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
- **CRC** – Class Responsibility Collaborator
- **FR** – Functional Requirement
- **GB** – Giga Byte
- **iOS** – iPhone Operating System
- **Linux** – Lovable Intellect Not Using XP
- **MacOS** – Macintosh Operating System
- **NFR** – Non-Functional Requirement
- **OO –** Object Oriented
- **OOA** – Object Oriented Analysis
- **OOD** – Object Oriented Design
- **OS** – Operating System
- **PC** – Personal Computer
- **RAM** – Random Access Memory
- **SD –** Sequence Diagram
- **UI** – User Interface
- **UML –** Unified Modeling Language
- **WINDOWS** – Wide Interactive Network Development for Office Work Solution.

List of Figures

Table of Figures

List of Tables

Index of Tables

Table 1: List of Team members with their Respective Tasks.....ii
Table 2: Description of Functional Requirement.....12
Table 3: Data Integration Use Case Doc.....14
Table 4: Machine Learning Use Case Doc.....14
Table 5: High-Throughput Screening Use Case Doc.....15
Table 6: Web-Based Interface Use Case Doc.....15
Table 7: Business Rule Documentation.....16
Table 8: Change Cases of the Project.....23
Table 9: Access Control and Security of the System.....28

Table of Contents

Declaration....i
Roles and Responsibilities....ii
Acknowledgment....iii
List of Acronym....iv
List of Figures....v
List of Tables....vi
Abstract....1
1. Chapter One: Introduction....2
1.1. Background....2
1.2. Objectives....3
1.2.1 General Objectives....3
1.2.2 Specific Objectives....3
1.3. Statement of the Problem....4
1.4. Beneficiaries of the Project....4
1.5. Limitations of the Project....5
1.6. Scope of the Project....5
1.7. Methodology....6
1.7.1 Data Acquisition....6
1.7.2 Data Pre-processing....6
1.7.3 Feature Engineering....7
1.7.4 Machine Learning Algorithms....7
1.7.5 Molecular Docking Simulations....7
1.7.6 Website Development....7
1.8. Feasibility Study....7
1.8.1 Technical Feasibility....7
1.8.2 Economic Feasibility....8
1.8.3 Legal Feasibility....8
1.8.4 Operational Feasibility....8
1.9. Organization of the Project....8
1.9.1 Planning Phase....8
1.9.2 Data Acquisition and Pre-processing Phase....8
1.9.3 Feature Engineering Phase....8
1.9.4 Machine Learning Algorithms Phase....9
1.9.5 Molecular Docking Simulations Phase....9
1.9.6 Website Development Phase....9
1.9.7 Testing and Validation Phase....9
1.9.8 Documentation and Reporting Phase....9
2. Chapter Two: System Features....10
2.1. The Existing System....10
2.2. Proposed System....10
2.3. Requirement Analysis....11
2.3.1 Functional Requirements....12

2.3.2 Non-Functional Requirements....13
2.3.3 Constraints and Limitations....13
2.3.4 System Use case....13
2.3.5 Business Rule Documentation....16
2.3.6 User Interface Prototype....17
2.3.7 Activity Diagram....18
2.3.8 Sequence Diagram....19
2.3.9 Analysis Class Model....20
2.3.10 Logic Model....21
2.4. Key Abstraction with CRC Analysis....21
2.5. Change Case....23
3. Chapter Three: System Design....25
3.1. Introduction....25
3.2. Architectural Design....25
3.2.1 Component Modeling....26
3.3. Deployment Modeling....27
3.4. Detail Design....27
3.4.1 Design Class Model....27
3.5. User Interface Design....27
3.6. Access Control and Security....28
4. Chapter Four: Implementation....30
4.1. Development Environment....30
4.2. Coding Standards....30
4.3. Testing....30
4.4. Deployment....31
4.5. Maintenance....31
References....32
Appendices....33

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 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 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 patient 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

[Use Case Diagram Goes Here]

In this diagram, there are three actors: the researcher, the clinician, and the system administrator. The system itself has four use cases: data integration, machine learning, high-throughput screening, and web-based interface. Each use case represents a distinct action or activity that the system must be capable of performing.

Here's a brief description of each use case:

- **Data Integration:** The system must be able to integrate data from multiple sources, including genomic, transcriptomic, proteomic, and clinical data.
- **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-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.
- **Web-Based 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.

Each use case is further elaborated in a use case document, which includes the following information:

Table 3: Data Integration Use Case Doc

Use Case Name	Data Integration
Actor	System
Description	Integrate multiple data sources to provide a more comprehensive understanding of breast cancer biology and potential drug targets.
Pre-conditions	Relevant data sources are available and accessible.
Post-conditions	Data from multiple sources is integrated and available for analysis.
Flow of Events	- The system identifies relevant data sources. - The system retrieves data from the identified sources. - The system transforms the data into a common format. - The system integrates the transformed data. - The system stores the integrated data.
Exceptions	- Data sources are not accessible. - Data from sources is incomplete or inconsistent.
Priority	High

Table 4: Machine Learning Use Case Doc

Use Case Name	Machine Learning
Actor	System
Description	Apply machine learning algorithms to identify potential drug candidates from integrated data sources.
Pre-conditions	Integrated data sources are available.
Post-conditions	Potential drug candidates are identified and ranked based on predicted efficacy and safety.
Flow of Events	The system preprocesses and prepares the data for machine learning

	analysis. - The system trains machine learning models using the prepared data. - The system applies the trained models to the integrated data to identify potential drug candidates. - The system ranks the potential drug candidates based on predicted efficacy and safety.
Exceptions	- The prepared data is incomplete or inconsistent. - The machine learning models fail to identify any potential drug candidates.
Priority	High

Table 5: High-Throughput Screening Use Case Doc

Use Case Name	High-Throughput Screening
Actor	System
Description	Conduct high-throughput screening to rapidly test potential drug candidates in vitro and in vivo.
Pre-conditions	Potential drug candidates have been identified through machine learning analysis.
Post-conditions	Efficacy and safety data for potential drug candidates is available for further analysis.
Flow of Events	- The system prepares the potential drug candidates for high-throughput screening. - The system conducts in vitro high-throughput screening of the prepared drug candidates. - The system conducts in vivo high-throughput screening of the top-ranked drug candidates from in vitro screening. - The system collects efficacy and safety data from the high-throughput screening assays.
Exceptions	- The potential drug candidates fail in the high-throughput screening assays. - The high-throughput screening assays produce incomplete or inconsistent data.
Priority	High

Table 6: Web-Based Interface Use Case Doc

Use Case Name	Web-Based Interface
Actor	Researcher, Clinician
Description	Provide a user-friendly web-based interface to allow researchers and clinicians to easily access and interpret the results of the drug discovery pipeline.

Pre-conditions	Data from all previous stages of the drug discovery pipeline is available.
Post-conditions	Researchers and clinicians are able to access and interpret the results of the drug discovery pipeline.
Flow of Events	- The user navigates to the web-based interface. - The user logs in with their credentials. - The user selects the relevant data and analyses from the drug discovery pipeline. - The user is presented with the relevant data and analyses in an easily understandable format.
Exceptions	The use is unable to log in due to incorrect credentials.
Priority	Medium

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 7: Business Rule Documentation

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

Introduction:

The purpose of this document is to provide an overview of the User Interface Prototype for the proposed system. The User Interface Prototype is a visual representation of how the system will look and function for the end user.

User Interface Prototype:

The User Interface Prototype consists of a series of screen mockups that demonstrate the system's functionality and user flow. The screens are designed to be user-friendly, intuitive, and visually appealing. The User Interface Prototype is divided into the following sections:

1. Login Screen
2. Home Screen
3. Search Screen
4. Results Screen
5. Detail Screen
6. Analysis Screen

Each section is described in more detail below.

1. **Login Screen:** The Login Screen is the first screen that the user sees when they access the system. It consists of a username and password field, as well as a "Forgot Password" link.
2. **Home Screen:** The Home Screen is the main screen of the system. It consists of a search bar, a list of recent searches, and a menu bar.
3. **Search Screen:** The Search Screen allows the user to search for data based on various criteria such as drug name, chemical structure, target protein, and disease type.
4. **Results Screen:** The Results Screen displays a list of results based on the search criteria entered by the user. Each result is displayed with a brief summary of the data.
5. **Detail Screen:** The Detail Screen provides detailed information about a specific result, including chemical properties, protein interactions, and relevant research articles.
6. **Analysis Screen:** The Analysis Screen allows the user to perform various data analysis tasks such as clustering, classification, and regression analysis.

Conclusion: The User Interface Prototype provides a visual representation of the proposed system's functionality and user flow. It is designed to be user-friendly, intuitive, and visually

appealing. The User Interface Prototype will be used as a guide for the development of the actual system, ensuring that the end product meets the needs of the users.

2.3.7 Activity Diagram

The Activity Diagram 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. User Search Activity
3. Data Retrieval Activity
4. Data Analysis Activity
5. Results Display Activity

Each section is described in more detail below.

1. **Login Activity:** Represents the process of getting into the system with user credentials.
2. **User Search Activity:** The User Search Activity represents the user's input of search criteria, such as drug name, chemical structure, target protein, and disease type. Once the user has entered the search criteria, the system proceeds to the Data Retrieval Activity.
3. **Data Retrieval Activity:** The Data Retrieval Activity represents the system's retrieval of data based on the search criteria entered by the user. The system searches the database for relevant data and retrieves it for analysis. Once the data has been retrieved, the system proceeds to the Data Analysis Activity.
4. **Data Analysis Activity:** The Data Analysis Activity represents the system's analysis of the retrieved data. The system performs various data analysis tasks such as clustering, classification, and regression analysis. Once the data analysis is complete, the system proceeds to the Results Display Activity.
5. **Results Display Activity:** The Results Display Activity represents the system's display of the analyzed data to the user. The system displays the results in an easy-to-understand format, such as a list or a chart. Once the results are displayed, the user can choose to perform another search or exit the system.

Conclusion: The Activity Diagram provides a visual representation of the flow of activities within the proposed system, from the user's perspective. It is designed to ensure that the system meets the needs of the users and that the activities are logically connected. The Activity Diagram will be used as a guide for the development of the actual system, ensuring that the end product meets the needs of the users.

[Activity Diagram Goes Here]

2.3.8 Sequence Diagram

Introduction: The purpose of this document is to provide an overview of the Sequence Diagram for the proposed system. The Sequence Diagram is a graphical representation of the interactions between different system components, from the user's perspective.

Sequence Diagram:

The Sequence Diagram 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. User Search Sequence
3. Data Retrieval Sequence
4. Data Analysis Sequence
5. Results Display Sequence

Each section is described in more detail below.

1. **Login Sequence:** Represents the sequence of the user authenticate them and access the system for use.
2. **User Search Sequence:** The User Search Sequence represents the user's input of search criteria, such as drug name, chemical structure, target protein, and disease type. The user input is sent to the system for processing.
3. **Data Retrieval Sequence:** The Data Retrieval Sequence represents the system's retrieval of data based on the search criteria entered by the user. The system searches the database for relevant data and retrieves it for analysis. The retrieved data is sent to the Data Analysis Sequence.
4. **Data Analysis Sequence:** The Data Analysis Sequence represents the system's analysis of the retrieved data. The system performs various data analysis tasks such as clustering, classification, and regression analysis. Once the data analysis is complete, the analyzed data is sent to the Results Display Sequence.
5. **Results Display Sequence:** The Results Display Sequence represents the system's display of the analyzed data to the user. The system displays the results in an easy-to-understand

format, such as a list or a chart. The user can interact with the displayed results and choose to perform another search or exit the system.

Conclusion: The Sequence Diagram provides a visual representation of the interactions between different system components within the proposed system, from the user's perspective. It is designed to ensure that the system meets the needs of the users and that the components are logically connected. The Sequence Diagram will be used as a guide for the development of the actual system, ensuring that the end product meets the needs of the users.

2.3.9 Analysis Class Model

Introduction: The purpose of this document is to provide an overview of the Analysis Class Model for the proposed system. The Analysis Class Model is a high-level representation of the system's objects, classes, and their relationships. It shows how the classes interact with each other and the data they hold.

Analysis Class Model:

The Analysis Class Model includes the following classes:

1. User
2. Search Criteria
3. Data Retrieval
4. Data Analysis
5. Results Display

Each class is described in more detail below.

1. **User:** The User class represents the user of the system. It holds the user's information such as name, email, and password. The User class has a relationship with the Search Criteria class.
2. **Search Criteria:** The Search Criteria class represents the user's input of search criteria, such as drug name, chemical structure, target protein, and disease type. It holds the search criteria entered by the user and has a relationship with the Data Retrieval class.
3. **Data Retrieval:** The Data Retrieval class represents the system's retrieval of data based on the search criteria entered by the user. It searches the database for relevant data and retrieves it for analysis. The Data Retrieval class has a relationship with the Data Analysis class.
4. **Data Analysis:** The Data Analysis class represents the system's analysis of the retrieved data. It performs various data analysis tasks such as clustering, classification, and

regression analysis. The Data Analysis class has a relationship with the Results Display class.

5. **Results Display:** The Results Display class represents the system's display of the analyzed data to the user. It displays the results in an easy-to-understand format, such as a list or a chart. The Results Display class has a relationship with the User class.

Conclusion: The Analysis Class Model provides a high-level representation of the system's objects, classes, and their relationships. It shows how the classes interact with each other and the data they hold. The Analysis Class Model will be used as a guide for the development of the actual system, ensuring that the end product meets the needs of the users.

2.3.10 Logic Model

Introduction: The purpose of this document is to provide an overview of the Logic Model for the proposed system. 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.

Logic Model: The Logic Model consists of four main components: inputs, activities, outputs, and outcomes.

1. **Inputs:** The inputs are the resources required for the system's development and implementation, including the AI algorithms, programming languages, and hardware.
2. **Activities:** The activities are the tasks that the system will perform to achieve its objectives. These include data retrieval, data analysis, and results display.
3. **Outputs:** The outputs are the tangible results of the system's activities. These include the analyzed data and the results displayed to the user.
4. **Outcomes:** The outcomes are the intended impact of the system on the users and the broader community. These include improved accuracy and efficiency of breast cancer drug discovery, reduced development time and cost, and ultimately, better patient outcomes.

The Logic Model also includes several assumptions and external factors that may influence the success of the system, such as the availability of data, the quality of the AI algorithms, and the level of user adoption.

Conclusion: The Logic Model provides a clear and concise overview of the system's inputs, activities, outputs, and outcomes, as well as the assumptions and external factors that may affect its success. It will serve as a roadmap for the system's design, implementation, and evaluation, ensuring that it achieves its intended impact on the breast cancer drug discovery process.

2.4. Key Abstraction with CRC Analysis

Introduction: This document provides an overview of the CRC Analysis and key abstractions for the proposed system. The CRC Analysis is a technique for identifying the classes and their responsibilities in a software system. It also identifies the collaboration between these classes to achieve the system's objectives.

Key Abstractions:

The key abstractions for the system are:

1. **Breast Cancer Dataset:** The dataset will contain the patient's data, including their medical history and diagnostic test results.
2. **AI Algorithms:** The AI algorithms will analyze the dataset and provide insights into the breast cancer drug discovery process.
3. **User Interface:** The user interface will allow users to interact with the system, input data, and view the results.
4. **Database:** The database will store the analyzed data and provide a platform for future analysis.

CRC Analysis:

The CRC Analysis identified the following classes and their responsibilities:

1. Data Retrieval Class:

Responsibilities:

- Retrieve the breast cancer dataset from external sources
- Store the dataset in the database

Collaboration:

- Collaborates with the Database Class

2. AI Algorithms Class:

Responsibilities:

- Analyze the breast cancer dataset to identify drug discovery patterns
- Generate drug discovery insights

Collaboration:

- Collaborates with the Data Retrieval Class and Database Class

3. User Interface Class:

Responsibilities:

- Display the breast cancer dataset and drug discovery insights
- Allow the user to input data for analysis

Collaboration:

- Collaborates with the AI Algorithms Class and Database Class

4. Database Class:

Responsibilities:

- Store the breast cancer dataset
- Store the analyzed data
- Provide a platform for future analysis

Collaboration:

- Collaborates with the Data Retrieval Class, AI Algorithms Class, and User Interface Class

Conclusion:

The CRC Analysis and key abstractions provide a clear and concise overview of the system's classes, responsibilities, and collaboration. It will guide the system's design and implementation, ensuring that it achieves its intended impact on the breast cancer drug discovery process.

2.5. Change Case

Change cases 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 8: 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 interface	The user interface should be continually evaluated and 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

Introduction:

The purpose of this document is to provide an overview of the architecture of the drug discovery project using AI for breast cancer. The document is aimed at developers and technical stakeholders, and serves as a guide for implementing the system.

Architecture:

The architecture of the system is based on a client-server model, with a web-based user interface for accessing the system. The server-side component of the system is responsible for data processing and analysis, while the client-side component provides a user-friendly interface for interacting with the system.

Server-side Architecture:

The server-side component of the system is built using Python, and relies on a number of third-party libraries for data processing and analysis. The system architecture includes the following components:

- **Data Ingestion:** This component is responsible for ingesting data from various sources, including clinical trials and genomic data repositories.
- **Data Processing:** This component is responsible for processing and cleaning the data, preparing it for analysis.
- **Machine Learning:** This component is responsible for running machine learning algorithms on the processed data to identify potential drug candidates.
- **Database:** This component is responsible for storing and retrieving data used by the system.

Client-side Architecture:

The client-side component of the system is built using HTML, CSS, and JavaScript, Streamlit, and is designed to be accessed through a web browser. The system architecture includes the following components:

- **User Interface:** This component provides a user-friendly interface for interacting with the system, allowing users to submit queries and view results.
- **Visualization:** This component is responsible for displaying data and analysis results in a visual format, making it easier for users to interpret the results.

Implementation:

The system is implemented using a number of tools and technologies, including:

- **Python:** Used for building the server-side component of the system.
- **Python:** Used as the web framework for building the server-side component of the system.
- **HTML, CSS, Streamlit, and JavaScript:** Used for building the client-side component of the system.
- **Pandas:** Used for data visualization.

Conclusion:

The architecture of the drug discovery project using AI for breast cancer is designed to be flexible and scalable, allowing for easy integration of new data sources and machine learning algorithms as they become available. The implementation of the system relies on a number of industry-standard tools and technologies, ensuring that it is both reliable and efficient.

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.

The main components of the drug discovery system using AI for breast cancer are:

1. **Data Collection and Processing:** This component is responsible for collecting and processing various types of data such as genomic, proteomic, clinical, and demographic data of patients.
2. **Machine Learning Model:** This component uses the collected data to train a machine learning model, which can predict the presence of breast cancer based on various factors.
3. **Prediction and Evaluation:** This component uses the trained model to predict the presence of breast cancer for new patients and evaluate the accuracy of the predictions.

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.

The deployment model of the drug discovery system using AI for breast cancer is as follows:

1. **Data Collection and Processing Component:** This component is deployed on a cloud-based server.
2. **Machine Learning Model Component:** This component is deployed on a cloud-based server.
3. **Prediction and Evaluation Component:** This component is deployed on a web server.
4. **User Interface Component:** This component is deployed on a web server.

Here is a graph to illustrate the deployment model:

[The Diagram Goes Here]

The cloud-based server can handle a large amount of data and processing power required for training the machine learning model. The web server provides an interface for users to access the system from their web browsers. By separating the components, the system can be scaled up or down based on the user demand.

3.4. Detail Design

The detail class model represents the specific classes and their attributes and methods that will be implemented in the system. This model is based on the analysis class model and is further refined to include more specific details.

3.4.1 Design Class Model

The class diagram below shows the detailed class structure of the system, including the classes, their attributes, and methods:

[The Diagram Goes Here]

3.5. 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:

[Insert image of the UI diagram here]

The UI diagram shows the different screens of the system, including the main dashboard, the search screen, the drug details screen, and the report screen. The main dashboard displays the key performance indicators (KPIs) for the system, such as the number of drugs in the system, the number of searches performed, and the number of reports generated. The search screen allows the user to search for drugs by various criteria, such as drug name, target protein, and disease indication. The drug details screen displays the details of a selected drug, including its chemical structure, mechanism of action, and clinical trial results. The report screen allows the user to generate reports based on the data in the system.

The UI components are designed to be intuitive and user-friendly, with clear and concise labels and instructions. The layout and design of the screens are optimized for ease of use and readability, with a consistent color scheme and typography. The system is designed to be responsive, with the UI components adapting to different screen sizes and resolutions, including desktop, tablet, and mobile devices.

In conclusion, the UI diagram provides a visual representation of the user interface components of the drug discovery system. The UI components are designed to be intuitive, user-friendly, and responsive, with a consistent layout and design across different screens and devices.

3.6. Access Control and Security

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

Table 9: 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, the developers 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
- Integrated development environment (IDE): Notebooks
- Version control: Git
- Package manager: Pipenv
- Web framework: Streamlit
- Machine learning libraries: Scikit-learn, Keras, TensorFlow

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.

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.
- **Artificial Intelligence(AI)** – is intelligence → perceiving, synthesizing, and inferring information → demonstrated by machines, as opposed to intelligence displayed by non-human animals and humans.
- **Class Responsibility Collaborator(CRC)** – is cards that are a brainstorming tool used in the design of object-oriented software. They originally proposed by Ward Cunningham and Kent Beck as a teaching tool but are also popular among expert designers and recommended by extreme programming.
- **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.
- **Object Oriented(OO)** – is a programming paradigm based on the concept of "objects", which can contain data and code. The data is in the form of fields, and the code is in the form of procedures. A common feature of objects is that procedures are attached to them and can access and modify the object's data fields.
- **Object Oriented Analysis and Design (OOA & OOD)** – is a technical approach for analyzing and designing an application, system, or business by applying object-oriented programming, as well as using visual modeling throughout the software development process to guide stakeholder communication and product quality.
- **Operating System(OS)** – is system software that manages computer hardware, software resources, and provides common services for computer programs.
- **Personal Computer(PC) -** is a multi-purpose microcomputer whose size, capabilities, and price make it feasible for individual use.
- **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.
- **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.
- **Unified Modeling Language(UML) -** is a general-purpose, developmental modeling language in the field of software engineering that is intended to provide a standard way to visualize the design of a system.
- **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. Documents and downloadable media are made available to the network through web servers and can be accessed by programs such as web browsers.