. • The interpretation is passed back to the system. • The system displays the interpretation to the user.
Alternative Flow	N/A
Exceptions	N/A
Priority	Medium

Table 13: View Prediction History Use Case Doc

Use Case Name	**View Prediction History**
Actor	User, Admin
Description	User and admin view the history of previously submitted input CSV data and their predicted drug chemical.
Pre-conditions	User is logged in.

Post-conditions	User has successfully viewed their predicted drug chemical history.
Normal Flow	• User clicks on "View Predicted History" button. • System displays a list of the user's previous predicted diagnoses.
Alternative Flow	N/A
Exceptions	N/A
Priority	Medium

Table 14: Admin Login Use Case Doc

Use Case Name	Admin Login
Actor	Admin
Description	Allows an admin to log in to the system using Firebase authentication.
Pre-conditions	Admin has a valid Firebase authentication account.
Post-conditions	Admin has successfully logged in.
Normal Flow	• Admin navigates to the login page. • Admin enters their Firebase authentication credentials. • System verifies the credentials and logs the admin in.
Alternative Flow	N/A
Exceptions	• If the admin enters invalid Firebase authentication credentials, the system displays an error message and prompts the admin to try again. • If the admin's Firebase authentication account has been disabled or deleted, the system displays an error message and prompts the admin to contact the system administrator.
Priority	High

2.3.5 Business Rule Documentation

The purpose of this document is to define the business rules that govern the behavior of the proposed system. These business rules will help ensure that the system functions in a consistent and predictable manner, and that all users of the system are held to the same standards.

Table 15: Business Rule Documentation(BRD)

No.	Description
1	Users can only access data that they are authorized to view.
2	All user input must be validated before it is processed by the system.

3	The system must be available 24/7 with a maximum downtime of 1 hour per month for maintenance.
4	All system errors and exceptions must be logged for debugging and audit purposes.
5	The system must comply with all relevant data protection and privacy regulations.
6	The system must be capable of handling a minimum of 1000 concurrent users.
7	Any changes to the system must be approved by the designated project manager.
8	The system must provide adequate data backup and recovery mechanisms to prevent data loss in case of system failure.
9	All system users must adhere to the ethical and professional standards set by their respective professional organizations.

The above business rules will ensure that the proposed system operates effectively, efficiently and consistently. All users of the system must adhere to these rules to ensure that the system is reliable and secure.

2.3.6 User Interface Prototype

The User Interface Prototype is a visual representation of how the system will look and function for the end user. The User Interface Prototype consists of a series of screen mockups that demonstrate the system's functionality and user flow. The User Interface Prototype is divided into the following sections:

1. Landing Page Screen
2. Register Screen
3. Login Screen
4. Results Screen
5. Settings Screen
6. Feedback Screen
7. Prediction History Screen
8. Logout Screen

Each section is described in more detail below.

1. **Landing Page Screen**: The Landing Page is the first screen the user sees when they open the application. It includes a logo or brand name, a brief description of the system, and a call-to-action button to either register or login.

2. **Register Screen**: The Register Screen is where new users can create an account by providing their name, email address, and password. There should be a validation check to ensure that the email address is valid and the password meets the required complexity criteria. Upon successful registration, the user should be redirected to the login screen.
3. **Login Screen**: The Login Screen allows users to access their account by entering their email address and password. There should be a validation check to ensure that the email address is valid and the password is correct. If the login is successful, the user should be directed to the Results Screen.
4. **Results Screen**: The Results Screen is where users can input input data and receive a predicted drug. The screen should include input fields for relevant input data. Once the prediction is complete, the results should be displayed on the screen.
5. **Settings Screen**: The Settings Screen is where users can customize the app to their preferences. This includes options to change the app's theme or color scheme, adjust notification settings, and change their account information.
6. **Feedback Screen**: The Feedback Screen is where users can provide feedback on the app's performance or report any issues or bugs they encounter. The screen should include input fields for the user's name, email address, and a description of the feedback or issue.
7. **Prediction History Screen**: The Prediction History Screen displays a list of past predictions made by the user. Each prediction should be displayed with its corresponding date and time stamp, as well as the predicted drug chemical.
8. **Logout Screen**: The Logout Screen allows users to log out of their account and exit the app. The screen should include a confirmation message asking the user if they want to log out. If the user confirms, they should be directed back to the Landing Page.

2.3.7 Activity Diagram

The Activity Diagram(AD) is divided into several sections, each representing a specific activity within the system. The sections are connected by arrows to show the flow of activity. The Activity Diagram includes the following sections:

1. Login Activity
2. Register Activity
3. View Result Activity
4. Change Settings Activity – here we have two setting’s activities namely, Change Light Mode and Update User Info like password and email.
5. Download Prediction Activity

6. View Prediction History Activity
7. Send Feedback Activity
8. View Feedback Activity

Each section is described in more detail below.

1. **Login Activity**: The Login Activity starts with the user entering their login credentials. If the credentials are correct, the system will validate and authenticate the user. Once the user is authenticated, they will be directed to the View Result Activity.
2. **Register Activity**: The Register Activity starts with the user providing their registration details, such as their name, email, and password. Once the user submits their registration details, the system will verify and create a new user account. Once the account is created, the user will be directed to the Login Activity.
3. **View Result Activity**: The View Result Activity starts with the user requesting to view their result. The system will retrieve and display the user's result. If the user wants to download the prediction output, they will be directed to the Download Prediction Activity. If the user wants to interpret the output, they will be directed to the Interpret Output Activity.
4. **Change Settings Activity**: The Change Settings Activity starts with the user requesting to change their settings. The user can choose to change their light mode or update their user information. Once the user submits their changes, the system will update the user's settings and redirect them back to the View Result Activity.
5. **Download Prediction Activity**: The Download Prediction Activity starts with the user requesting to download their prediction output. The system will retrieve and download the prediction output for the user.
6. **View Prediction History Activity**: The View Prediction History Activity starts with the user requesting to view their prediction history. The system will retrieve and display the user's prediction history.
7. **Send Feedback Activity**: The Send Feedback Activity starts with the user providing their feedback about the system. Once the user submits their feedback, the system will store and process the feedback.
8. **View Feedback Activity**: The View Feedback Activity starts with the admin requesting to view the user feedback. The system will retrieve and display the user feedback for the admin to review.

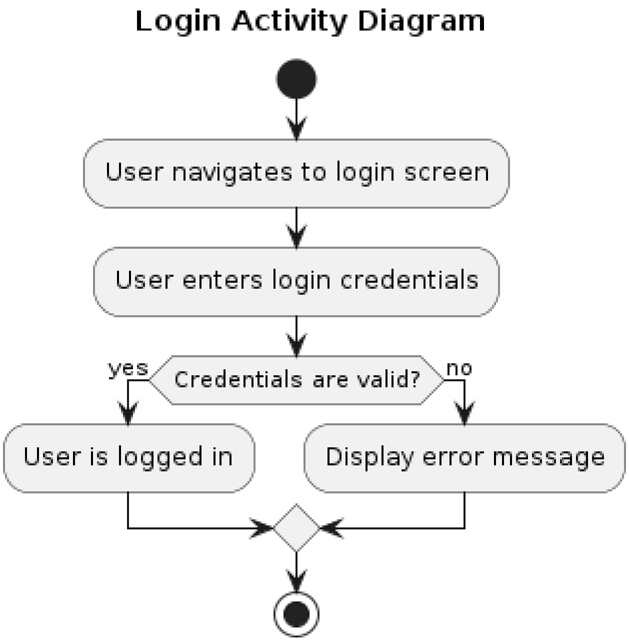


Figure 2: Login Activity Diagram

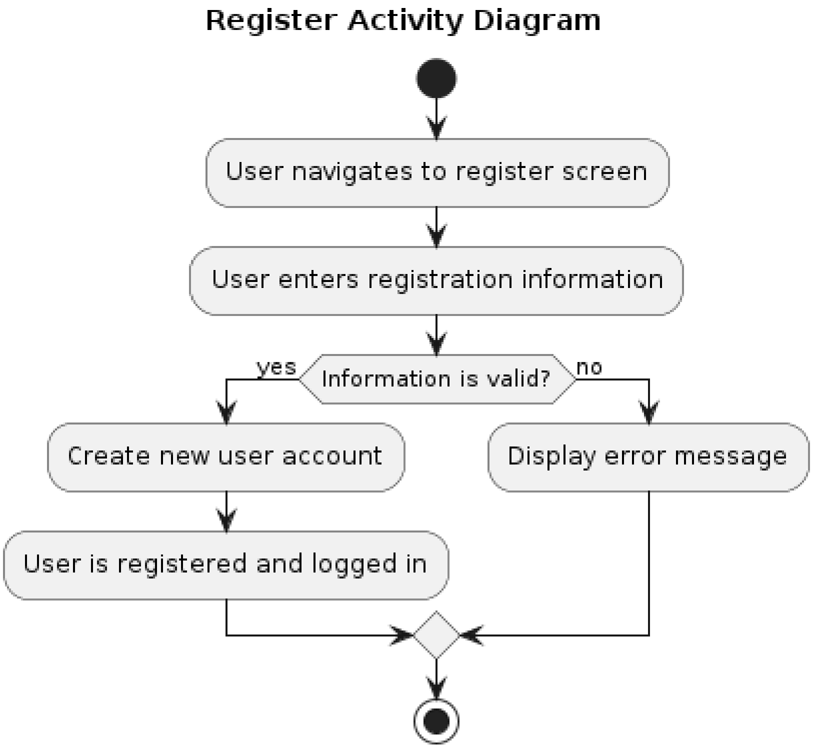

Register Activity Diagram
User navigates to register screen
User enters registration information
Information is valid?
yes
no
Create new user account
Display error message
User is registered and logged in

Figure 3: Register Activity Diagram

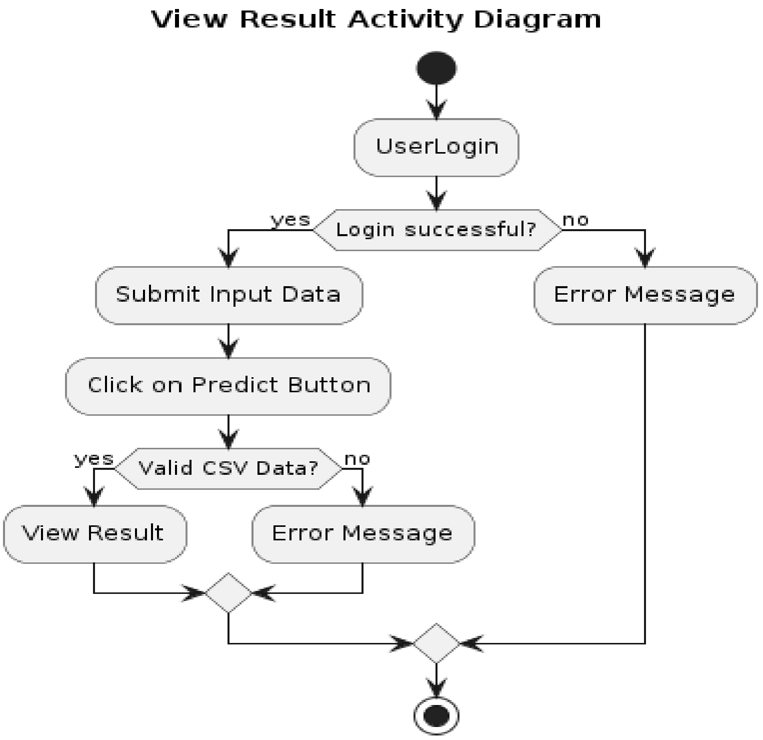


Figure 4: View Result Activity Diagram

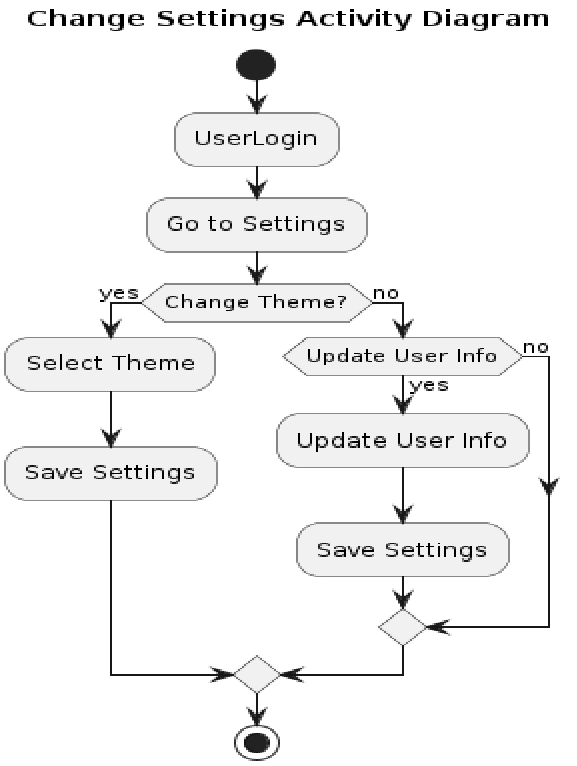


Figure 5: Change Settings Activity Diagram

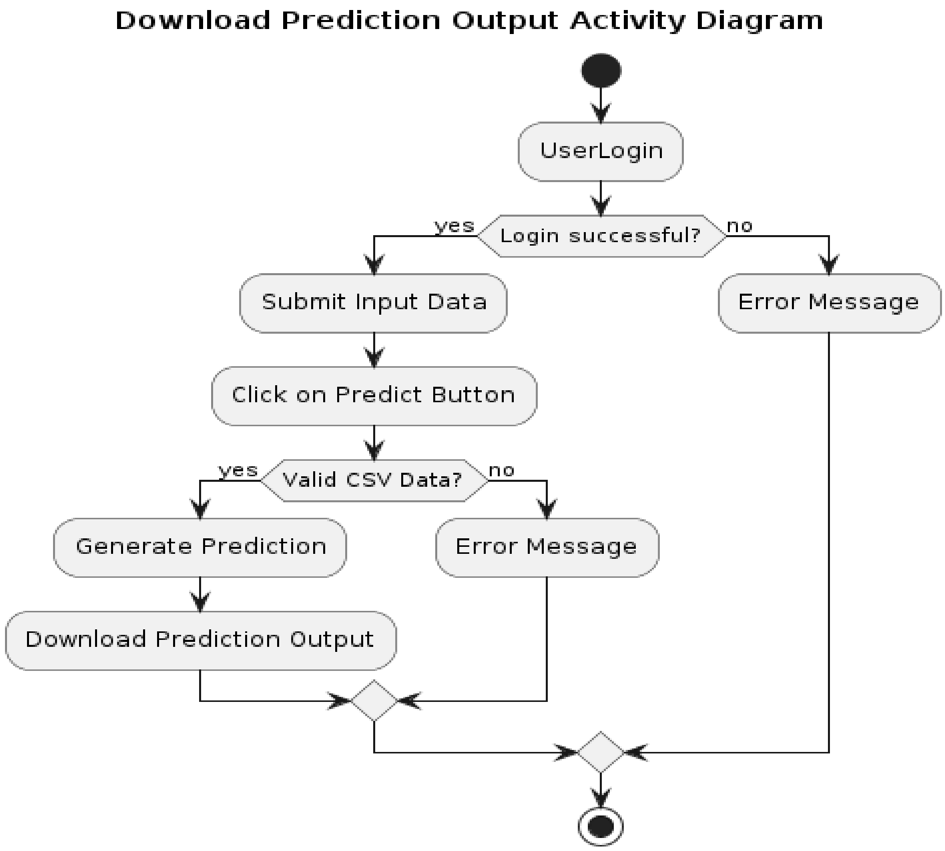


Figure 6: Download Prediction Output Activity Diagram

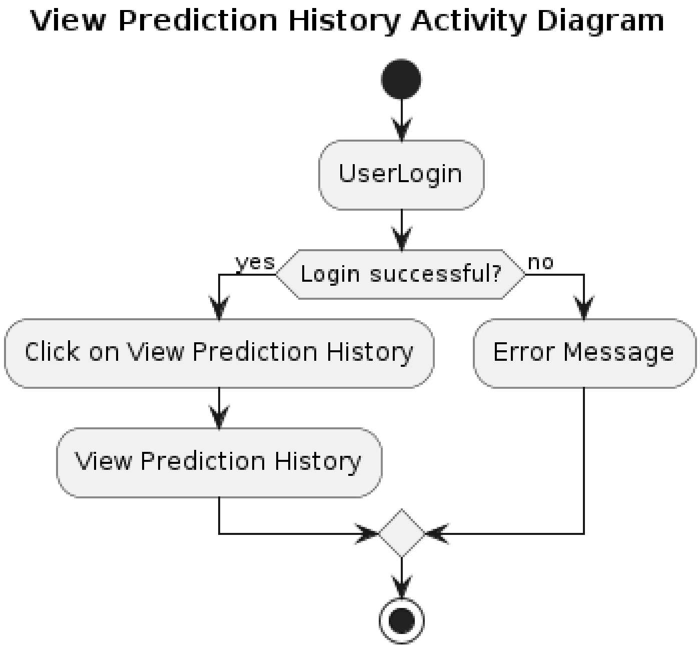

View Prediction History Activity Diagram
UserLogin
Login successful?
yes
no
Click on View Prediction History
Error Message
View Prediction History

Figure 7: View Prediction History Activity Diagram

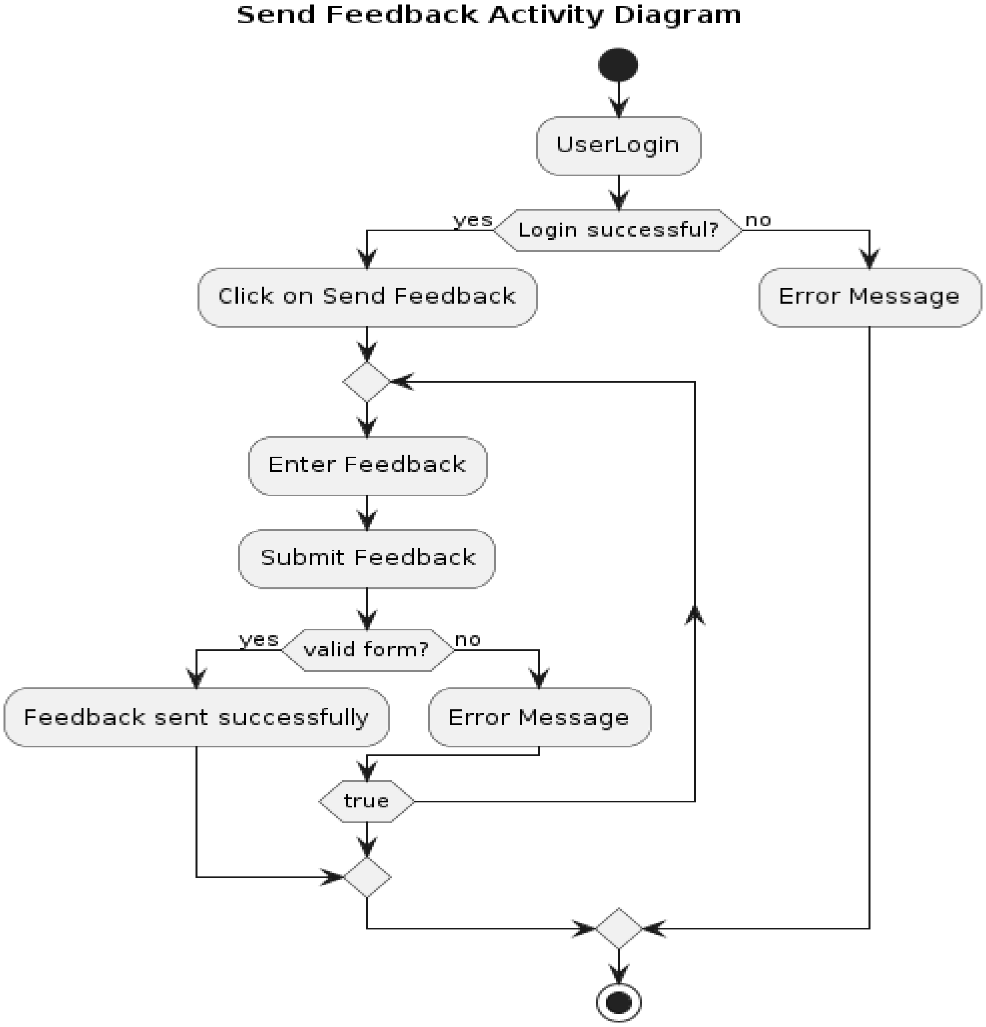


Figure 8: Send Feedback Activity Diagram

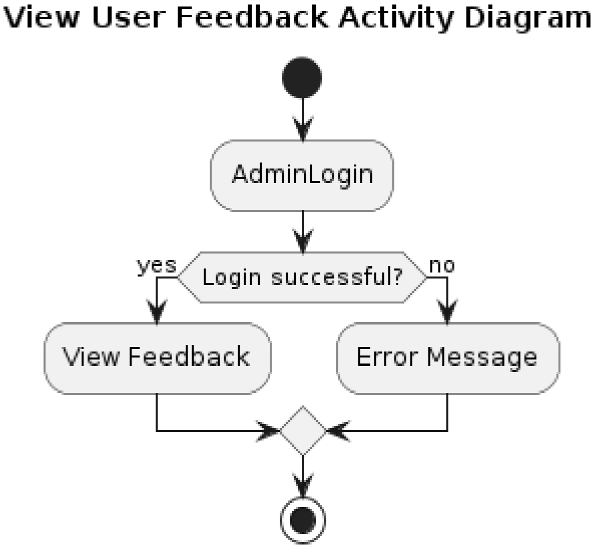


Figure 9: View User Feedback Activity Diagram

2.3.8 Sequence Diagram

The Sequence Diagram(SD) is divided into several sections, each representing a specific interaction between system components. The sections are connected by arrows to show the flow of interaction. The Sequence Diagram includes the following sections:

1. Login Sequence
2. Register Sequence
3. View Result Sequence
4. Change Settings Sequence
5. Download Predicted Output Sequence
6. View Prediction History Sequence
7. Send Feedback Sequence
8. View Feedback Sequence

Each section is described in more detail below.

1. **Login Sequence:** Shows the sequence of events that occur when a user logs into the system, including sending login credentials, validating the credentials, and granting access to the system.
2. **Register Sequence:** Shows the sequence of events that occur when a user registers for the system, including submitting registration information, validating the information, and creating a new user account.
3. **View Result Sequence:** Shows the sequence of events that occur when a user views the results of a prediction, including requesting the prediction, processing the request, and displaying the results.
4. **Change Settings Sequence:** Shows the sequence of events that occur when a user changes their settings, such as updating their user profile or changing the theme of the interface.
5. **Download Predicted Output Sequence**: Shows the sequence of events that occur when a user downloads the output of a prediction, including requesting the output, processing the request, and providing the output to the user.
6. **View Prediction History Sequence**: Shows the sequence of events that occur when a user views their prediction history, including requesting the history, processing the request, and displaying the history.

7. **Send Feedback Sequence**: Shows the sequence of events that occur when a user sends feedback, including submitting the feedback, processing the feedback, and confirming the feedback has been sent.
8. **View Feedback Sequence**: Shows the sequence of events that occur when an administrator views user feedback, including requesting the feedback, processing the request, and displaying the feedback.

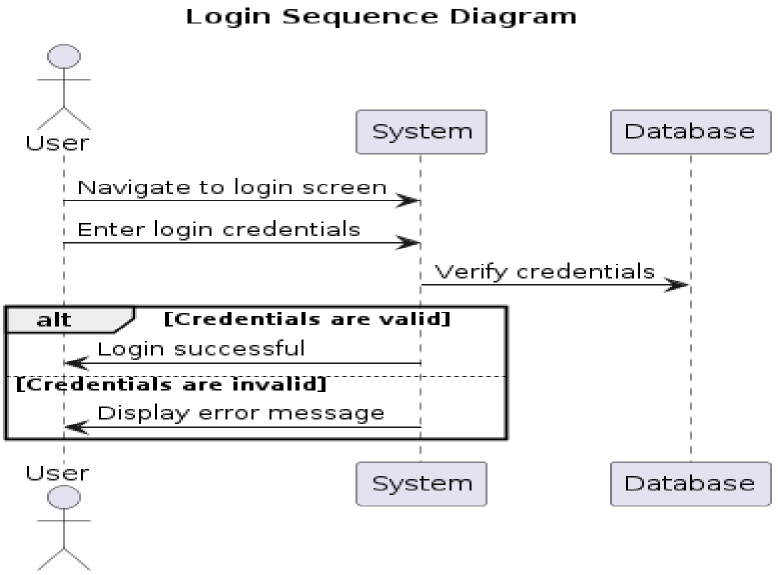


Figure 10: Login Sequence Diagram

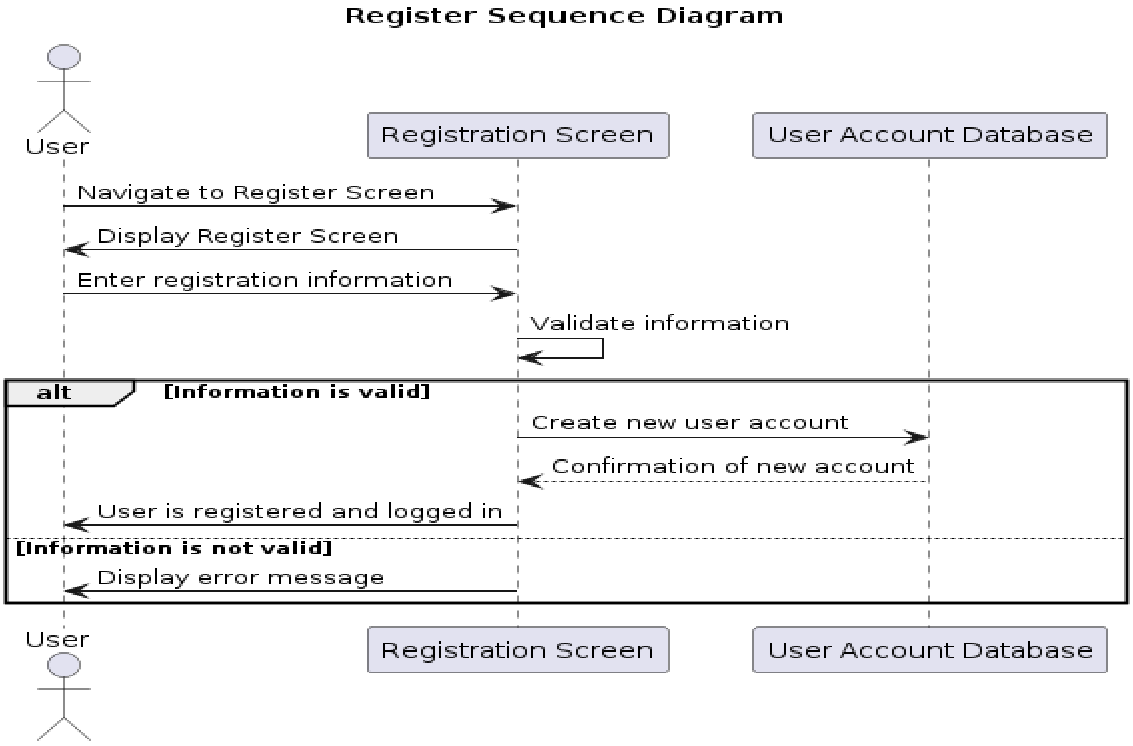


Figure 11: Register Sequence Diagram

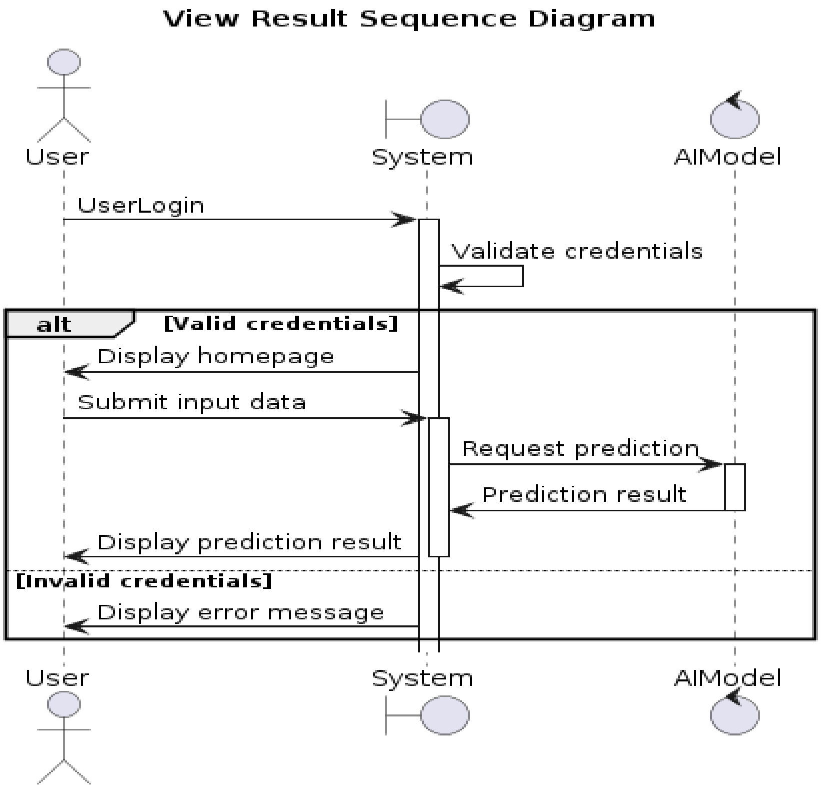


Figure 12: View Result Sequence Diagram

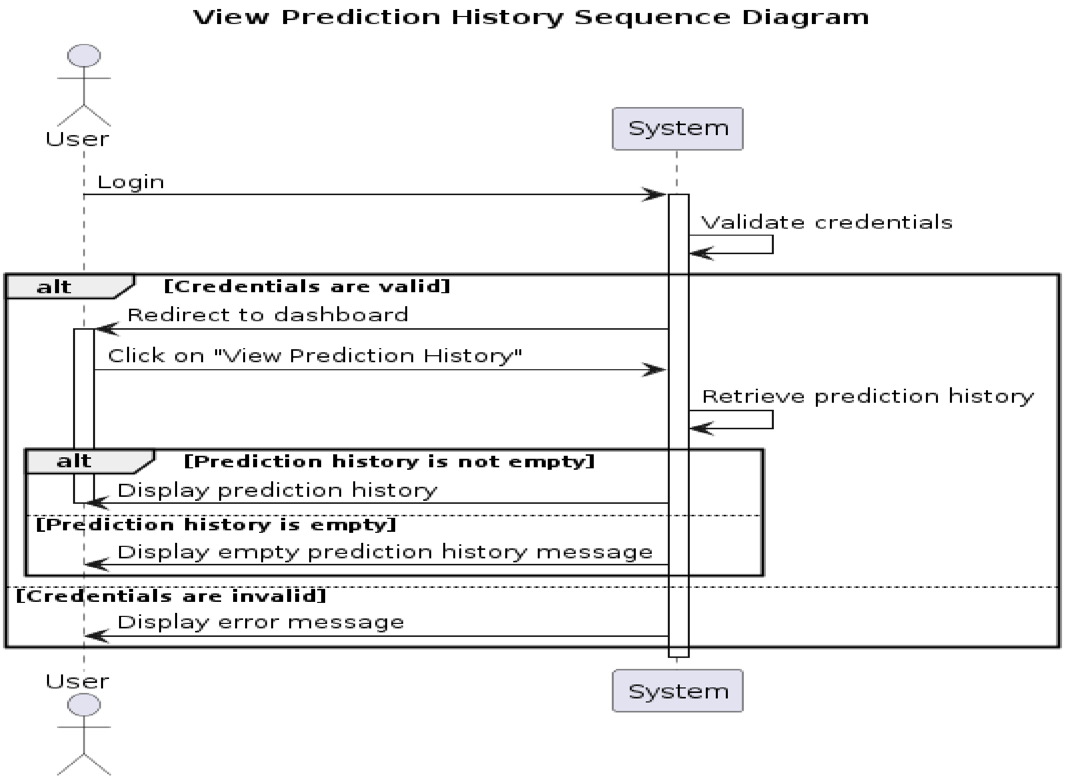

View Prediction History Sequence Diagram
User
System
Login
Validate credentials
alt
[Credentials are valid]
Redirect to dashboard
Click on "View Prediction History"
Retrieve prediction history
alt
[Prediction history is not empty]
Display prediction history
[Prediction history is empty]
Display empty prediction history message
[Credentials are invalid]
Display error message
User
System

Figure 13: View Prediction History Sequence Diagram

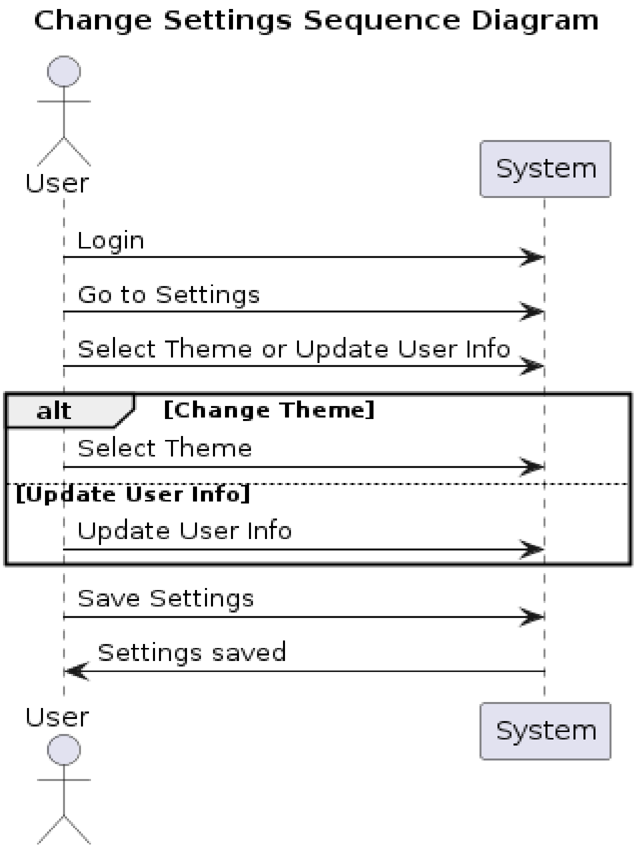


Figure 14: Change Settings Sequence Diagram

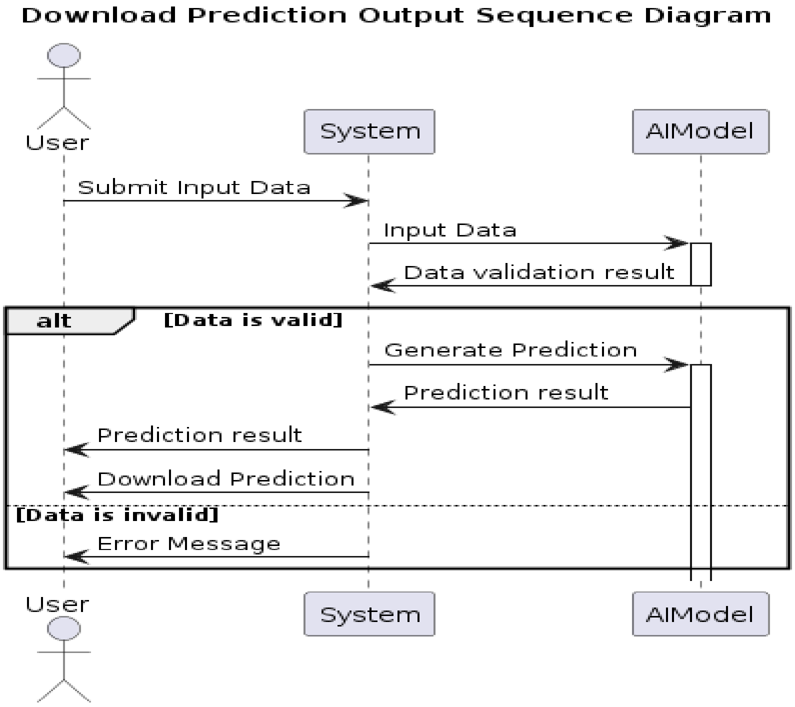


Figure 15: Download Prediction Output Sequence Diagram

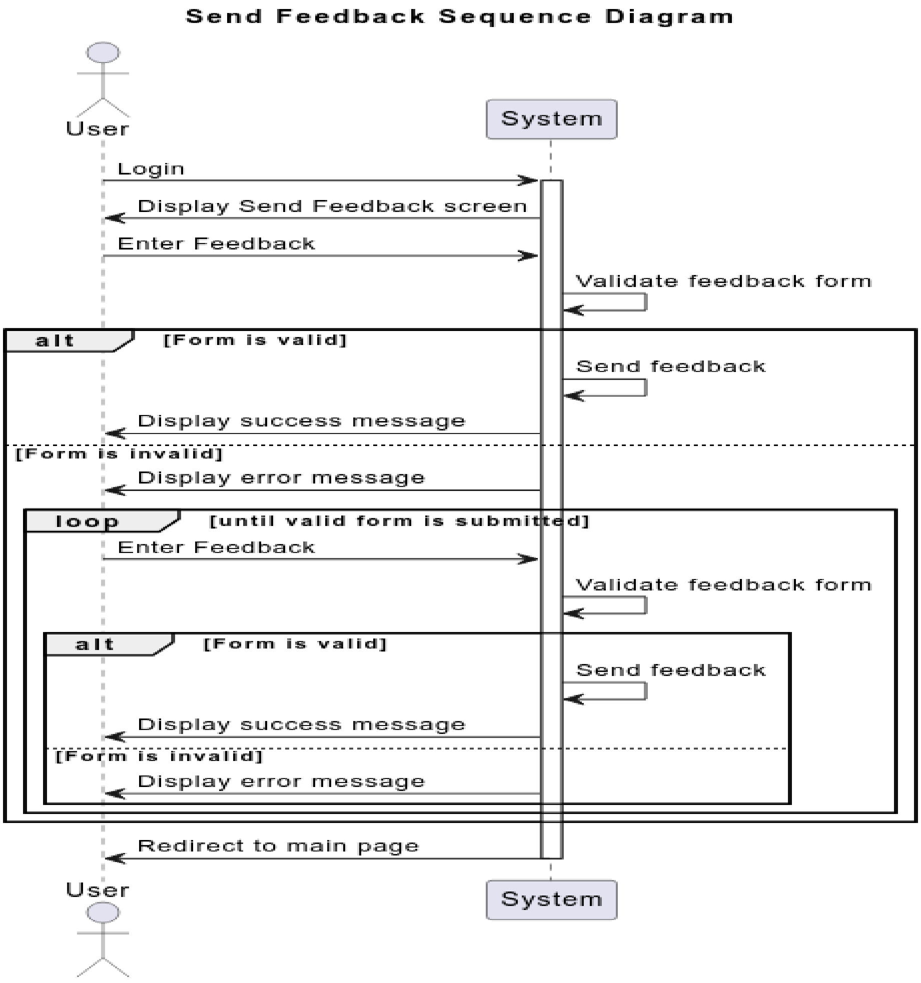


Figure 16: Send Feedback Sequence Diagram

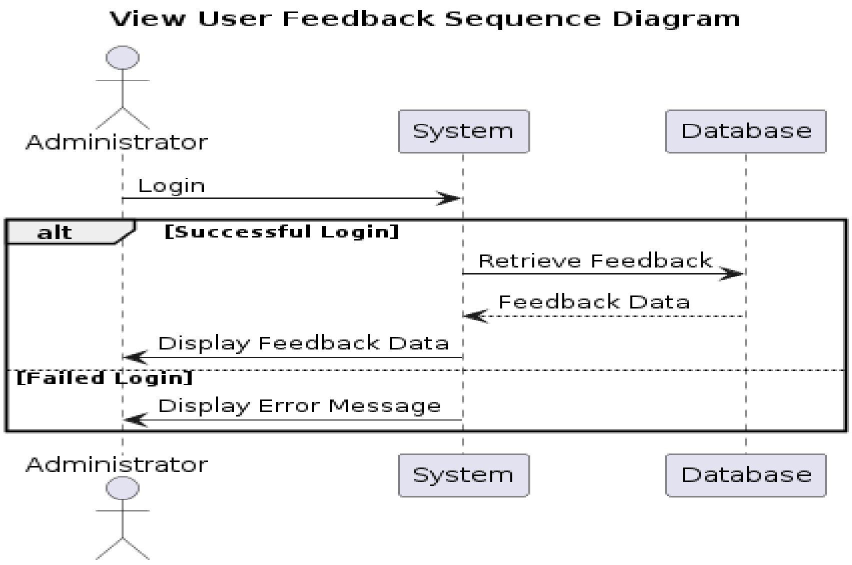


Figure 17: View User Feedback Sequence Diagram

2.3.9 Logic Model

The Logic Model is a visual representation of how the inputs, activities, outputs, and outcomes of the system are related. It helps to clarify the program's logic and serves as a roadmap for the system's design, implementation, and evaluation.

1. **Login:**

   - The user navigates to the login screen.
   - The user enters their login credentials.
   - The system checks if the entered credentials are valid.
   - If the credentials are valid, the user is logged in and granted access to the system.
   - If the credentials are invalid, an error message is displayed and the user is not granted access.

2. **Register:**

   - The user navigates to the registration screen.
   - The user enters their registration information.
   - The system checks if the entered information is valid.
   - If the information is valid, a new user account is created and the user is automatically logged in.
   - If the information is invalid, an error message is displayed and the user is not registered.

3. **View Result:**

   - The user logs in to the system.
   - The user submits input data and clicks on the predict button.
   - The system checks if the entered data is valid.
   - If the data is valid, the system generates the prediction and displays it to the user.
   - If the data is invalid, an error message is displayed and the user is prompted to enter valid data.

4. **Change Settings:**

   - The user logs in to the system.
   - The user navigates to the settings screen.
   - If the user chooses to change the theme, the system allows the user to select a new theme and saves the settings.
   - If the user chooses to update their user information, the system allows the user to update their information and saves the settings.

5. **Download Predicted Output:**

   - The user logs in to the system.
   - The user submits input data and clicks on the predict button.

- The system checks if the entered data is valid.
- If the data is valid, the system generates the prediction and allows the user to download the output file.
- If the data is invalid, an error message is displayed and the user is prompted to enter valid data.

6. **View Prediction History:**

- The user logs in to the system.
- The user clicks on the "View Prediction History" button.
- The system displays the user's prediction history.

7. **Send Feedback:**

- The user logs in to the system.
- The user clicks on the "Send Feedback" button.
- The user enters feedback and submits the form.
- The system checks if the entered form is valid.
- If the form is valid, the system sends the feedback.
- If the form is invalid, an error message is displayed and the user is prompted to enter valid feedback.

8. **View Feedback:**

- An admin logs in to the system.
- The admin clicks on the "View Feedback" button.
- The system displays all user feedback.

2.4. Change Case

Change cases(CC) are used to document the proposed changes to a system's requirements, design, or implementation. The table below outlines the different change cases identified for this project:

Table 16: Change Cases of the Project

ID	Title	Description
CC-01	Integration with additional datasets	The system needs to be able to integrate with new datasets that may become available for breast cancer research.
CC-02	Addition of new machine learning models	As new machine learning models are developed and become available, they should be incorporated into the system to improve accuracy and efficiency.
CC-03	Expansion to other types of cancer	The system may be expanded to include the detection and discovery of other types of cancer beyond breast cancer.
CC-04	Enhancement of user	The user interface should be continually evaluated and

	interface	improved to ensure ease of use and accessibility for all users.
CC-05	Integration with electronic health records	The system could be integrated with electronic health records to improve the accuracy and efficiency of patient diagnosis and treatment.

These change cases will help guide future development and improvement of the system, ensuring that it continues to meet the evolving needs of researchers and healthcare providers in the field of breast cancer detection and treatment.

3. Chapter Three: System Design

3.1. Introduction

The System Design document outlines the technical aspects of the drug discovery project using AI for breast cancer. This document is aimed at developers and technical stakeholders, and provides an overview of the system architecture, design decisions, and implementation details. The document also outlines the key features and functionality of the system, as well as the tools and technologies used in its development. The System Design document serves as a blueprint for the development team, providing guidance on how to build, test, and maintain the system. It is an important resource for ensuring that the system is developed to meet the project objectives, and that it is both effective and efficient.

3.2. Architectural Design

The application is built using a client-server architecture with the front-end developed using HTML pages and Streamlit app while the backend is based on Firebase as the serverless computing platform.

Functional Architecture: The functional architecture of the system consists of the following components:

- **Client-Side Components:** The client-side of the application is developed using HTML pages and Streamlit app that provide an interactive user interface for the application. The HTML pages allow users to view and interact with the website while the Streamlit app provides the user interface for data visualization and exploration.
- **Server-Side Components**: The server-side of the application is developed using Firebase, which provides serverless computing and storage infrastructure. Firebase Authentication is used to register and store user accounts, while Firebase Cloud Firestore is used to store user feedback and prediction history for later use.
- **AI Prediction Engine**: The AI prediction engine is responsible for processing user input data and providing drug recommendations. The engine is built using machine learning algorithms such as neural networks that have been trained using large datasets of breast cancer patient information.

Technical Architecture: The technical architecture of the system consists of the following components:

- **HTML Pages**: The HTML pages are used to provide the user interface for the application. The pages are designed to be responsive and user-friendly, providing a seamless experience for the user.

- **Streamlit App:** The Streamlit app is used to provide interactive data visualization and exploration features. The app allows users to explore data and generate insights using visualizations and other interactive features.
- **Firebase Authentication**: Firebase Authentication is used to register and store user accounts. The authentication system is secure and easy to use, allowing users to quickly create an account and log in to the system.
- **Firebase Cloud Database**: Firebase Cloud Firestore is used to store user feedback and prediction history for later use. The data is stored in a structured format, making it easy to access and analyze using machine learning algorithms.
- **AI Prediction Engine:** The AI prediction engine is built using machine learning algorithms such as neural networks. The engine is responsible for processing user input data and providing drug recommendations.

3.2.1 Component Modeling

Component modeling is a technique used to describe and organize the software components of a system, and their relationships with each other. It helps to understand the functionalities of the system at a higher level of abstraction.

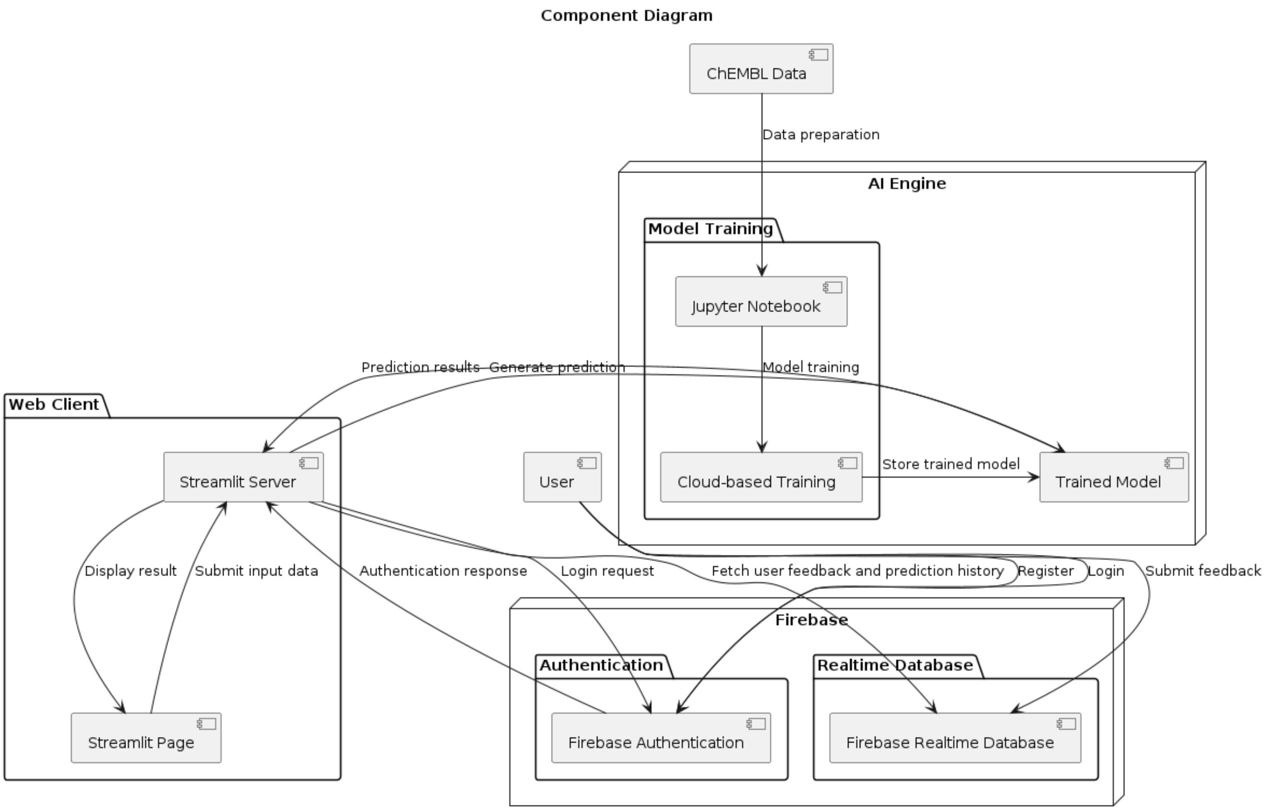


Figure 18: Component Diagram

3.3. Deployment Modeling

Deployment modeling is a technique used to describe the hardware components and their relationships with the software components of a system. It helps to understand how the system will be deployed on different hardware components.

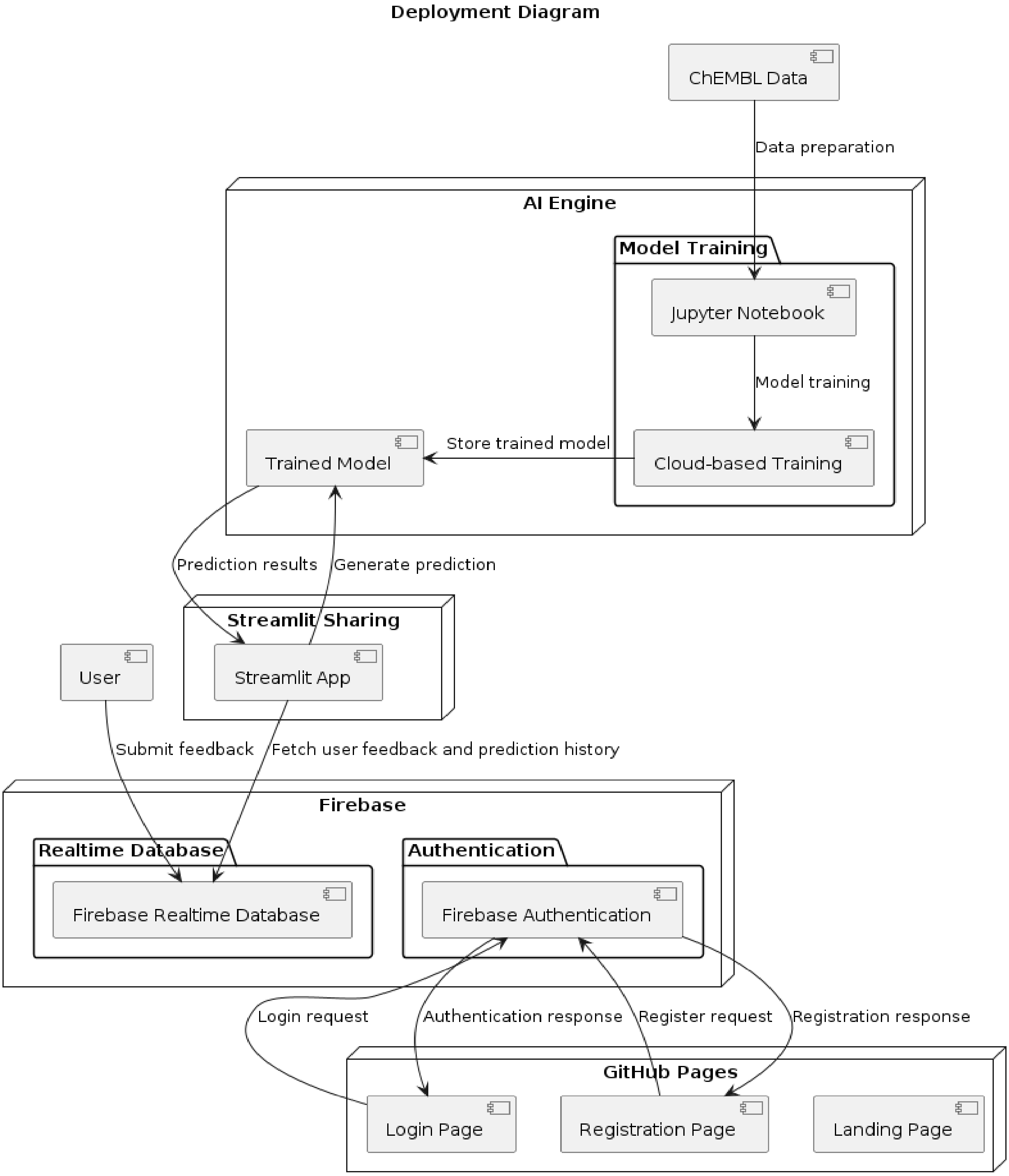
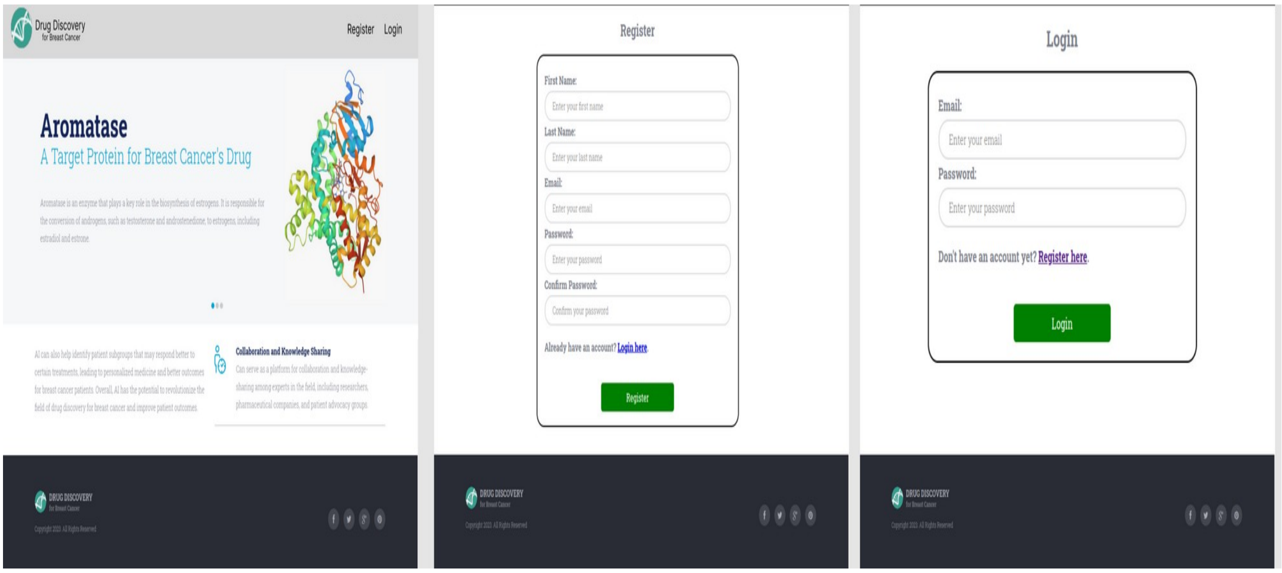


Figure 19: Deployment Modeling Diagram

3.4. User Interface Design

The user interface (UI) diagram depicts the layout of the user interface components that will be used by the user to interact with the system. The UI components include windows, dialog boxes, menus, buttons, and other controls. The UI diagram helps to visualize the flow of information and the navigation between the different screens of the system.

The following is the UI diagram for the drug discovery system:



Drug Discovery for Breast Cancer

Register Login

Aromatase

A Target Protein for Breast Cancer's Drug

Aromatase is an enzyme that plays a key role in the biosynthesis of estrogens. It is responsible for the conversion of androgens, such as testosterone and androstenedione, to estrogens, including estradiol and estrone.

AI can also help identify patient subgroups that may respond better to certain treatments, leading to personalized medicine and better outcomes for breast cancer patients. Overall, AI has the potential to revolutionize the field of drug discovery for breast cancer and improve patient outcomes.

Collaboration and Knowledge Sharing

Can serve as a platform for collaboration and knowledge-sharing among experts in the field, including researchers, pharmaceutical companies, and patient advocacy groups.

DRUG DISCOVERY for Breast Cancer

Copyright 2023. All Rights Reserved

Register

First Name: Enter your first name

Last Name: Enter your last name

Email: Enter your email

Password: Enter your password

Confirm Password: Confirm your password

Already have an account? Login here.

Register

Login

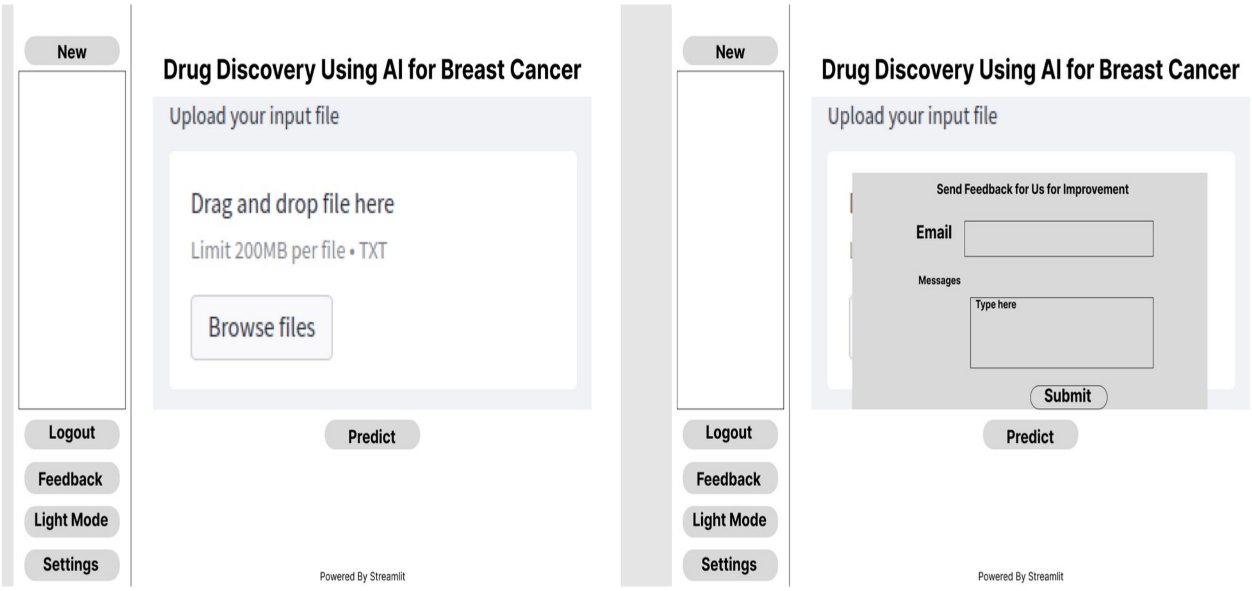
Email: Enter your email

Password: Enter your password

Don't have an account yet? Register here.

Login

Figure 20: Landing, Login and Register Pages



New

Logout

Feedback

Light Mode

Settings

Drug Discovery Using AI for Breast Cancer

Upload your input file

Drag and drop file here

Limit 200MB per file • TXT

Browse files

Predict

Powered By Streamlit

New

Logout

Feedback

Light Mode

Settings

Drug Discovery Using AI for Breast Cancer

Upload your input file

Send Feedback for Us for Improvement

Email

Messages

Type here

Submit

Predict

Powered By Streamlit

Figure 21: Streamlit App UI Part

3.5. Access Control and Security

The following table outlines the access control and security measures implemented in the system:

Table 17: Access Control and Security of the System

Access Control and Security Measures	Description
Authentication	Users are required to log in with a username and password to access the system. Passwords are encrypted and stored securely in the database.
Authorization	Users are assigned roles that determine their level of access to the system. Access is granted on a need-to-know basis, and users are only able to view and modify data that they have been granted permission to access.
Session Management	User sessions are managed to ensure that users are automatically logged out of the system after a period of inactivity, and that multiple users are not able to log in using the same account simultaneously.
Encryption	All data transmitted between the user and the system is encrypted using industry-standard encryption protocols to ensure that it cannot be intercepted or read by unauthorized parties.
Backup and Recovery	Regular backups of the system are taken to ensure that data is not lost in the event of a system failure or other disaster. Additionally, procedures are in place for recovering data in the event of a breach or other security incident.
Audit Trail	The system logs all user activity, including logins, modifications to data, and other events, in an audit trail. This information is used to track and investigate any security incidents that occur within the system.
Physical Security	The servers hosting the system are housed in a secure data center with 24/7 security and environmental controls to prevent unauthorized access or damage to the hardware.

These access control and security measures are designed to ensure the confidentiality, integrity, and availability of data in the system, as well as to protect against unauthorized access and data breaches.

4. Chapter Four: Implementation

The implementation phase is the process of turning the design into a working system. In this phase, we will write code, integrate and test components, and prepare the software for deployment.

4.1. Development Environment

The development environment includes the tools and technologies used for coding, debugging, testing, and version control. The following tools were used for the implementation of this project:

- Programming language: Python, HTML, CSS and JavaScript
- Integrated development environment (IDE): cloud-based Notebooks - Colabs
- Version control: Git
- Web framework: Streamlit
- Machine learning libraries: Scikit-learn

4.2. Coding Standards

To ensure consistency and maintainability of the code, the following coding standards were followed:

- PEP 8 for Python code style
- Docstrings for documenting functions and classes
- Clear and meaningful variable names

4.3. Testing

Testing is an important part of the implementation phase to ensure that the software meets the requirements and functions as intended. The following types of tests were performed:

- **Unit testing:** Individual functions and modules were tested using the Python unittest library.
- **Integration testing:** Multiple components were tested together to ensure they work correctly.
- **User acceptance testing:** The software was tested by end-users to ensure it meets the requirements.

4.4. Deployment

The software was deployed on a web server for online use. The following steps were taken for deployment:

- The code was pushed to a GitHub repository.
- The server environment was set up with the required packages and dependencies.
- The Streamlit web application was deployed using a GitHub server.
- The machine learning models were deployed using a Streamlit API.
- The HTML part of the code will be deployed on GitHub Page Functionality.

4.5. Maintenance

After the software has been deployed, it is important to ensure that it is maintained and updated as needed. The following tasks are performed for maintenance:

- **Bug fixes:** Any issues reported by users are addressed and resolved as quickly as possible.
- **Updates:** The software is updated with new features and improvements.
- **Security:** Regular security audits are performed to ensure the software is secure and protected from vulnerabilities.

Overall, the implementation phase is a critical step in the software development process. With careful planning, testing, and deployment, the software can be developed to meet the requirements and provide value to its users.

References

- BiT Project Documentation format of the students
- Wikipedia
- [1] Aliper, A., Plis, S., Artemov, A., Ulloa, A., Mamoshina, P., & Zhavoronkov, A. (2016). Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. Molecular pharmaceutics, 13(7), 2524-2530.
- [2] Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., ... & Greene, C. S. (2018). Opportunities and obstacles for deep learning in biology and medicine. Journal of The Royal Society Interface, 15(141), 20170387.
- [3] Huang, L., Ma, X., Wang, J., & Li, W. (2018). A convolutional neural network-based framework for predicting side effects of drugs. Bioinformatics, 34(13), i457-i467.
- [4] Keum, J., & Park, H. (2019). An overview of computational approaches for drug discovery. Computers in biology and medicine, 104, 296-305.
- [5] Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., ... & Pande, V. (2018). MoleculeNet: a benchmark for molecular machine learning. Chemical science, 9(2), 513-530.

Appendices

- **Activity Diagram(AD)** – is graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency.
- **Aromatase** - is an enzyme that plays a critical role in the biosynthesis of estrogens, which are hormones that are important for the development and function of the female reproductive system.
- **Artificial Intelligence(AI)** – is intelligence → perceiving, synthesizing, and inferring information → demonstrated by machines, as opposed to intelligence displayed by non-human animals and humans.
- **Breast Cancer** - is a type of cancer that occurs when cells in the breast tissue grow uncontrollably.
- **ChEMBL** - is a database of bioactive molecules and their drug-like properties, maintained by the European Bioinformatics Institute.
- **Colabs** - Google Colaboratory, commonly referred to as "Colab," is a cloud-based service provided by Google that allows users to create and run Jupyter Notebook-style documents for data analysis and machine learning tasks.
- **Drug** - is a substance that is used to diagnose, cure, treat, or prevent a disease or medical condition.
- **Firebase** - is a mobile and web application development platform developed by Google.
- **Non-Functional Requirement(NFR)** – is a requirement that specifies criteria that can be used to judge the operation of a system, rather than specific behaviors. They are contrasted with functional requirements that define specific behavior or function.
- **PEP 8** - is a style guide for writing Python code, which provides guidelines and best practices for formatting, naming, and organizing Python code.
- **Sequence Diagram(SD)** - shows process interactions arranged in time sequence in the field of software engineering. It depicts the processes involved and the sequence of messages exchanged between the processes needed to carry out the functionality.
- **Serverless** - is a term used to describe a cloud computing model where the cloud provider manages the infrastructure required to run an application, allowing developers to focus solely on building the application logic.
- **Scikit-learn** - also known as sklearn, is a popular machine learning library for the Python programming language.
- **Streamlit** – is an open-source app framework for Machine Learning and Data Science team and can create beautiful web apps in minutes.
- **User Interface(UI)** - In the industrial design field of human–computer interaction, a user interface is the space where interactions between humans and machines occur.
- **Web** – is World Wide Web, commonly known as the Web, is an information system enabling documents and other web resources to be accessed over the Internet.