PATHWAY GENIE

NAELAH SULTAN MIYAJAN
BADOR RAKKAN ALGHALIB
RAHAF MOHAMMED ALWAJEEH
MARYAM ABDULRHMAN BAMBEROOK

DEPARTMENT OF COMPUTER SCIENCE
KING ABDULAZIZ UNIVERSITY

February, 2023

PATHWAY GENIE

NAELAH SULTAN MIYAJAN
BADOR RAKKAN ALGHALIB
RAHAF MOHAMMED ALWAJEEH
MARYAM ABDULRHMAN BAMBEROOK

THIS REPORT IS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE BACHELOR DEGREE
IN COMPUTER SCIENCE

DR. NOFE ALGANMI

DEPARTMENT OF COMPUTER SCIENCE
FACULTY OF COMPUTERS INFORMATION TECHNOLOGY
KING ABDULAZIZ UNIVERSITY

February, 2023

Declaration of Originality

We hereby declare that this project report is based on our original work except for citations and quotations, which had been duly acknowledged. We also declare that it has not been previously and concurrently submitted for any other degree or award at KAU or other institutions.

Student Name	ID	Signature	Date
Naelah Sultan Miyajan	1905313		February, 2023
Bador Rakkan Alghalib	1905253		February, 2023
Rahaf Mohammed alwajeeh	1909711		February, 2023
Maryam Abdulrhman Bamberook	1906316		February, 2023

Acknowledgements

We would like to thank our supervisor Dr. Nofe Alghanmi who gave us a golden opportunity to work on this project, her continuous support and guidance have really helped us in this journey. we would also like to express our gratitude to Prof. Arwa Alaama wholeheartedly for her guidance. Lastly, we must also thank our parents and friends for their immense support and help during this project and our incredible team for their cooperation and support in difficult times and situations. We would like to thank Abdulaziz, Talah, and Abeer Hadadi for their encouragement.

Abstract

Pathway Genie is a web-based tool that reads gene expression data and biological pathways of these genes in order to generate a comprehensive report analyzing the data and producing precise results in a short amount of time. This tool will be useful for geneticists and all researchers. Our interest in pathway detection analysis was sparked by the notion that some bodies can have mutations or diseases while others are healthy. Additionally, finding the first compromised biological pathway might point us in the direction of other compromised biological pathways and make it easier for geneticists to retrieve data. With the help of using artificial intelligence, a model was successfully built to predict and distinguish between tumor and non-tumor samples with an accuracy of 96%.

المُستخلص

باث واي جيني هي أداة قائمة على الويب تقرأ بيانات التعبير الجيني والمسارات البيولوجية لهذه الجينات من أجل إنشاء تقرير شامل يحلل البيانات وينتج نتائج دقيقة في فترة زمنية قصيرة. ستكون هذه الأداة مفيدة لعلماء الوراثة وجميع الباحثين. الذي لفت انتباهنا إلى فكرة تحليل اكتشاف المسار أن بعض الأجسام يمكن أن يكون لها طفرات أو أمراض بينما يتمتع البعض الآخر بصحة جيدة.

بالإضافة إلى ذلك ، فإن العثور على أول مسار بيولوجي مخترق قد يوجهنا في اتجاه المسارات البيولوجية الأخرى المعرضة للخطر ويسهل على علماء الوراثة استرداد البيانات. بمساعدة استخدام الذكاء الاصطناعي ، تم بناء نموذج بنجاح للتنبؤ والتمييز بين عينات الورم وغير الورم بدقة ٩٦٪.

Contents

Declaration of Originality ii

Acknowledgements iii

Abstract iv

Table of Contents v

List of Tables viii

List of Figures ix

List of Abbreviations xi

List of Appendices xii

1 INTRODUCTION 1
- 1.1 Introduction . . . 1
- 1.2 Motivation . . . 3
- 1.3 Problem Statement . . . 3
- 1.4 Proposed Solution . . . 3
- 1.5 Aim and Objectives . . . 3
- 1.6 Target Users . . . 4
- 1.7 Scope . . . 4
- 1.8 Methodology . . . 4
- 1.9 Plan and Milestones . . . 7
- 1.10 Conclusion . . . 7

2 LITERATURE REVIEW 8
- 2.1 Introduction . . . 8
- 2.2 Genetic Pathways . . . 8
- 2.3 Bio-informatics . . . 10
- 2.4 Artificial Intelligence . . . 12
- 2.5 Data Science . . . 12
  - 2.5.1 Machine Learning . . . 13
  - 2.5.2 Data Mining . . . 14
  - 2.5.3 Phases of Data Mining . . . 15
  - 2.5.4 Differences Between Data Mining and Machine Learning . 17
- 2.6 Database . . . 17
  - 2.6.1 Distributed Database . . . 18
  - 2.6.2 Data Virtualization . . . 20
- 2.7 Precision Medicine . . . 20

2.8 Related Gene Expression Tools 21
2.8.1 GeoQuery Tool 21
2.8.2 LIMMA Tool 22
2.8.3 BART Tool 22
2.8.4 SAGE Tool 23
2.9 Related AI Researches 25
2.9.1 AI Technique for Gene Expression by Tumor RNA-Seq Data 25
2.9.2 Enhanced Deep Learning Model for Personalized Cancer Treatment 25
2.9.3 Deep Learning and Machine Learning via a Genetic Algorithm to Classify Breast Cancer DNA Data 26
2.9.4 Alzheimer's Patient Analysis Using Image and Gene Expression Data and Explainable-AI to Present Associated Genes 26
2.10 Conclusion 28

3 DATA COLLECTION AND REQUIREMENTS 29
3.1 Introduction 29
3.2 Dataset 29
3.3 Data Collection 31
3.3.1 Interviews 31
3.3.2 Questionnaire 32
3.4 Establishing Requirements 32
3.4.1 Functional Requirements 32
3.4.2 Non-Functional Requirements 33
3.5 Conclusion 33

4 ARCHITECTURE AND DESIGN 34
4.1 Introduction 34
4.2 Diagrams 35
4.2.1 Use cases 35
4.2.2 Class Diagram 40
4.2.3 Database diagrams 42
4.2.4 Sequence Diagram 44
4.2.5 Flowchart 46
4.2.6 Data flow 48
4.3 Prototype 48
4.4 Conclusion 53

5 IMPLEMENTATION 54
5.1 Summary of Previous Work 54
5.2 Activities and Actions of the Development 54
5.3 Tools and platform 55
5.3.1 HTML 55
5.3.2 CSS 55
5.3.3 MySQL 56
5.3.4 PHP 56
5.3.5 RStudio 57
5.3.6 Python 57

5.3.7 JavaScript . . . 57
5.3.8 Visual Studio . . . 57
5.3.9 Formspree . . . 57
5.4 Front-end Implementation . . . 58
5.5 Database Implementation . . . 69
5.6 AI Model Implementation . . . 70
5.6.1 Data Annotation . . . 70
5.6.2 Classification . . . 70
5.6.3 Results . . . 71
5.7 Conclusion . . . 75

6 TESTING 76
6.1 Introduction . . . 76
6.2 Purpose and Objectives . . . 76
6.3 System Components Testing . . . 77
6.4 Types of testing . . . 77
6.5 Unit Testing . . . 77
6.5.1 AI Model Unit Testing . . . 77
6.5.2 Database Unit Testing . . . 78
6.6 Usability Testing . . . 81
6.6.1 Participants . . . 81
6.6.2 Environment of the Test . . . 81
6.6.3 Evaluation Tasks . . . 82
6.6.4 Objective Measure Analysis . . . 83
6.6.5 Subjective Measure Analysis . . . 85
6.7 Conclusion . . . 88

7 CONCLUSION AND FUTURE WORK 89
7.1 Introduction . . . 89
7.2 Challenges . . . 89
7.3 Learned skills . . . 90
7.4 Future work . . . 90

References 90

1 Interview Questions 95

2 Survey Questions 97

3 Usability Testing Survey 100

List of Tables

1.1 Project Plan 7

2.1 Related Gene Expression Tools 24
2.2 Related AI Researches 27

4.1 Create account table 36
4.2 Log into account table 37
4.3 Submit data file table 38
4.4 Generate final report table 39

List of Figures

1.1 Transcription Vs. Translation 2
1.2 Waterfall Model Phases 5

2.1 Difference between DNA and RNA 10
2.2 Data science life cycle 13
2.3 Machine Learning Infographic 14
2.4 Data Mining Phases 15
2.5 5 ways centralized databases differ from distributed databases . . 19

4.1 Use Case diagram 35
4.2 UML Class Diagram 41
4.3 ER Class Diagram 43
4.4 Database Schema Diagram 44
4.5 Sequence Diagram 45
4.6 Flow Chart 47
4.7 Data Flow diagram 48
4.8 Home page 49
4.9 Sign Up page 49
4.10 Sign In page 50
4.11 Upload page 50
4.12 After Uploading a file 51
4.13 Changing settings 51
4.14 Running pipeline 52
4.15 Final Report Page 53

5.1 Home Page 58
5.2 Login Page 59
5.3 Register Page 59
5.4 Home Page when logged in 60
5.5 Login Empty Email 61
5.6 Login Empty Password 61
5.7 Wrong Password or Email 61
5.8 Contact Us Page 62
5.9 Username drop-down list 62
5.10 Account page 63
5.11 Update Account 63
5.12 Analysis Page 64
5.13 Listed Algorithms Choices 64
5.14 Only CSV files are allowed error 65
5.15 Algorithms information 66
5.16 History Page 67

5.17 Final Report page 68
5.18 Log out page 68
5.19 Pathway Genie File Table 69
5.20 Pathway Genie user Table 70
5.21 Classification Report 72
5.22 Confusion matrix plot 73
5.23 AUC - ROC Curve 75

6.1 New profile account 78
6.2 The account was created in the database 78
6.3 Inserting data 79
6.4 File Uploaded Successfully 79
6.5 The inserted data in the database 80
6.6 Password update 80
6.7 Updated password in the database 81
6.8 Minimum and maximum of the participant's clicks 83
6.9 Minimum and maximum of the participant's duration 84
6.10 Autonomy level of participants 85
6.11 Result of the overall satisfaction of users 85
6.12 Result of how easy the website to use 86
6.13 Pathway genie Application System Testing 87

1.1 Interview result 96

2.1 Questionnaire capture 1 98
2.2 Questionnaire capture 2 99

List of Abbreviations

AI	Artificial Intelligence.
BART	Bioinformatics Array R Tool.
BRCA	Breast Invasive Carcinoma.
CNN	Convolutional Neural Network.
CSS	Cascading Style Sheets.
DEGs	Differentially Expressed Genes.
DNA	Deoxyribo Nucleic Acid.
ER	Entity Relationship.
ETL	Extraction, Transformation, and Loading.
GEO	Gene Expression Omnibus.
HGP	Human Genome Project.
HTML	Hyper Text Markup Language.
ICU	Intensive Care Unit.
IDE	integrated development environment.
KIRC	Kidney Renal Clear Well Carcinoma.
KNN	K-Nearest Neighbors.
LIME	Local Interpretable Model-agnostic Explanations.
LUSC	Lung Squamous Cell Carcinoma.
ML	Machine Learning.
mRNA	Messenger Ribo Nucleic Acid.
MySQL	My Structured Query Language.
PCA	Principal component analysis.
PHP	Hypertext Preprocessor.
RNA	Ribo Nucleic Acid.
RNAseq	Ribo Nucleic Acid Sequencing.
SAGE	Serial Analysis of Gene Expression.
SDD	Software Design Document.
SRS	Software Requirement Specification.
SVC	Support Vector Classifier.
UCEC	Uterine Corpus Endometrial Carcinoma.
UML	Unified Modeling Language.

List of Appendices

Appendix 1 Interview Questions **95**

Appendix 2 Survey Questions **97**

Appendix 3 Usability Testing Survey **100**

CHAPTER 1

INTRODUCTION

1.1 Introduction

In biology, we study living organisms and the key to understand living organisms is to analyze and study their genes. What is a gene and what do we mean by analyzing the gene? A gene is a strand of DNA, and it has nucleotides, these nucleotides (A, C, T, G) are used to synthesize protein, this process is called gene expression. Gene expression involves two main stages: transcription and translation. Transcription is the process by which a DNA template synthesizes RNA where the code in the DNA is converted into a complementary RNA code while, translation is the process by which the mRNA template synthesis a protein where the code within mRNA is translated into an amino acid sequence in a protein.

Now how do we analyze gene expression? by our tool pathway genie Before we explain the tool, Let us first talk about gene expression analysis to understand our tool. Gene expression analysis is defined as the study of the process through which genes are transcribed into functional proteins. Studies can be categorized into four categories: Post-Translation modification, promoter analysis, protein expression, and RNA expression. We will focus on RNA expression specifically in DNA micro-arrays technology. DNA micro-arrays were only used as a tool for study. Massive population studies are still being conducted today by scientists, for example, to find out how frequently people with a specific mutation get breast cancer or to find the variations in gene sequences that are most frequently linked to diseases.

The degree to which particular genes are activated or inactivated in cells

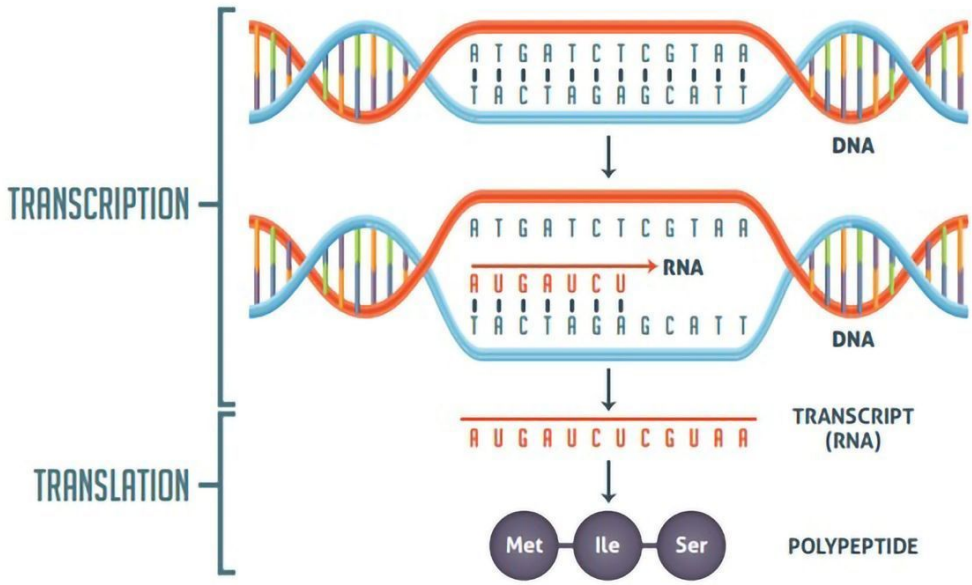


Figure 1.1: Transcription Vs. Translation
Source:(tec, 2021)

and tissues can also be examined using microarrays. In this instance, RNA (which is a transcript of the DNA) is separated and analyzed from the samples rather than isolating the DNA.

Since genes control how our bodies respond to the chemistry of those drugs, DNA micro-arrays are now used to figure out which medications could be best given to certain individuals. This is how we produced pathway genie, a pathway genie is a web-based tool that will use AI to analyze the gene expression data and once we get the list of candidate genes, we will detect the biological pathways using the online databases to also link them to suggested drug lists. There are a lot of drugs and biological pathways that need to be investigated and tested, and we provide pathway genie to help researchers minimize their options, time, and lab testing on research to give them their next suggested step while giving accurate results. in addition, AI technology in pathway genie will pipeline the data so it would provide easy access to the information. Lastly, our tool output is to generate a complete visualized accurate report for the scientists. (mic, 2022)

1.2 Motivation

The idea that some bodies can have mutations/diseases while others are normal, motivated us to work in pathway detection analysis. Furthermore, detecting the first corrupted pathway could lead us to other affected biological pathways and facilitate the data retrieval process for geneticists.

1.3 Problem Statement

Researchers and geneticists put efforts into gene analysis. They generate files that contain matrices of big data and then search to study many databases for specific information. afterward, they do mathematical equations and comparisons of different data sets, followed by manual testing and report writing to detect the pathway of the gene. This process takes a lot of time and is costly.

1.4 Proposed Solution

Our tool, pathway genie, resolves the stated problem by offering a web-based tool that reads gene expression data and the biological pathways of these genes to ultimately generate a full report analyzing the data with accurate results in a matter of minutes.

1.5 Aim and Objectives

This project aims to develop a web-based application that helps bioinformatics specialists to analyze their gene expression data accurately and efficiently.
The objectives of the project are:

- Speeding up the research study process.
- Automating the analysis of gene expression using AI.
- Developing a flexible web-based tool.
- Lowering the cost and effort of gene expression analysis.
- Providing a comprehensive scientific report that helps to increase accuracy.
- Improving the efficiency of the R&D in discovering new bio-markers

1.6 Target Users

Our tool's targeted users are Geneticists, data scientists, bioinformaticians, and data analysts.

1.7 Scope

The scope of our project involves multiple knowledge areas like Artificial Intelligence, Bio-informatics, databases, data science, and precision medicine. Each of these plays an important role in genomics; with AI, we will analyze the gene expression data and get the list of candidate genes after the analysis. Then, detect the Biological pathways using high-dimensional online databases, which will use AI technology to pipeline the data with bio-informatics languages and add deep learning algorithms to connect the data as a network-like database to be able to work with any database type that is large or small. The term "personalized medicine" (also known as "precision medicine") refers to a medical approach that divides people into several groups and tailors medical decisions, practices, procedures, and/or products to each patient depending on that patient's expected response or risk of disease. Our project output is to generate a visualized accurate report for the scientists and Geneticists researchers that has a complete analysis of each patient's pathway and gene expression.

1.8 Methodology

The methodology is a procedure that typically resolves a specific problem and includes a specified set of rules, techniques, tests, actions, deliverables, and processes. The methodology that is applied to our project is iterative Waterfall.the key advantages of this methodlogy, it allows the correction of the errors in every phase and it's simple to use.

1. **Requirement gathering & analysis**

   Understanding the exact requirements of the client and accurately documenting them is the purpose of this phase. Together with the client, the software developer documents all of the software's performance, interface, and function requirements. It explains the "what," not the "how," of the system to be created. During this phase, a large document known as the

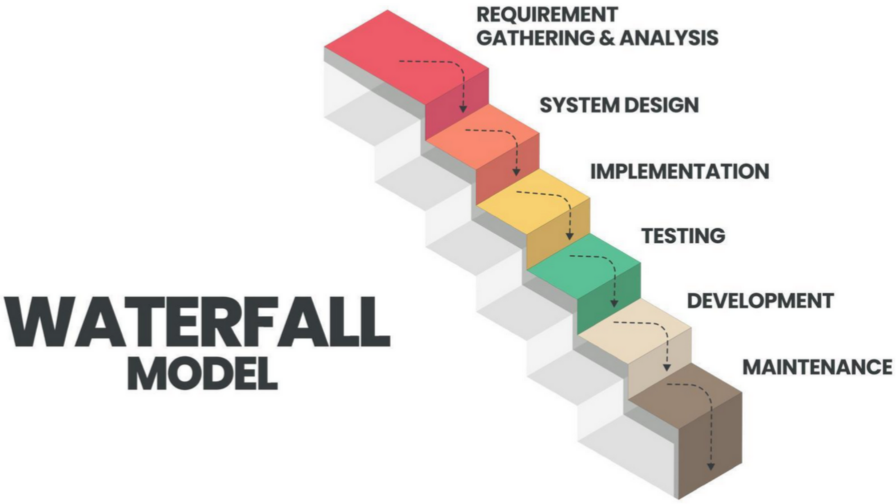


Figure 1.2: Waterfall Model Phases
Source:(ist, 2022)

Software Requirement Specification SRS was created with a clear description of what the system would accomplish.Kung (2013)

2. **System design**

   This phase expects to transform SRS requirements into an appropriate module design for further coding in a programming language. Moreover, The design does adapt as we improve our knowledge of our goals. Also, It provides a high-level and comprehensive design in addition to defining the overall program architecture. A Software Design Document provides documentation of all of this effort SDD. More importantly, a change in the design stage is easier to accomplish (if necessary) than a change in the implementation stage.(ifa, 2021)

3. **Implementation**

   Design is implemented during this phase. If the SDD is finished, the coding step goes without a problem since it contains all the data that software developers and clients require.(Kung, 2013)

4. **Testing**

   At this step, the entire system is under test acceptance to determine whether it has implemented all of the requirements effectively and whether all of the features and functions are operating flawlessly and efficiently.(Kung, 2013)

5. **Development**

   The process of developing software begins at this point. After conducting some study, we will develop the code using the appropriate programming language that is given later in the project. The software development will also be tested and examined during this stage by a module testing framework.(Kung, 2013)

6. **Maintenance**

   Maintenance is the work carried out by each user, Once the program has been provided to the client, installed, and operated.(ifa, 2021)

1.9 Plan and Milestones

Table 1.1: Project Plan

Term	Week #	Milestone #	Milestone Description
498	Week 8	1	• Introduction • Literature Review
	Week 8	2	• System Analysis • Dataset
	Week 10	3	• Design
499	Week 8	4	• Implementation
	Week 9	5	• System Integration
	Week 10	6	• Testing • Evaluation

1.10 Conclusion

In this chapter, we described the idea of the project, discussed the problem and solution, then we stated our aim, objectives, and the timeline of the project. Finally, we set up the scope and the target users of Pathway Genie.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Before we proceed with our report, it is crucial to understand the field as it is the core of our project. This chapter provides an overview of the knowledge areas related to pathway genie. The organization of this chapter is as follows. Explain DNA and Genetic Pathways in Section 2.2. Section 2.3, introduces an overview of Bio-informatics. Section 2.4, Explain in general Artificial Intelligence. Data Science, Machine learning, and Data mining are described in detail, and the difference between them is in section 2.5. Section 2.6, describes the database, types of databases, and data storage methods. Section 2.7 Precision Medicine. Section 2.8 summarizes related gene expression tools. And Section 2.9 summarizes related AI researches. Finally, a summary of this chapter is in section 2.10.

2.2 Genetic Pathways

The fundamental units of all living things are cells. There are many billions of cells in the human body. They give the body structure, absorb nutrients from meals, turn those nutrients into energy, and perform specific tasks. Additionally, cells can replicate themselves and contain the body's genetic material. The nucleus acts as the control center of the cell, giving instructions on how to develop, mature, divide, or die. DNA (deoxyribonucleic acid), the cell's genetic material, is also stored there. The genetic substance in humans and nearly every other organism is DNA. The DNA of an individual can be found in almost all of their cells. Adenine (A), guanine (G), cytosine (C), and thymine (T) are the four chemical bases that make up the code that stores the information in DNA. More

than 99 percent of the 3 billion bases that make up human DNA are the same in every person. The information accessible for constructing and maintaining an organism depends on the order, or sequence, of these bases. (Suza and Lee, 2021)

The fundamental functional component of heredity is a gene. DNA is the component of genes. Some genes serve as instructions for the synthesis of proteins. Humans are thought to have between 20,000 and 25,000 genes, according to the Human Genome Project, an international research project that aimed to decode the human genome's sequence and catalogs its genes. Genes are given unique names so that scientists can keep track of them. Genes are also given symbols, which are brief string combinations of letters and sometimes digits that serve as an abbreviation for the gene name because gene names can be quite long. (med, 2021)

DNA transmits its genetic code to an intermediary molecule called RNA, which is frequently transcribed into the amino acid sequences needed to construct proteins. RNA is a nucleic acid like DNA, but there are some key differences between the two. In contrast to DNA's deoxyribose, RNA has a ribose sugar group. Instead of having two strands, RNA molecules only have one. There is no thymidine base in RNA; instead, there is a base called uridine (U). DNA and RNA both contain the additional nucleotides (A, C, and G). Base pairing with A is a characteristic of U. (Suza and Lee, 2021)

Cells are continually receiving chemical signals from both within and outside the body that are activated by things like injury, illness, stress, or simply the presence or absence of food. Cells use genetic pathways to transmit and receive signals in order to respond to and adapt to these signals. In order to do their intended activities, the molecules that make up genetic pathways interact with signals as well as with one another. There are both short and long distance pathways. To repair localized damage, such as a scratch on the knee, certain cells, for instance, send signals to surrounding cells. Other cells produce chemicals, such hormones, that reach distant target cells through the blood. (gen, 2020)

These pathways regulate how a person reacts to their environment. For example, some pathways have a little impact on how the body body processes medications, while others have a significant impact on how a fertilized egg devel-

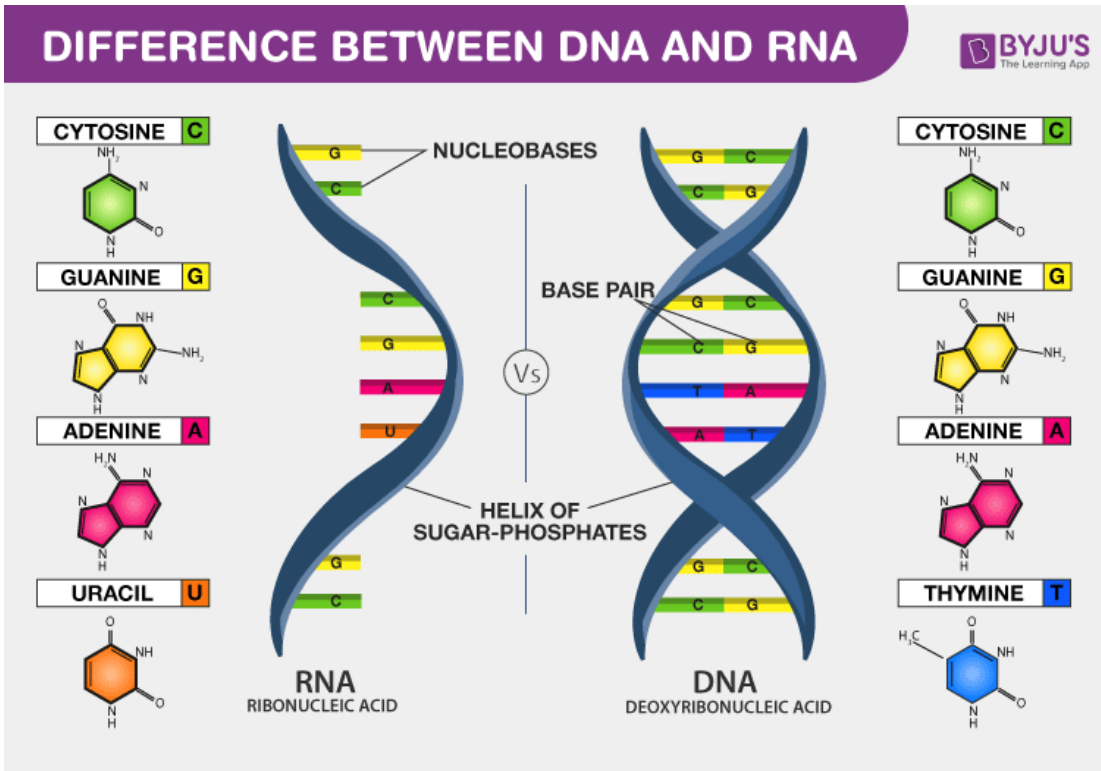


Figure 2.1: Difference between DNA and RNA
Source: byj (2019)

ops into a baby. Other pathways control how and when the eye opens or closes in reaction to light, control how the skin reacts to changes in temperature, and maintain balance while a person is walking. Genetic pathways sometimes fail. Whenever something goes wrong in a pathway, an illness like diabetes or cancer may occur. By examining biological pathways, researchers can discover a lot about human disease. Knowing which genes, proteins, and other components are a part of a pathway can help scientists understand what goes wrong when a disease develops. To identify the causes of the disease, researchers compare certain biochemical pathways in a healthy person to the same pathways in a person with a disease. (gen, 2020)

2.3 Bio-informatics

Bio-informatics collects data from multiple experiments in one location, allowing you to ask questions and find answers, We can process and understand massive amounts of data with the help of bio-informatics. The processing, storage, and analysis of biological data are the focal points of bio-informatics. Coming to ask

ourselves, in bio-informatics, what kinds of biological data can be used? There are three biological data. First, Transcriptomics is the study of Transcriptome, which is a cell's entire set of RNA transcripts. Second, Proteomics is the study of the entire protein set within a cell or system. However, The study of phenotype at the genome level is known as phenomics. A phenotype is how scientists describe something in a person that can be measured. Diabetes risk or eye color are examples of phenotypes. In addition, Bio-informatics allows us to search for alleged connections between our DNA and our phenotypes. Third, Chemo-informatics: Computer-assisted chemical and biochemical data analysis.

A large amount of experimental data is generated during pharmaceutical research. Moreover, Large drug information databases aid scientists in the development of new drugs by providing examples of chemicals that target specific proteins. There are two general approaches in order bio-informatics can help with genomics research with any of the data types listed above. The first approach is to begin with the genome, Scientists can use tools to start with a gene and see which proteins it is the blueprint for. From there, they investigate where the proteins are located in the body and what diseases are associated with them. The second approach is for Health researchers are conducting large-scale studies on volunteers who agree to share their phenotypic measurements and genetic samples. Furthermore, Researchers can use this population data to confirm whether a phenotype is associated with a disease or to identify genes that may influence the phenotype.

Bio-informatics faces Challenges and some of these challenges are Prediction of protein structure Analysis of genomics sequences and gene-finding Analysis of genomics sequences and gene-finding Comparisons of entire genomes Precise, predictive transcription initiation and termination model: the ability to predict where and when transcription will occur in a genome Precise, predictive RNA splicing/alternative splicing model: predicts the splicing pattern of any primary transcript.

Lastly, bio-informatics has Goals to accomplish and some of which are the Development and implementation of computer programs that allow efficient access, management, and use of various types of information. Development and

implementation of computer programs that allow efficient access, management, and use of various types of information. Development of new algorithms (formulas) and statistical measures to analyze connections between members of large data sets. (you, 2022)

2.4 Artificial Intelligence

Artificial intelligence AI is the imitation of human intelligence in technologies that have been designed to behave and think like humans. Any machine that demonstrates characteristics of a human mind, such as learning and problem-solving, may also be considered to be a "machine mind."

The ability of artificial intelligence to think and develop action plans that have the best probability of achieving a particular goal would be its ideal characteristic. Machine learning ML, a subset of artificial intelligence, is the idea that computer programs can automatically identify from as well as adapt to new data without human assistance. By consuming large amounts of unstructured data, including text, photos, and video, deep learning algorithms make it possible for this automatic learning to take place.

The terms artificial intelligence and data science are most frequently used interchangeably. Data science might influence some aspects of AI, but it does not cover all of it. The most renowned field in the world nowadays is data science.

2.5 Data Science

The term "data scientist" was first coined in 2008 when corporations realized the necessity for professionals who are skilled in organizing and analyzing massive amounts of data. "In 2009 McKinsey & Company article, Hal Variant, Google's chief economist and UC Berkeley professor of information sciences, business, and economics predicted the importance of adapting to technology's influence and reconfiguration of different industries".(isc, 2022)

Data Science includes sets of principles, algorithms, problem definitions, and processes for extracting beneficial and non-obvious patterns from large data sets. Successful Data scientists need to possess a level of flexibility and awareness to optimize benefits at each stage of the process and comprehend the complete

spectrum of the data science life cycle in order to uncover meaningful intelligence and improve decision-making through the analysis of data for their organizations. In addition, it has been developed in numerous fields such as data mining and machine learning. Moreover, the terms Data Science, Data mining, and Machine learning are used interchangeably.(Kelleher and Tierney, 2018)

The Data science life cycle:

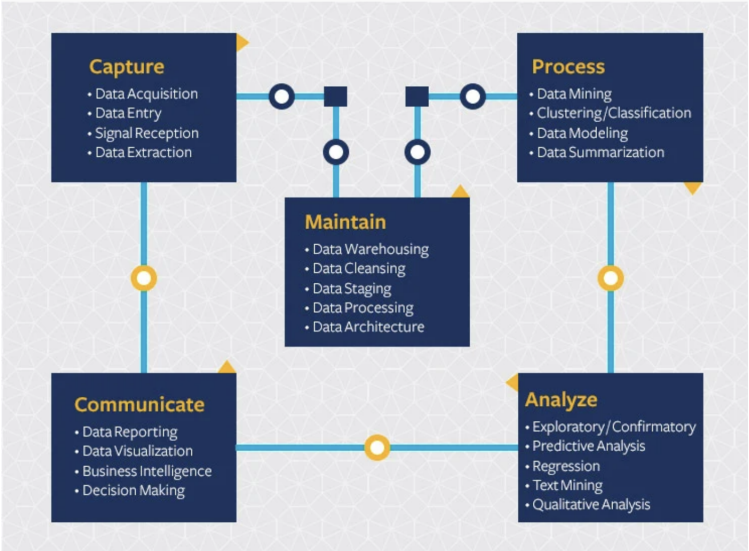


Figure 2.2: Data science life cycle
Source:(isc, 2022)

2.5.1 Machine Learning

"Machine learning is a subset of AI. It was defined in the 1950s by AI pioneer Arthur Samuel as (the field of study that gives computers the ability to learn without explicitly being programmed.)".

It is a sub-field of artificial intelligence, which is widely defined as a machine's ability to mimic intelligent human behavior. In addition, Artificial intelligence (AI) systems are used to carry out complicated tasks in a manner similar to how humans solve issues. Machine learning provides designed statistical algorithms to explore the data for pattern extraction. So, it starts with any kind of data such as numbers, pictures, bank transactions...etc, after that, A machine learning model will be trained on these data, which has been collected and prepared for use as training data. The performance of programs depends on data, the more data, the better performance.(Brown, 2021)

1. Supervised Machine Learning

   It is a past Labeled data set utilized to train models, enabling them to learn and become more accurate over time.(Brown, 2021)

2. Unsupervised Machine Learning

   An unlabeled data program that uses clustering techniques to group data based on their similarities. (Brown, 2021)

3. Semi-supervised(Reinforcement) Machine Learning

   A combination of supervised and unsupervised machine learning. Thus, some parts have supervised data while the other part is unsupervised.(Brown, 2021)

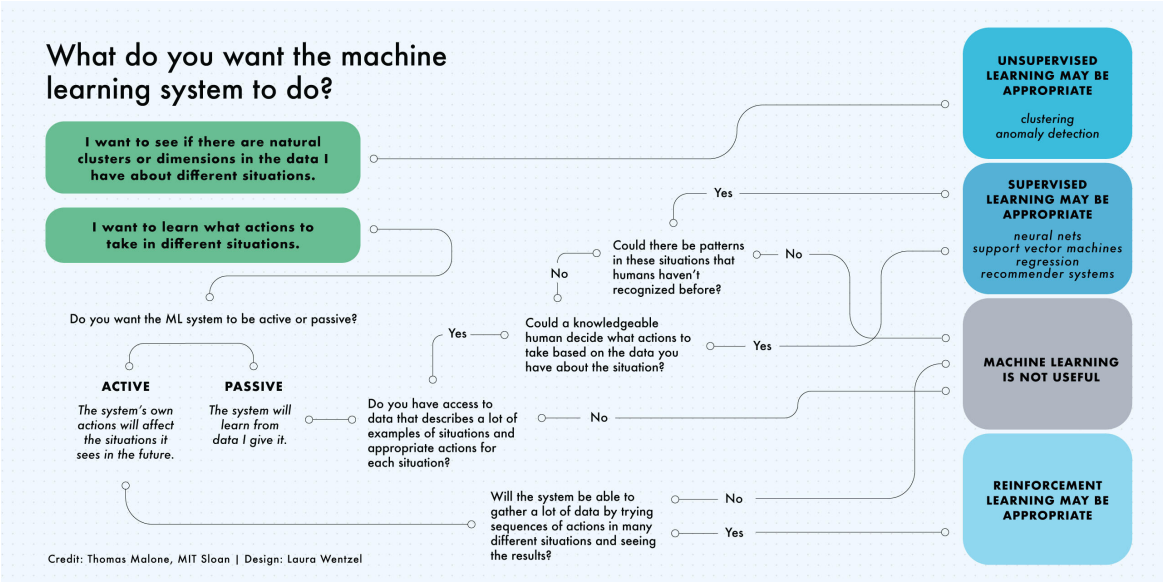


Figure 2.3: Machine Learning Infographic
Source: Brown (2021)

There are approaches to machine learning which are divided into three categories:

2.5.2 Data Mining

Standard methods for analyzing data primarily focus on the extraction of quantitative and statistical data features. Although these methods have been widely used to solve many practical problems such as regression analysis, multidimensional analysis, stochastic models. There is a usage limitation in the Standard

methods. Thus, researchers are utilizing concepts and techniques from machine learning in order to meet the expanding need for new data analysis tools that are resolving and overcome the limitation. The research has led to a new field which is Data Mining.(Sumathi and Sivanandam, 2006)

Data Mining is " a multidisciplinary field drawing works from Statistics, Database technology, Artificial Intelligent, Pattern Recognition, Machine Learning, Information Theory, Knowledge Acquisition, Information Retrieval, High-performance Computing, and Data Visualization ".(Sumathi and Sivanandam, 2006)

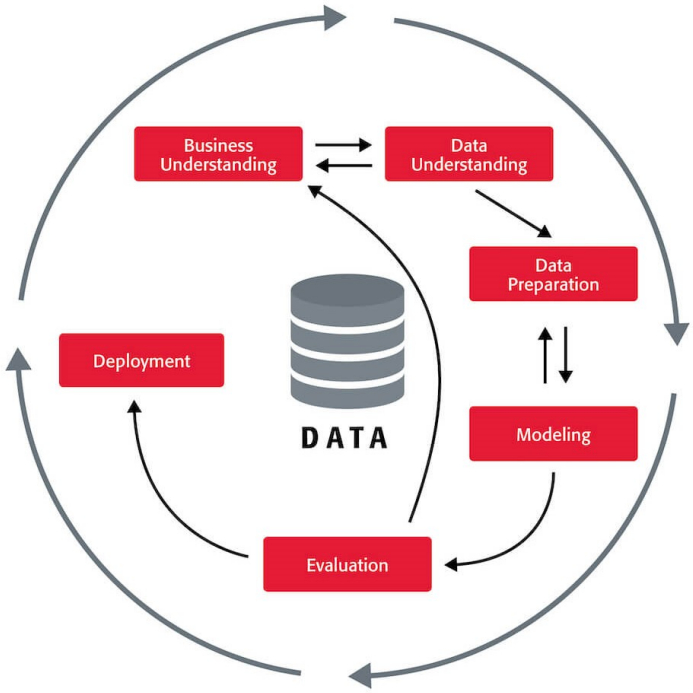


Figure 2.4: Data Mining Phases
Source: boo (2022)

2.5.3 Phases of Data Mining

It is frequently essential to repeat processes in order to adjust for changing data or to introduce new variables.

1. Business Understanding if the proper data to mine is not clearly understood, projects can lead to errors, results that do not answer the precise questions, or inaccurate results. Therefore, asking questions involve around the prob-

lems, aims, and overall objectives is essential to have a clear understanding of the project.(boo, 2022)

2. Data Understanding After the main goal has been established, accurate data must be gathered. The data, which typically originates from a range of sources like sales records, customer surveys, and Geo-location data, must be related to the topic matter. The objective of this phase is to verify the information accurately includes all data sets required to address the objective.(boo, 2022)

3. Data Preparation The preparation phase, which consumes the most time, comprises three steps: extraction, transformation, and loading (also known as ETL ETL). Data is first extracted from various sources and placed in a staging area. The data is then cleaned, null sets are populated, eliminate duplicate data, errors are fixed, and all data is allocated to tables during the transformation step. Finally, at the loading step, the formatted data is loaded into the database for use.(boo, 2022)

4. Modeling The relevant data set is addressed by data modeling, which also takes into account the most appropriate statistical and mathematical strategy (s). There are numerous modeling techniques available, which include classification, clustering, and regression analysis. It's also not infrequent applying different models to the same data to achieve different goals.(boo, 2022)

5. Evaluation It's time to examine the models' effectiveness in responding to the question posed during the business understanding phase after they have been built and tested. This is a human-driven phase, as the project manager must determine whether the model output adequately meets their objectives. If not, a new model can be created or new data can be prepared.(boo, 2022)

6. Deployment Finally, utilizing the data mining model once it has been determined to be accurate and successful in resolving the objective question. A report providing insights or a visual presentation are both examples of

deployment. It can also result in action, like developing a new sales strategy or implementing risk-reduction measures into place.(boo, 2022)

2.5.4 Differences Between Data Mining and Machine Learning

Machine learning and data mining are distinct techniques that are frequently used interchangeably. Although they both work well for pattern recognition in big datasets, they function quite differently.

Data Mining is a procedure to recognize patterns in data. The benefit of data mining is actively recognizing non-intuitive data patterns using algorithms, it assists in providing answers to questions we didn't even know to ask "(e.g., consumers who buy peanut butter are more likely to buy paper towels)". However, these insights still need to be understood by humans in order to be applied to business choices.

Meanwhile, Machine Learning refers to the process of teaching a computer to learn similarly to how humans do. Through data analysis and machine learning, computers can learn to estimate probabilities and make predictions. Additionally, while data mining is occasionally a part of the machine learning process, it is not always necessary for humans' continued involvement (e.g., a self-driving car determines where to stop, accelerate, and turn by relying on data mining).(boo, 2022)

2.6 Database

A database is a collection of data organized for simple management, updating, and access. Data records or files containing information, including sales transactions, customer information, financial information, and product information, are often aggregated and stored in computer databases. Any type of data may be stored, maintained, and accessed using databases. They gather data on individuals, locations, or objects. It is gathered in one location so that it may be seen and analyzed. You might think of databases as a well-organized collection of data. (ora, 2022)

The types of databases are many. In computing, databases are frequently categorized according to their organizational approach. Relational databases,

Object-oriented databases, Distributed databases, and Graph databases are a few of the most common organizational databases. Distributed databases will be used in our tool Pathway Genie. pho (2021a)

2.6.1 Distributed Database

A distributed database is one that consists of two or more files spread over many places, whether they are connected by the same network or not. The database is split up into several physical places for storage and processing, and there are many database nodes involved. Distributed databases are logically connected to one another when they are part of a collection, and they frequently form a single logical database. Data is physically stored across several sites and is separately handled in distributed databases.

Utilizing distributed databases has a lot of benefits. Since distributed databases provide modular development, systems may be expanded by putting new computers and local data in a new location and easily connecting them to the distributed system. In centralized databases, failures result in a total shutdown of the system. Distributed database systems, however, continue to operate with lower performance when a component fails until the issue is resolved. If the data is near where it is most often utilized, administrators can reduce transmission costs for distributed database systems. Centralized systems are unable to handle this. (pho, 2021b)

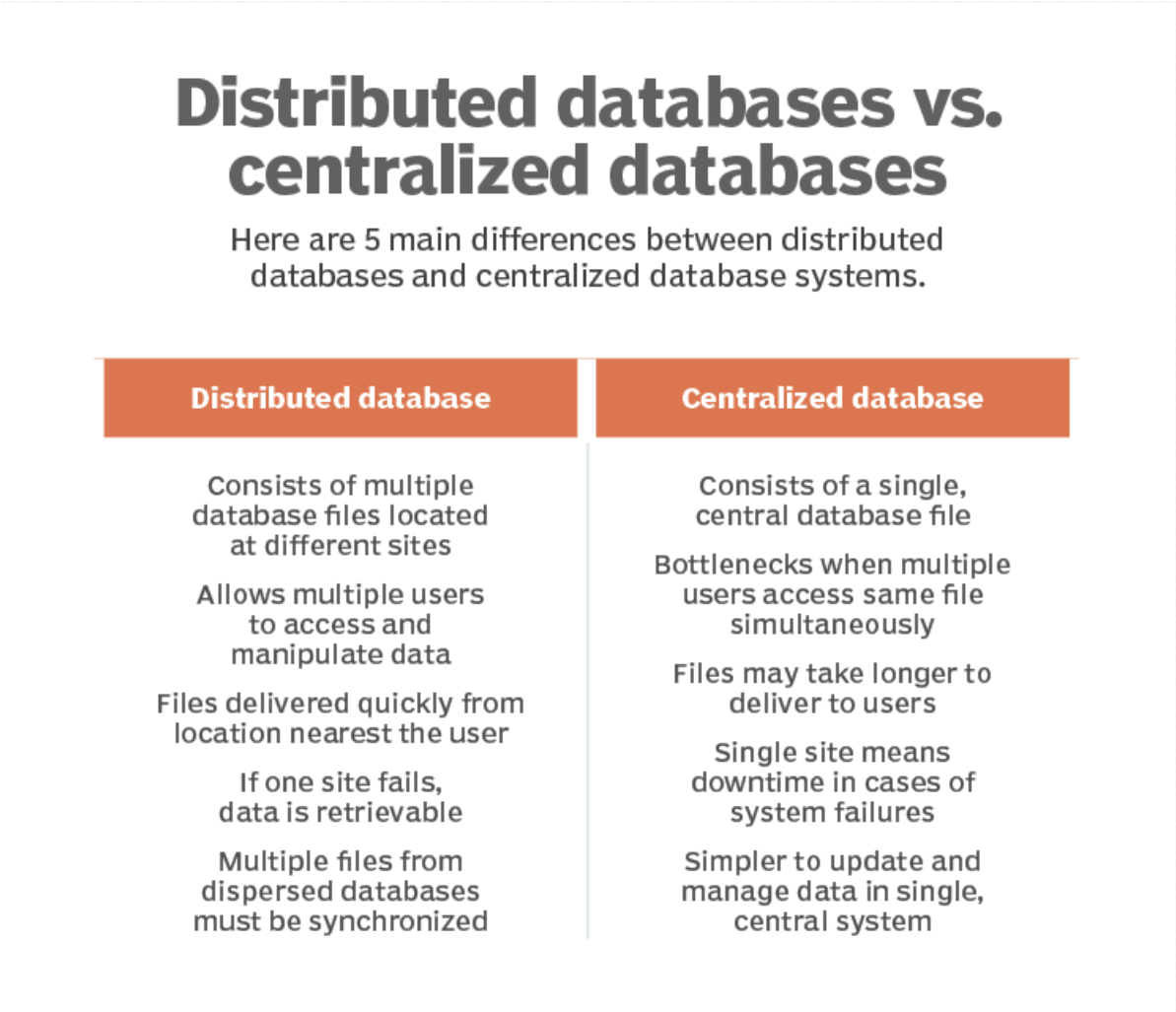


Figure 2.5: 5 ways centralized databases differ from distributed databases

Source: pho (2021b)

There are data storage methods for distributed databases, such as data transformation, data preparation, data replication, data fragmentation, and data virtualization. Systems that use database replication keep copies of the data at many locations. A database is fully redundant if every single record is accessible from multiple locations. The relations are fragmented, which means they are divided into smaller parts when it comes to distributed database storage. Each fragment is kept in a separate location depending on where it is needed. While data virtualization is an approach to data management that allows an application to retrieve and manipulate data without requiring technical details about the data, such as how the data is formatted or where it is physically located. In our

tool, we will use the data virtualization method to store distributed databases. (fau, 2020)

2.6.2 Data Virtualization

Data virtualization is a methodology of organizing data access without the need to obtain information about data structure or placement in a specific information system. The main goal of data virtualization is to simplify the access and use of data by converting it into a service (Data as a Service), which essentially shifts the paradigm from storage to usage. Users are able to access all data through a single view due to data virtualization. Data virtualization employs pointers to those blocks rather than transferring large data blocks, which have a lower storage footprint and offers high-performance access to stored data. (Joshi, 2021)

2.7 Precision Medicine

Precision medicine commonly referred to as "personalized medicine," is a cutting-edge strategy for tailoring illness prevention and treatment that considers variations in people's genes, environments, and lifestyles. One size fits are the underlying principle of modern healthcare. Unfortunately, the effects of a defined treatment plan are not always consistent. For example, a medication that reduces the tumor in one lung cancer patient might not do the same for another lung cancer patient. (fda, 2018)

The idea of precision medicine is beneficial in this situation. Although the goal of precision medicine is to personalize all treatments, it is especially well adapted to treating diseases like cancer and genetic disorders, which have so far defied effective treatment due to genetic variances between individuals. Precision medicine, which holds great potential, is rapidly evolving and becoming more extensively used. (sof, 2022)

The future and present of critical care depend greatly on precision medicine. But various obstacles prevent it from being used for all ICU ICU patients. These include a lack of agreement on nomenclature, difficulty classifying patients into distinct groups, the difficulties of multiple health conditions, scale, and the requirement for fast interventions. (spr, 2017)

Genomic research has greatly benefited from technology, which we can no longer imagine science without. The knowledge produced by the HGP HGP has altered genomics and is now being used in health care settings. With The new precision (personalized) medicine, giving customized diagnosis and therapy depends on genome sequencing and data analysis. The understanding of environment-biology interactions, underlying illness mechanisms, and evidence-based therapies are also essential for this new approach to medicine. With the use of all this data, it should be possible to trace a person's progression from health to disease. (nat, 2017)

Precision medicine has several approaches that belong to genetics and health care which include: (aca, 2015)

- Human genome projects
- Sequencing technologies

2.8 Related Gene Expression Tools

This section summarizes all the gene expression tools. Table 2.1 reflects the most common criteria which are gene expression Tool, description, pros, and cons.

2.8.1 GeoQuery Tool

Gene Expression Omnibus GEO is a public archive for a wide variety of high-throughput experimental data. There are four fundamental entities in GEO's most fundamental organizational structure. The first three (Sample, Platform, and Series) are contributed by users, while the fourth (the dataset) is assembled and processed by GEO personnel from user-submitted data. GEOquery defines a class representation for each of these entity categories.

Once the data from GEO has been parsed into the GEOquery data structures, methods for accessing the different pieces of the GEO entity are available from GEOquery, allowing access to the data simple and quick. However, it take a while, approximately 5 days to proceeds data back and forth.

GEOquery's major objective is to download and parse SOFT format files from GEO while preserving all of the data found in the GEO records.GEOquery's layout makes it simple to obtain data from GEO. Hence, GetGEO is the only

command that is required. The custom data structures and associated methods that arrange data from a GEO download are two of GEOquery's most advanced functionality.

GEOquery data structures are classified into two major categories that are equivalent to the structure used by GEO to convey data. Moreover, the associated techniques affect each in a similar way.

GDS2MA and GDS2eSet are two functions that transform the GDS into one of the two data structures. For example:

$> eset \leftarrow GDS2eSet(gds, do.log2\frac{1}{4}TRUE)$

File stored at: /tmp/RtmpitNEHc/GPL5.soft

By applying this function we convert the GDS objective to an expression set. (Davis and Meltzer, 2007)

2.8.2 LIMMA Tool

LIMMA stands for "linear models for microarray data", such models include linear regression and analysis of variance. LIMMA is a software program for R/Bioconductor, offers an integrated method for analyzing data from gene expression experiments. It has a variety of tools for dealing with complicated experimental designs and information borrowing to resolve the issue of small sample numbers. Recent developments have greatly increased limma's capabilities in two important areas.

First, the package can now analyze RNA sequencing (RNAseq RNAseq) data for differential expression and differential splicing. Second, the package could now analyze expression profiles in terms of co-regulated gene sets or higher-order expression signatures, moving beyond the typical gene-wise expression analyses in a variety of ways. Hence, it significantly improves opportunities for biological understanding of variations in gene expression. (Ritchie et al., 2015)

2.8.3 BART Tool

BART (Bioinformatics Array R Tool) is a web application that analyzes microarray data. Simplifies the process of downloading and analyzing data from various microarray platforms. Can use an expression table the user has uploaded

or a GSE accession ID from GEO. In addition, BART produces hierarchically clustered heat maps and PCA PCA plots to simplify sample comparison, while also creating pathway enrichment and volcano plots analysis to conclude differential expression testing. Also, BART enables researchers to derive information from their own microarray data or microarray studies made available by GEO. In comparison to other microarray analysis tools, BART is functional on more microarray experiments and offers more thorough analyses. (Maria Luisa Amaral, 2018)

2.8.4 SAGE Tool

SAGEis a shortcut for Serial Analysis of Gene Expression, is an experimental method established to provide a precise measurement of gene expression. The tool a very capable technology that not only provides a global gene expression profile of a certain type of cell or tissue, but also enables us to compare the profiles created for a pair of cells maintained under various conditions to find a collection of specific genes that are related to those conditions. Moreover, The SAGE approach relies on isolating unique sequence tags from individual mRNAs and serially concatenating those tags into lengthy DNA molecules for lump-sum sequencing. Studies examining almost any type of biological phenomenon for which changes in cellular transcription are involved, may use the SAGE approach. (Yamamoto et al., 2001)

Table 2.1: Related Gene Expression Tools

Gene Expression Tool	Pros	Cons	Support AI?
GeoQuery Tool	- Only one command is needed which is getGeo. - Free, simple, and ground-breaking tool.	Takes a long time to receive information back	No
LIMMA Tool	- Useful for continuous data such as microarray data. - Resolve the issue of small numbers.	-	Yes
BART Tool	- Simplifies downloading and analyzing data from many microarray platforms. - Use an expression table downloaded by the user.	Time-consuming to score new data in the large data sets	No
SAGE Tool	- provides a global gene expression profile. - Enables comparison of the profiles created for a pair of cells.	- The length of gene tag is extremely short which makes analysis difficult. - sequencing error and valid tag that does not, in reality, exist.	No

2.9 Related AI Researches

This section summarizes some of the research related to AI and Gene expression. Table 2.2 reflects the most common criteria which are methods, preprocessing, dataset, and the results of the papers in this section.

2.9.1 AI Technique for Gene Expression by Tumor RNA-Seq Data

This paper described a method for classifying several types of cancer, including breast invasive carcinoma BRCA, kidney renal clear well carcinoma KIRC, E: at Adenocarcinoma (LUAD), Lung Squamous Cell Carcinoma LUSC and Uterine Corpus Endometrial Carcinoma UCEC. The proposed approach consisted of three phases. Pre-processing is the first phase, followed by data augmentation, which expanded the size of the original dataset by five times. The deep convolutional neural network architecture is used in the third phase to categorize the 5 main cancer types. According to the results that were shown and the performance metrics used in this research, the suggested approach had a total testing accuracy of 96.90% for 5 classes of tumors. (Khalifa et al., 2020)

2.9.2 Enhanced Deep Learning Model for Personalized Cancer Treatment

This paper introduces both a data federation method and a model that predicts the response of 265 anti-cancer drugs. The introduced Enhanced Deep-DR enhances the model generalization by increasing the number of mutated genes that may be responsible for the response to different drugs. The fundamental goal is to generalize the predictor so it will be able to predict the response to different drugs accurately. The results show that data federation enhanced the accuracy and decreased the Mean Square Error by almost 25%. Convolutional Neural Network (CNN) based-model is also implemented; it achieves higher accuracy in predicting the drug response than in some other DL-based models but less than the Enhanced Deep learning. The Enhanced Deep DR achieves better accuracy within the range of 5% to 12% than other DL models. (Ahmed et al., 2022)

2.9.3 Deep Learning and Machine Learning via a Genetic Algorithm to Classify Breast Cancer DNA Data

This study employs algorithm analysis based on Artificial Intelligence AI to categorize breast cancer (DNA). The core idea is to focus on the use of the machine and deep learning techniques. Furthermore, a genetic algorithm is used to diagnose gene expression to reduce the number of unclassified cancers. After the entry of patients' genetic data, processing processes that call for filling in the missing values using various methods are utilized. The genetic algorithm is used to combine each approach, and the results are compared for accuracy to determine which produces the best data for the classification process. (Hussein and Al-Sarray, 2022)

2.9.4 Alzheimer's Patient Analysis Using Image and Gene Expression Data and Explainable-AI to Present Associated Genes

In this paper, gene expression and image data are used to identify and detect AD using machine learning techniques. To categorize AD from MRI images, SpinalNet and convolutional neural networks (CNN) were first applied. Then, we classified the diseases using the k-nearest neighbors KNN, support vector classifier SVC, and Xboost classifiers using micro-array gene expression data. We developed an explainable artificial intelligence (XAI) technique to produce reliable predictive modeling. Local interpretable model-agnostic explanations LIME, for a straightforward human interpretation, is the XAI approach we applied in this case. LIME explains how genes were predicted and identify the genes that are especially at fault in an AD patient. With a 97.6% accuracy rate, CNN is 10.96% more accurate than the SpinlNet method. SVC provides more accuracy when analyzing gene expression data than other methods. LIME shows how the most important genes were chosen for a specific AD patient based on the patient's gene expression data. (Kamal et al., 2021)

Table 2.2: Related AI Researches

Reference	Methods	Preprocessing	Datasets	Results
Khalifa et al. (2020)	Deep learning, KNN method	data augmentation	matrix contains 2,086 rows and 972 columns	overall testing accuracy of 96.90%
Ahmed et al. (2022)	data federation, DL-based model	data integration	total mutated genes are 19, 702 genes	accuracy within the range of 5% to 12% better than other DL-models
Hussein and Al-Sarray (2022)	machine learning, PNN, KNN, deep learning, and Long Short-Term Memory	Cleaning of the dataset and using the value techniques to fill in missing data	series of 295 consecutive patients	The accuracy ratio was arranged as follows: GA-KNN, GA-CNN, GA-PNN, GA-LSTM 0.92881, 0.97627, 0.94915 and 0.8339.
Kamal et al. (2021)	SpinalNet, CNN, KNN, support vector classifier (SVC), and Xboost	data augmentation	gene expression and image data	CNN with a 97.6% accuracy rate

2.10 Conclusion

This chapter provides an overview of the five main knowledge areas related to the pathway genie project, Bioinformatics, AI, Data Science, databases, and Precision Medicine. It discussed what will be applied to our tool PathwayGenie based on the literature review. Also, the related gene expression tool and AI research. The next chapter will discuss the data collection.

CHAPTER 3

DATA COLLECTION AND REQUIREMENTS

3.1 Introduction

This chapter discusses data collection including questionnaires and interviews, the project's dataset, as well as, the functional and non-functional requirements of Pathway genie.

3.2 Dataset

GEO is an international public repository that archives and freely distributes research community-submitted microarray, next-generation sequencing, and other forms of high-throughput functional genomics data.

GEO's three main objectives are: Provide a robust, versatile database for storing high-throughput functional genomic data, Provide simple submission procedures and formats to support complete and well-annotated data deposits from the research community and, Provide simple mechanisms for users to search for, locate, review, and download studies, gene expression profiles of interest.

The following is how GEO records are organized: A Platform record consists of a summary description of the array or sequencer, as well as a data table defining the array template for array-based Platforms. A Platform may refer to a large number of Samples submitted by multiple submitters. A Sample record describes the conditions under which a single Sample was handled, the manipulations that were performed on it, and the abundance measurement of each element derived from it. A Sample entity may be included in multiple Series and must reference only one Platform. A Series record connects a group of related Samples and serves as the focal point and description of the entire study. Tables describing

extracted data, summary conclusions, or analyses may also be included in series records.

Each of the three is given a distinct and stable GEO accession number. Selected primary records are rendered at a higher level into DataSet and gene Profile records: A GEO Series record, as defined above, is an original submitter-supplied record that summarizes an experiment. A DataSet is a curated collection of biologically and statistically comparable GEO Samples that serves as the foundation for GEO's data display and analysis tools.Samples within a DataSet all refer to the same Platform, which means they have the same set of array elements. Value measurements for each Sample within a DataSet are assumed to be calculated in the same way, i.e., background processing and normalization are consistent across the DataSet. DataSet subsets provide information reflecting experimental factors.

Although both Series and DataSets can be searched using the GEO DataSets interface, only DataSets serve as the foundation for GEO's advanced data display and analysis tools, such as gene expression profile charts and DataSet clusters. Because not all data submitted is suitable for DataSet assembly and we have a backlog in DataSet creation, not all Series have corresponding DataSet records.

DataSets are used to generate profiles. A Profile is made up of the expression measurements for each gene across all Samples in a DataSet. The GEO Profiles interface can be used to search for profiles. Original submitter-supplied records (Series, Samples, and Platforms) as well as curated DataSets are stored in the GEO DataSets database. Curated DataSets serve as the foundation for GEO's advanced data visualization and analysis features, such as tools for detecting differences in gene expression levels and cluster heatmaps. GEO DataSets are used to create GEO Profiles. Not all of the original submitter-supplied records have yet been compiled into curated DataSets.

The GEO DataSets database can be searched using a variety of criteria such as keywords, organism, DataSet type, and authors. The Querying GEO DataSets and GEO Profiles page contains examples and complete instructions for searching for GEO DataSets of interest.

Our project focuses on breast cancer, it's a disease that when the cells in

the breast mutate uncontrollably. There are various types of breast cancer. The type of breast cancer depends on which cells in the breast develop into cancer.

Breast cancer begins in any part of the breast. A breast is composed of lobules, ducts, and connective tissue. The lobules are the milk-producing glands. Ducts transport milk to the nipple. The connective tissue holds everything together. The majority of breast cancers begin in the ducts or lobules. Breast cancer can mutate outside of the breast via blood and lymph vessels. Breast cancer spreads to other parts of the body. (ncb, 2022) cdc (2022)

3.3 Data Collection

The process of collecting the information required to determine the system requirements is known as data collection. Although data collection methods differ, all highlight the collection of precise and truthful data. Data collection improves product quality, aids in customer understanding, and facilitates decision-making. This project employs two data collection methods: an interview and a questionnaire.

3.3.1 Interviews

An interview is a data collection method that uses verbal responses to gather detailed information. They are especially useful for eliciting the story behind a participant's perception. The Research team can use a series of questions to learn more about a topic, or they can dig deeper into responses or findings.

We sent the interview via email to a professional in the field. They have been in the field for 3 years specializing in AI and Bio-informatics, their current research is on AI and health informatics, machine/deep learning modeling, drug repositioning, diagnostic, and information integration. Appendix 1 contains the interview questions.

From the interview results, we were able to conclude features and improvements to add and develop in our tool based on the professional's responses. including a full report and an updated pathway in the system.

3.3.2 Questionnaire

A questionnaire is a data collection technique that structures a set of questions that participants must answer. Pathway Genie's questionnaire includes six questions relating to gene expression analysis tools used by researchers. The questionnaire responses came from professionals with at least 5+ years of experience in the field. Appendix 2 contains the questionnaire questions.

And Due to our product being in a relatively uncommon field, getting insights from scientists specializing in the field on other tools was very helpful.

From the questionnaire analysis, we discovered that approximately 60 percent did not know if the tool they worked with supported AI technology or not, and 20 percent said that no AI technology had been used.

As a result, our goal is to use AI machine learning to improve the tool and speed up the analysis process while giving the scientists a result that's fully ready for their research saving them time and cost.

3.4 Establishing Requirements

At the beginning of the project, requirements must be carefully considered and understood. Functional and non-functional requirements are two different categories. Functional requirements describe what a system should be or be able to perform. Additionally, they are establishing the qualities and characteristics that the system must have in order for users to find value and utility in it. Non-functional requirements outline the system's operation and illustrate behavior. Additionally, they demonstrate the system's quality. Here are the functional and non-functional requirements:

3.4.1 Functional Requirements

FR 1. Genie shall Allow users to create an account.

FR 1.1 Genie shall allow the user to enter their personal information (first and last name).

FR 1.2 Genie shall allow the user to enter their email address.

FR 1.3 Genie shall allow the user to enter and confirm their password.

FR 2. Genie shall allow the user to Log in.

FR 2.1 Genie shall allow the user to enter his/her E-mail.

FR 2.2 Genie shall allow the user to enter his/her password.

FR 3. Genie shall allow the user to submit their data to be analyzed.

FR 4. Genie shall generate a report with pathways of differential mutated genes.

FR 5. Genie shall allow the user to Log out

3.4.2 Non-Functional Requirements

NFR 1. Genie shall be available 24/7.

NFR 2. Genie shall be available in light and dark modes.

NFR 3. Genie shall support the English language

NFR 4. Genie shall provide a user-friendly, clear, and easy-to-use interface.

NFR 5. Genie shall be able to secure patient private data.

NFR 6. Genie shall save histories of previous pathway analysis.

3.5 Conclusion

This chapter mainly focuses on the dataset and collecting data using two methods which are interview and questionnaire. GEO is a database platform that assists in providing datasets. The dataset's chosen category is breast cancer. in addition, the project's functional and non-functional requirements were discussed.

CHAPTER 4

ARCHITECTURE AND DESIGN

4.1 Introduction

Software engineering refers to the design, development, and maintenance of software. It ensures your applications are built consistently, accurately, on time, and within constraints and budget. The ease of use of a software product and the features it offers to its users determine its quality. Applications must meet the following criteria:

1. **Operational**

   This indicates how well the software performs in operations such as budget, ease of use, efficiency, accuracy, functionality, reliability, security, and protection.

2. **Transitional**

   Migration is important when moving an application from one platform to another. So portability, reusability, and adaptability all fall into this category.

3. **Maintenance**

   Demonstrate how well the software performs in changing environments. Modularity, maintainability, flexibility, and scalability go into the maintenance part.

This section discusses the system's diagrams ER, relational database, use case, dataflow, flowchart, sequence, and class diagrams. Finally, the prototype of the system is presented.

4.2 Diagrams

Diagrams are used to represent information and techniques. As we move forward with the project, we created diagrams for viewing the system. Draw.io was used to create all of the diagrams, they key advantages of the website that it was simple to use and allowed us to work together.

4.2.1 Use cases

A use case diagram is a graphical representation of how a user might interact with a system.

The main actor in the system is scientist controlling and managing the interactions with the system which is interacting with the database, patient data and the data file. The database here is in charge of storing and analyzing the data by reading it and saving it's history. The main actor will log in, log out,upload file and cancel and reload the file.

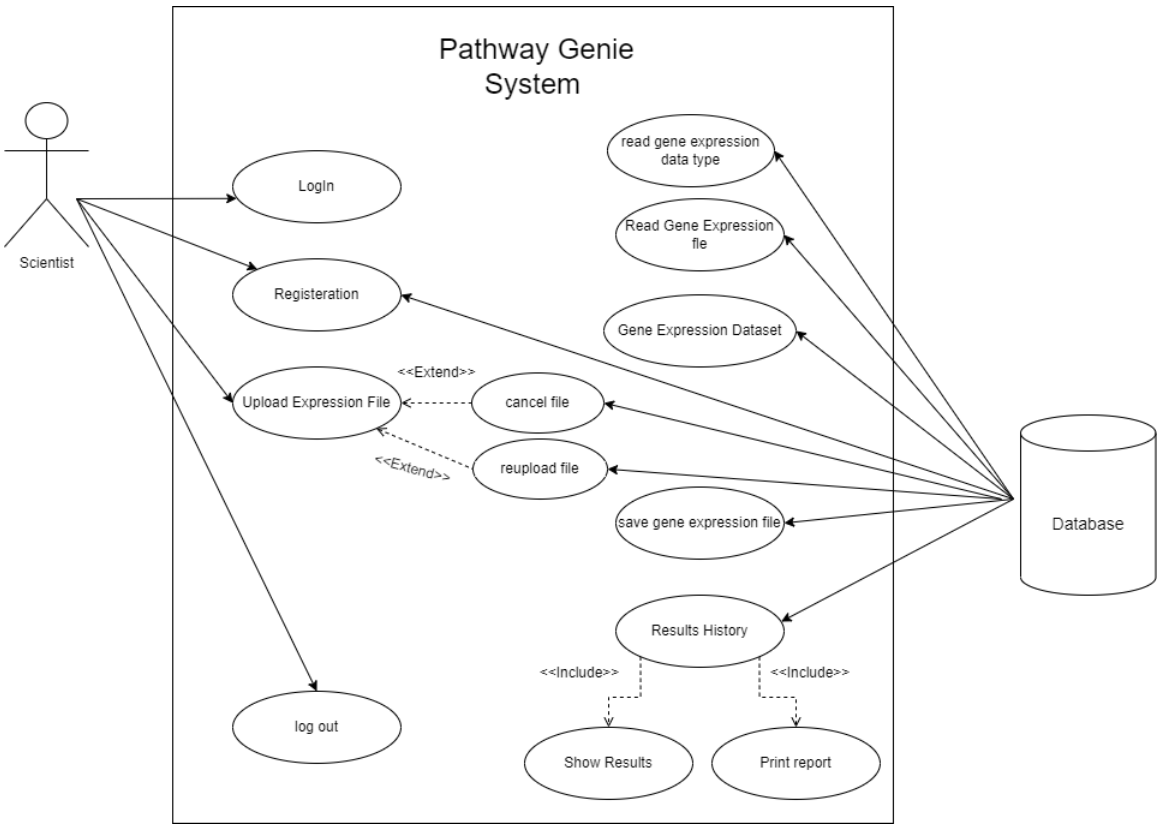


Figure 4.1: Use Case diagram

Table 4.1: Create account table

Create Account
Brief description: The Actor must create an account to use the program.
Actor: Scientists
Precondition: The actor must enter the correct personal information to get into account.
Basic flow of events: • The actor can open the program web application from their device. • The actor creates an account. • The actor submits his necessary correct details to the system. (Name, email). • The account is saved in the system and successfully created.
Extensions: • The system fails to create an account. • The actor filled in the wrong information. • The user has already created an account.
Post condition: The account must be created to use the programs advantage.
Special requirements: To create the account the actor must be connected to the internet.

Table 4.2: Log into account table

Log into Account
Brief description: The Actor must log in to account to use the application.
Actor: Scientists
Precondition: The actor must enter the correct personal information to get into account.
Basic flow of events: • The actor can open the program web application from their device. • The actor log's into the account • The actor enters all their necessary correct details into the system. (Name, email). • The actor gets into account.
Extensions: • The system fails to log into account. • The actor filled in the wrong information. • The actor did not create an account.
Post condition: The actor must log in to use the programs advantage.
Special requirements: To log in to the account the actor must be connected to the internet.

Table 4.3: Submit data file table

Submit data file
Brief description: For the application to read the data the actor must submit file
Actor: Scientists
Precondition: The actor must be logged into the account.
Basic flow of events: • The actor can open the program web application from their device. • The actor log's into the account. • The actor submits the expression file to the system. • The system reads and analysis the file. • The system loads the analyzed result to the actor.
Extensions: • The actor is not logged into account. • The actor uploads wrong file. • The actor uploads a file that is too big for the system.
Post condition: The system reads the file.
Special requirements: For the system to read the submitted file the actor must be connected to the internet.

Table 4.4: Generate final report table

Generate final report
Brief description: The system finalizes with a final report after processing and analyzing the data.
Actor: Scientists
Precondition: The actor must be logged in and submitted a gene expression file to get the final analyzed report.
Basic flow of events: • The actor can open the program web application from their device. • The actor log's into the account. • The actor submits the expression file to the system. • The system reads and analysis the file. • The system prints the final report.
Extensions: • The actor is not logged into account. • The system fails to read the gene expression file. • The system fails to print the final report.
Post condition: The final report is printed for the actor.
Special requirements: To get the final report the actor must be connected to the internet.

4.2.2 Class Diagram

UML is a common modeling language for system design visualization that simplifies the process for software developers to build systems. Class diagram is used to illustrate the static structure of classes in the system to show the interaction between objects. Also, UML class diagram represent the system classes, attributes of objects, operations, and relations between the classes. (Gustriansyah et al., 2019)

The diagram 4.2 shows the relationship between the main classes in the system. such as the scientist class, data file class, gene expression class and final report class.

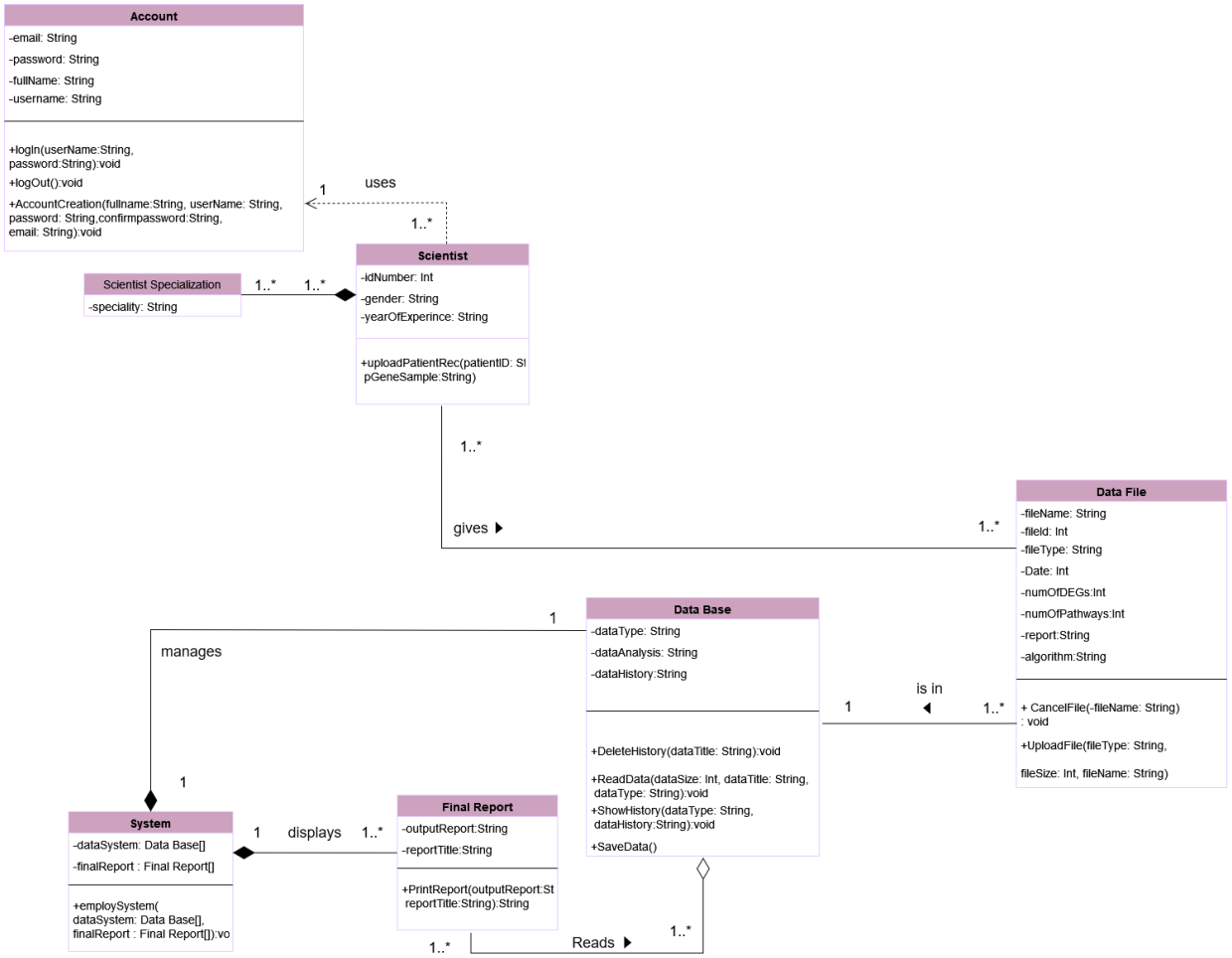


Figure 4.2: UML Class Diagram

4.2.3 Database diagrams

Database diagrams consist of ER ER and schema, these two diagrams are critical to understand the system.

4.2.3.1 Entity-Relationship Diagram

The entity relationship diagram is a well known high-level conceptual data model describes all the entities and attributes in our system and their relationships. Many database design tools use the concepts of this model and its modifications when conceptually designing database applications. (Elmasri and Navathe, 2011)

Diagram 4.3 shows the relation between the scientist and the data file is one-to-many relationship as multiple files can be uploaded by one user (scientist). Another relation between the gene expression and final report is one-to-one relationship as one final report contain result of a one gene expression analysis.

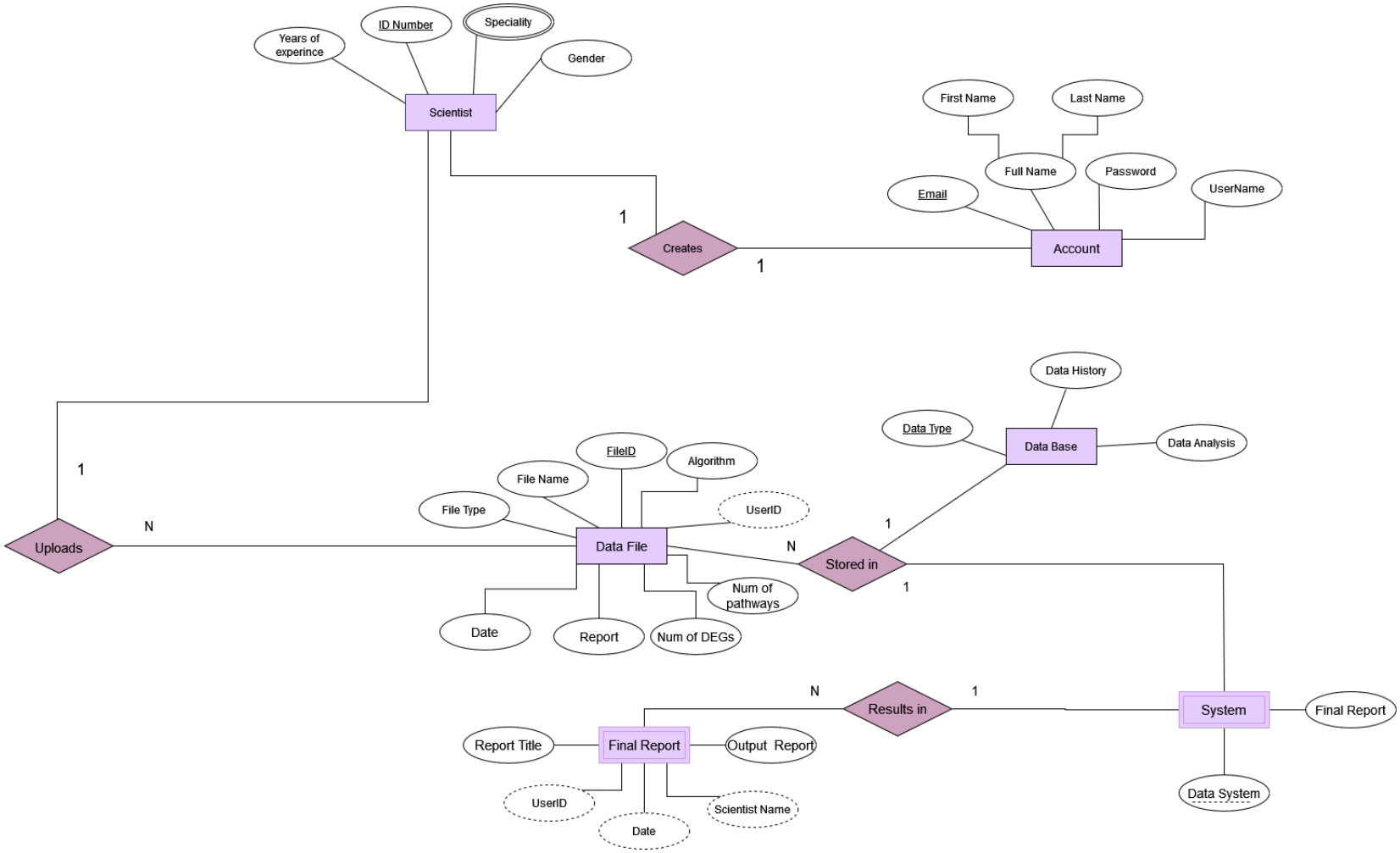


Figure 4.3: ER Class Diagram

4.2.3.2 Database schema Diagram

Diagram 4.4 describes the tables and the corresponding attributes in the database.

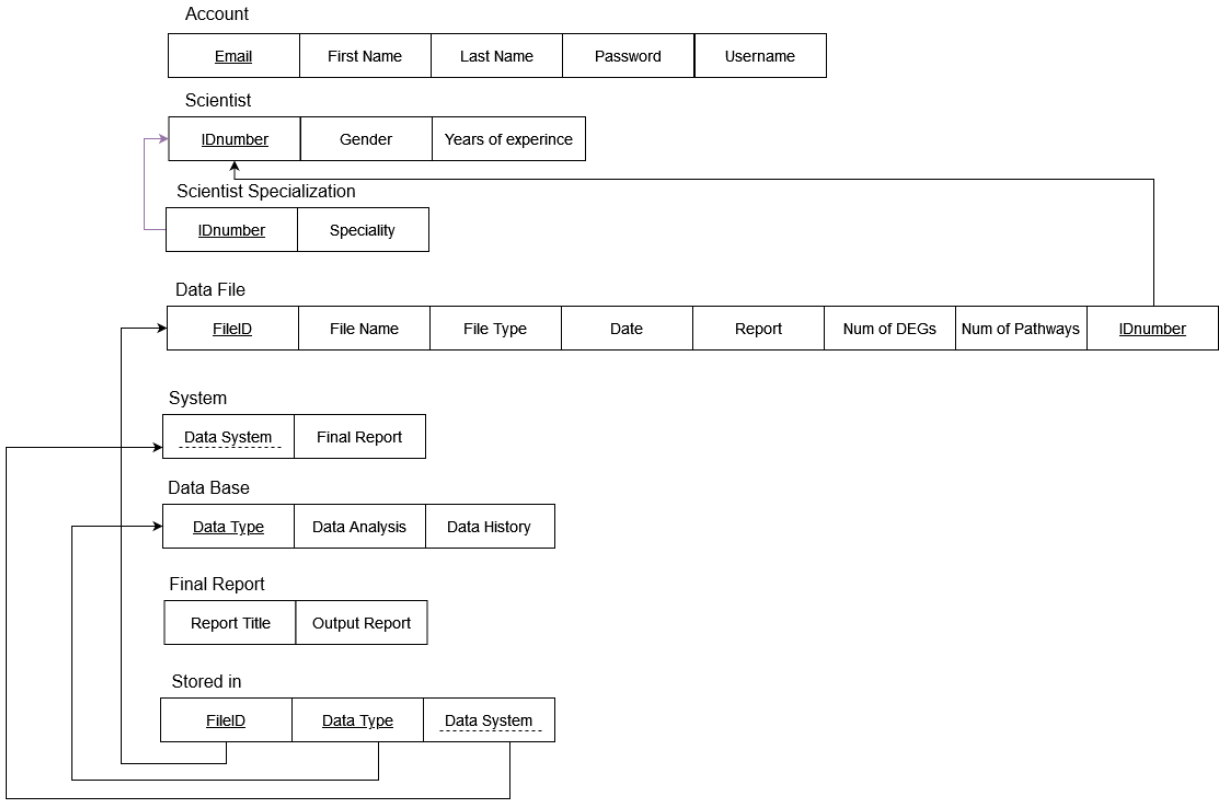


Figure 4.4: Database Schema Diagram

4.2.4 Sequence Diagram

Figure 4.5 shows the process of the user interaction with the system. The scientist will try to login into the account, the system will send either it's valid or invalid. If it's valid will login and if it's not it will ask again for information. Then the uploaded file, database will make sure if the file is invalid or valid. If it's valid will print a final report and if it's not, will ask again for the information.

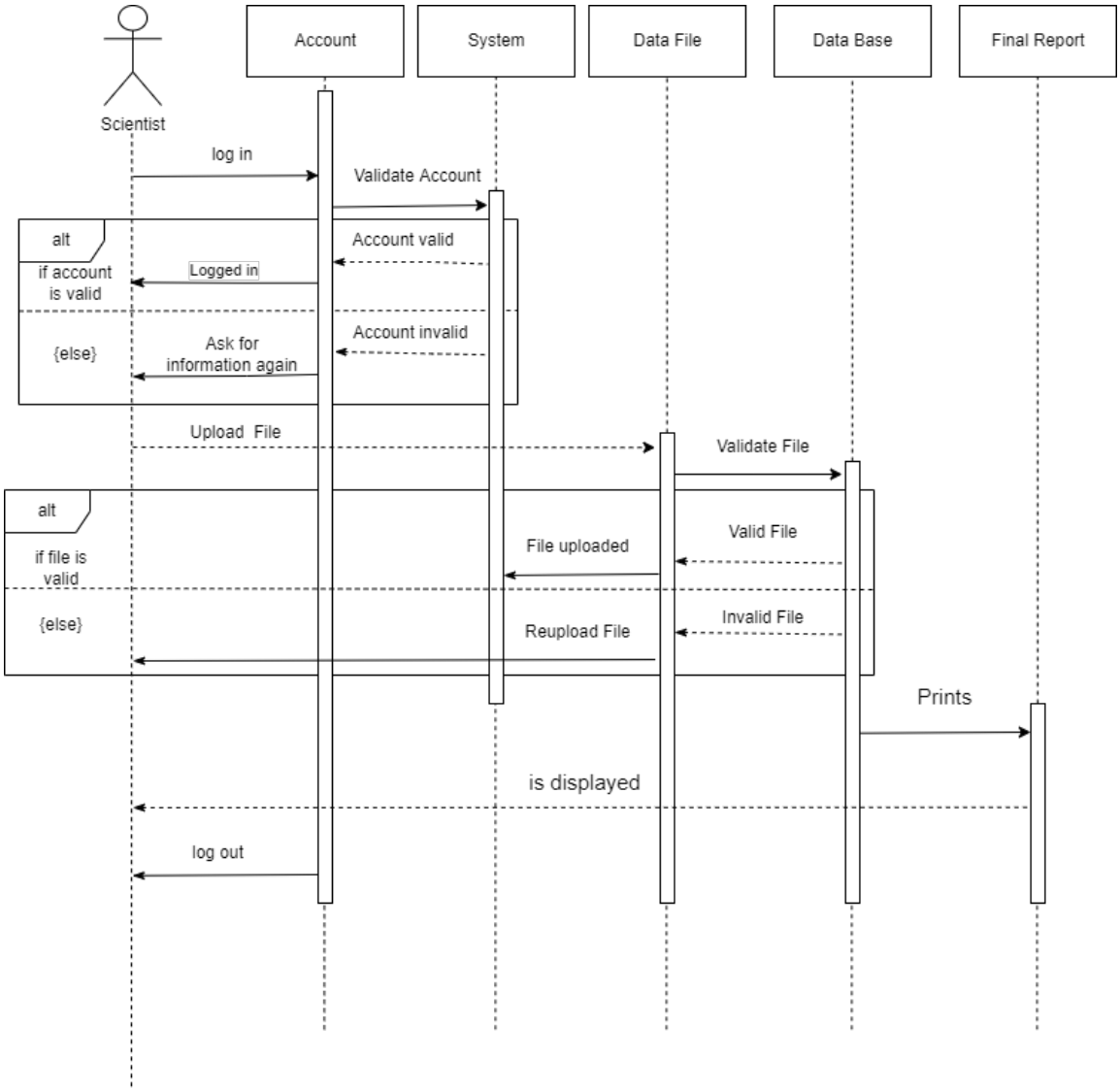


Figure 4.5: Sequence Diagram

4.2.5 Flowchart

Represents the workflow of the system. Figure 4.6 shows that Home page gets displayed, four options will be in the page, contact us, login, show upload history and register.contact us page will have the team's numbers and emails. Moving forward to registering, it is made for newly registers whereas, login is for people who already have accounts, lastly show upload history will show old uploaded files. Once you log in, the user will upload the file, afterwards, will either start the process or cancel the upload. When the user decides to proceed with the process, the user will either choose a database or a default database and the user will be able to edit with the settings as he wishes. Afterwards, the system prints a final report.

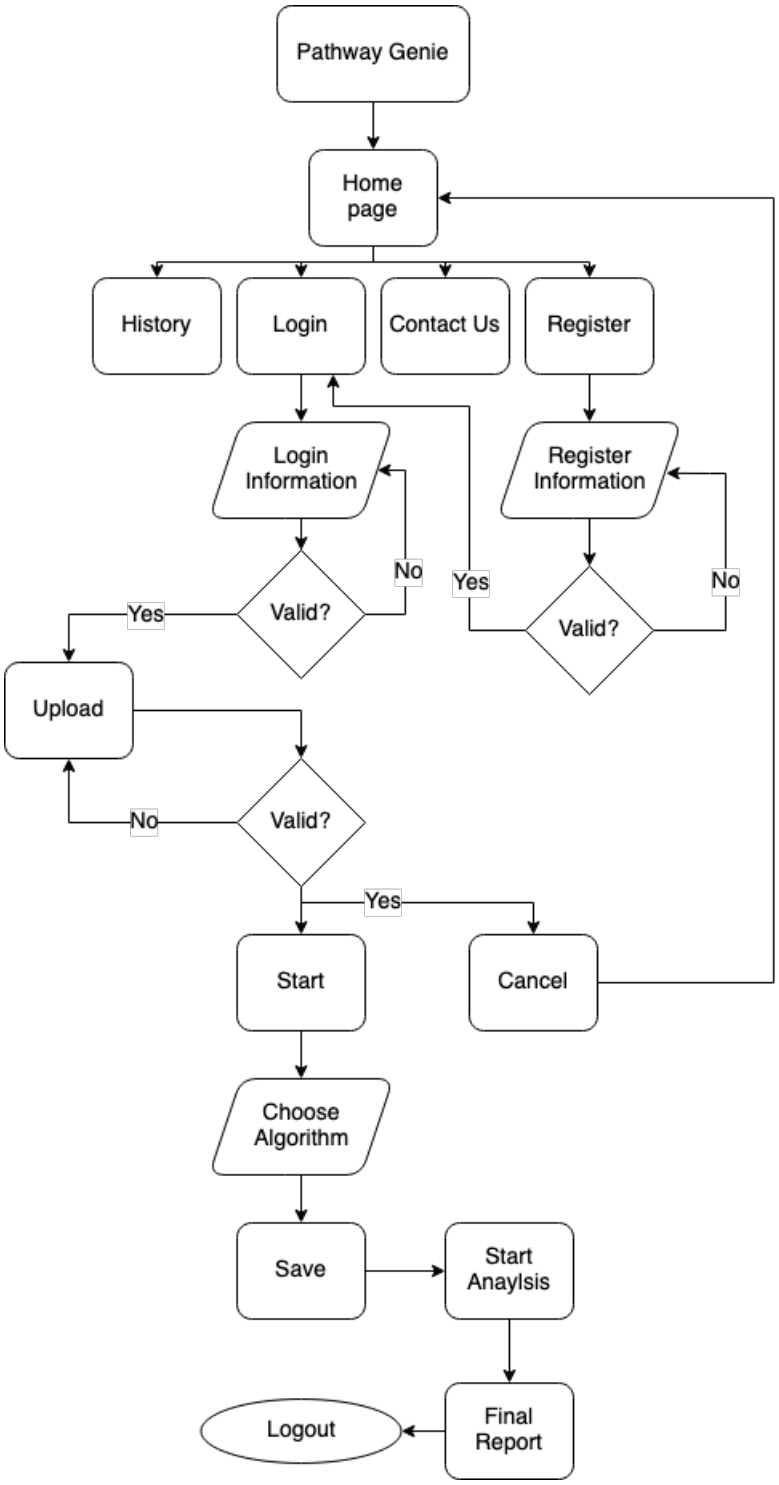


Figure 4.6: Flow Chart

4.2.6 Data flow

Figure 4.7 shows how data flows, the zoomed in process in the diagram was to explain how the data will move from the scientist to the pipeline until final report and going back.

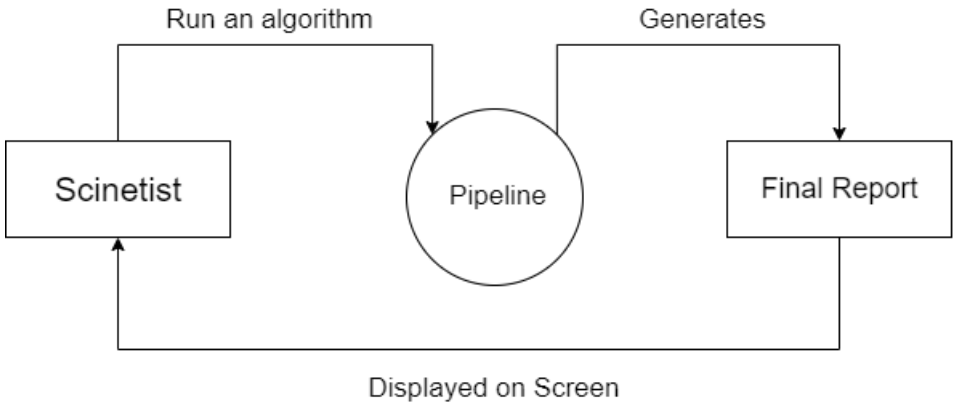


Figure 4.7: Data Flow diagram

4.3 Prototype

A medium-fidelity prototype was created based on the requirements of the Pathway Genie tool using Proto.io. Figure 4.8 shows the home screen when you first open the website. Then if you click on get started button, sign up window will pop up as shown in Figure 4.9. Figure 4.10 shows the sign in page. Figure 4.11 shows the pathway analysis page which contains 3 steps. First step is uploading file, after uploading a file user can change settings in step 2 and 3 as shown in Figure 4.12 and 4.13. Figure 4.14 shows pipeline steps after starting the analysis. Lastly Figure 4.15 show the final report.

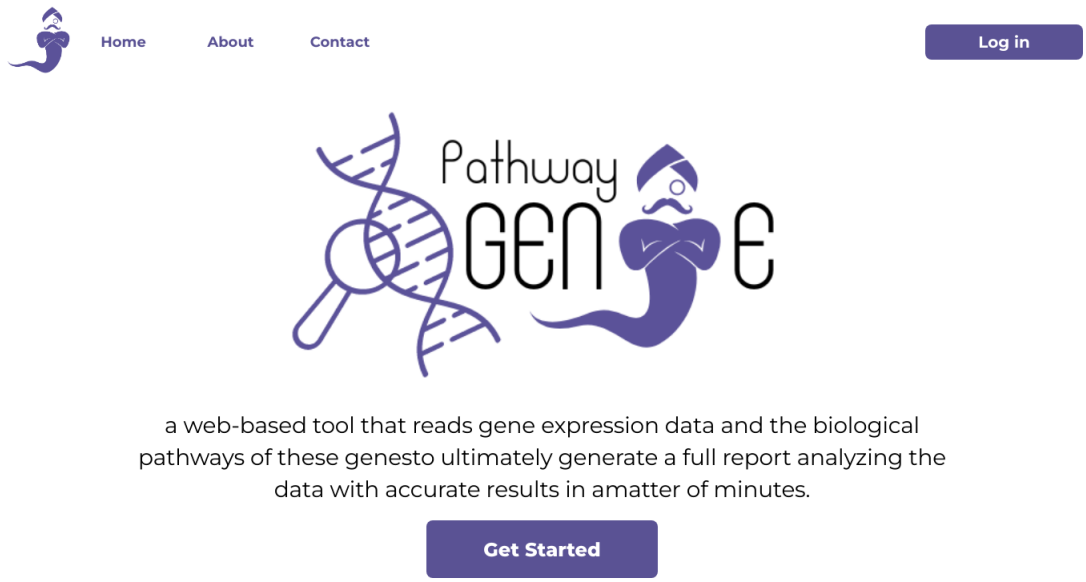


Figure 4.8: Home page

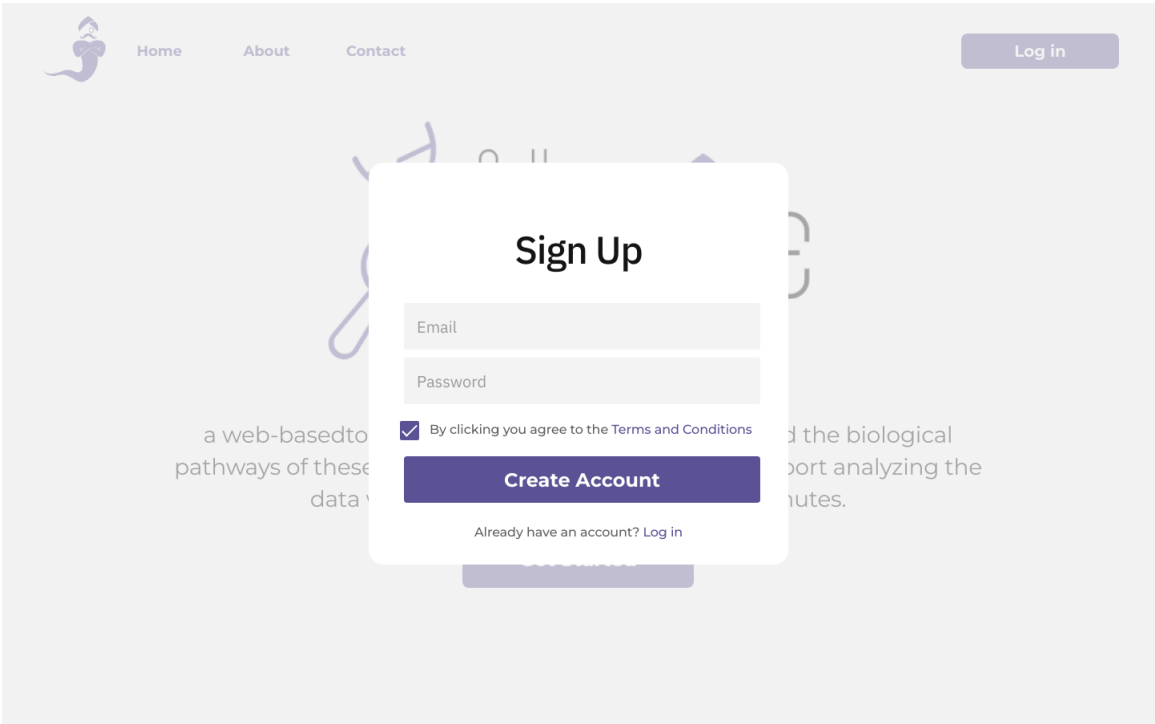


Figure 4.9: Sign Up page

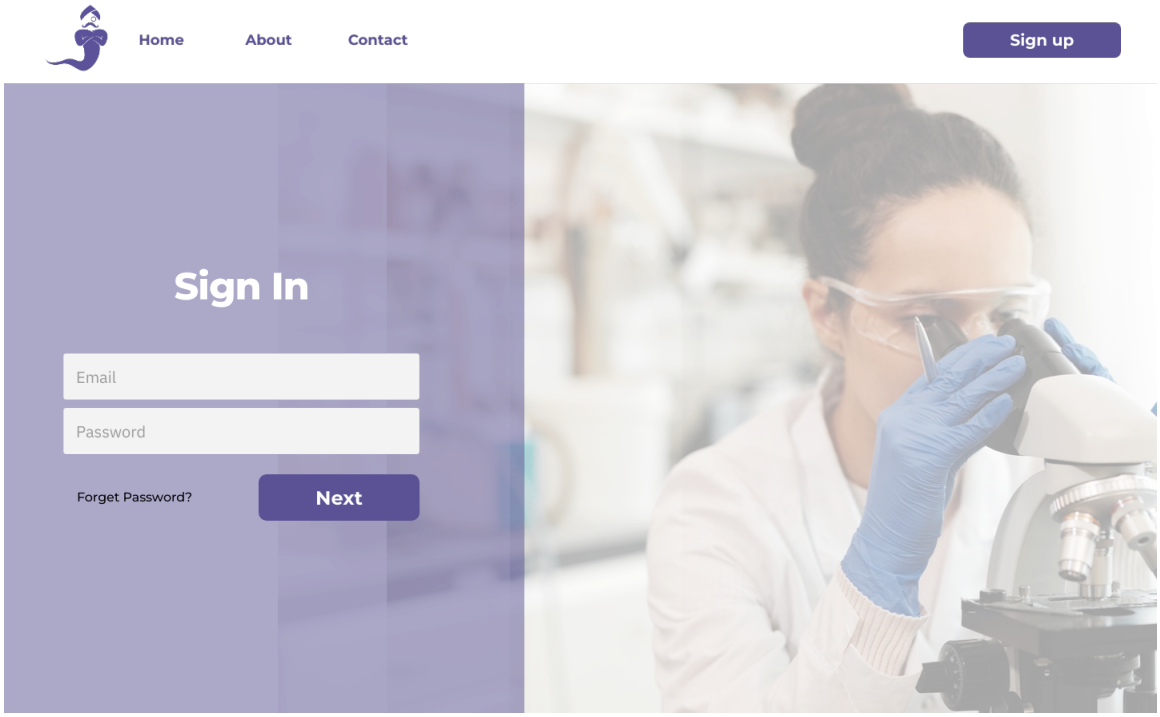


Figure 4.10: Sign In page

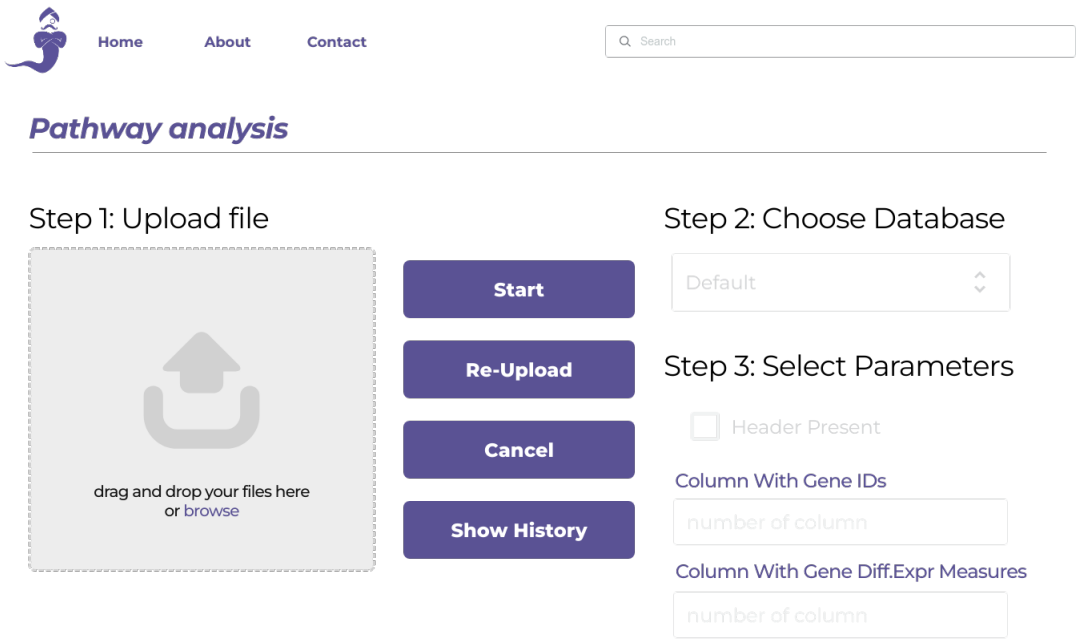


Figure 4.11: Upload page

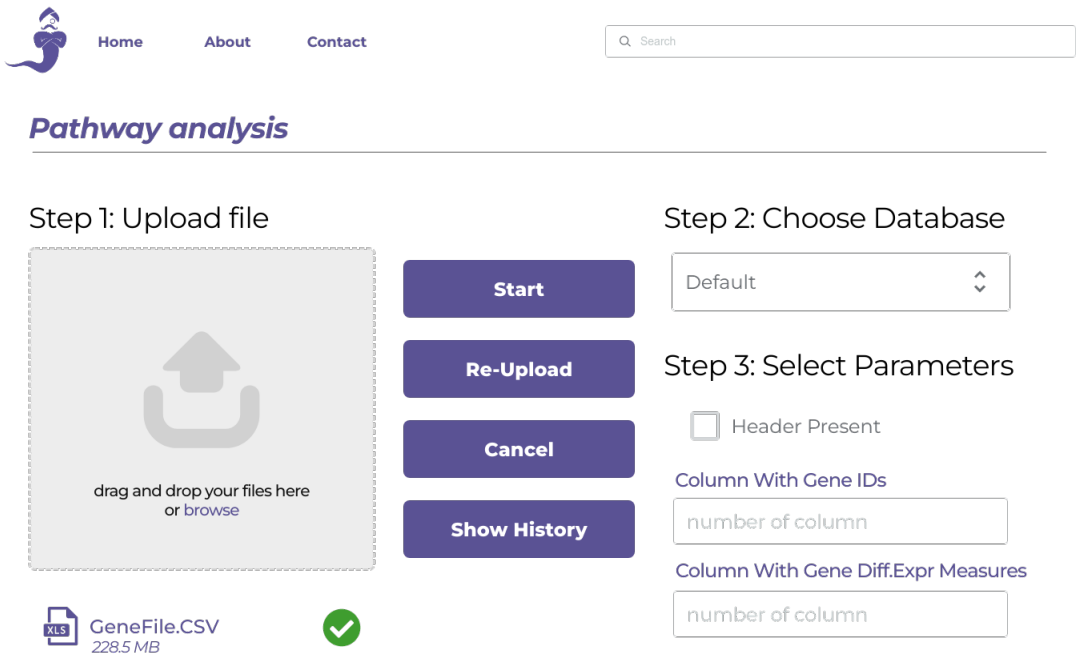


Figure 4.12: After Uploading a file

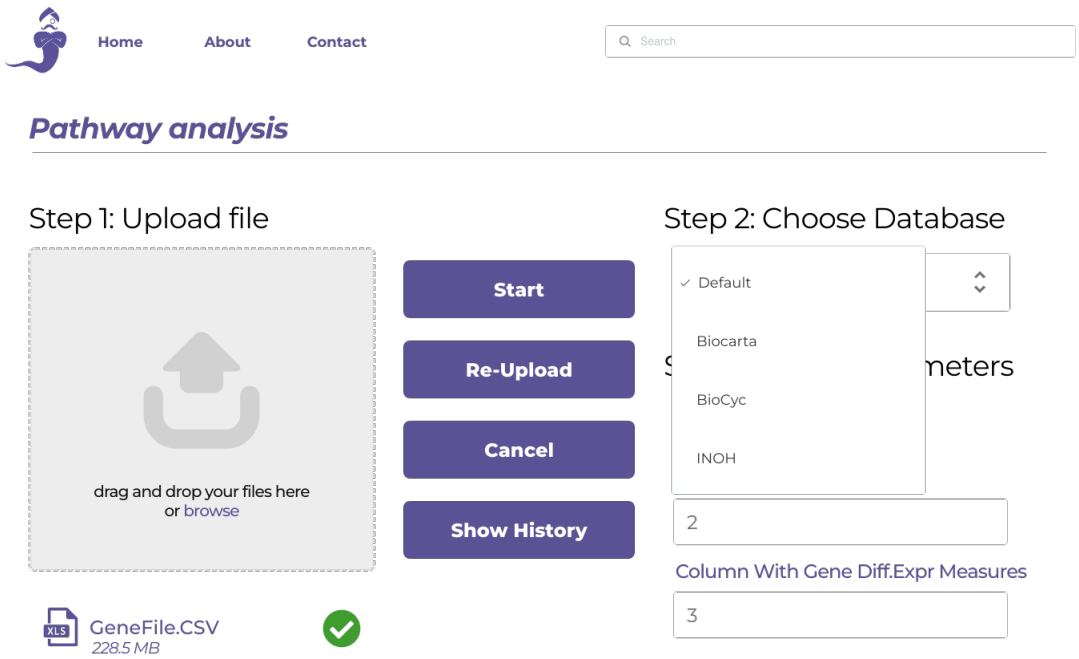


Figure 4.13: Changing settings

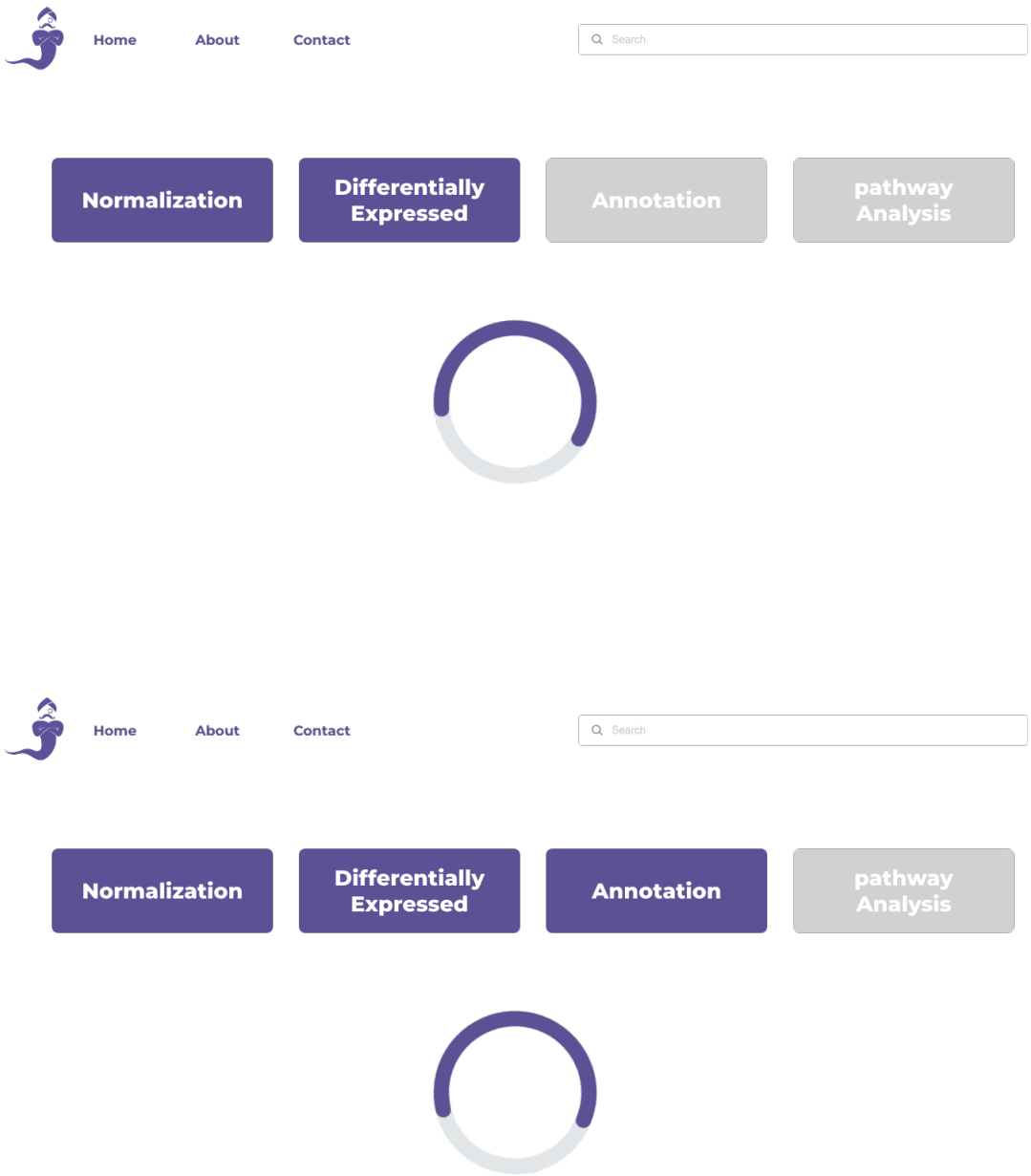


Figure 4.14: Running pipeline

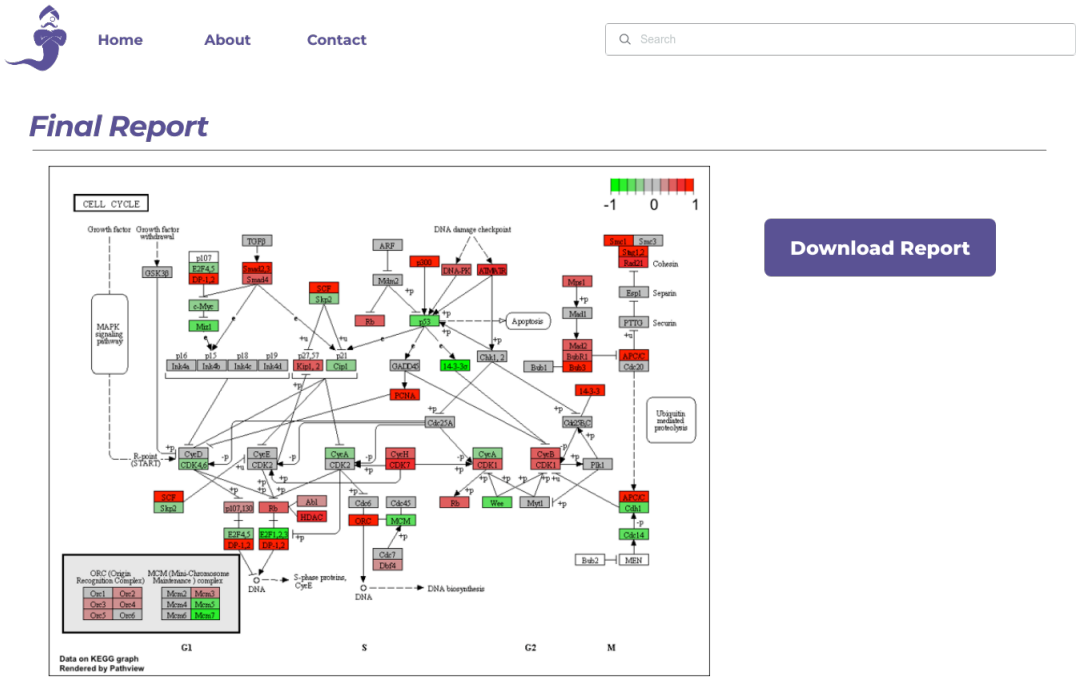


Figure 4.15: Final Report Page

4.4 Conclusion

Concluding this chapter, use cases, class diagrams, database diagrams, sequence diagram, flowchart, and data flow help with understanding the system and learn more about how data flow. The implementation will be shown in the next chapter.

CHAPTER 5

IMPLEMENTATION

5.1 Summary of Previous Work

Pathway Genie is a web-based tool that reads gene expression data and the biological pathways of these genes to ultimately generate a full report analyzing the data with accurate results. In order to understand the needs of the users better, a stakeholder survey and interview were conducted. Furthermore, thorough discussions about different Fields related to pathway genie. In addition, UML diagrams, Sequence diagrams, flow charts, and other more were made to give us a hint about the pathway genie's structure, model, and system.

The upcoming chapter describes the development of the pathway genie. Starting with section 5.2 discussing activities and actions, and section 5.3 tools and platforms used in implementation. Section 5.4 front-end implementation, Section 5.5 database implementation, AI model implementation in section 5.6. Lastly, section 5.7 conclusion.

5.2 Activities and Actions of the Development

Pathway Genie consists of three layers:

1. **Front-end layer:** This layer includes everything the end-user sees and needs to interact with the system.

2. **Back-end layer:** This layer has the database, The database contains all of the system's information that can be saved and accessed.

3. **AI model layer:** This layer is considered the system's core. More information about each layer's development will be provided in the following

sections.

5.3 Tools and platform

The following tools were used to implement the project:

5.3.1 HTML

The most popular language on the Internet for creating web pages is called Hyper Text Markup Language HTML. some of HTML applications are: HTML is used in the development of web pages that are displayed online. Almost every web page has html tags that allow it to be rendered in a browser. Internet Navigation - HTML offers tags that are used to move between pages, and is a major component of internet navigation. Because of the responsive design approach, HTML pages today perform effectively on all platforms, including mobile, tablets, desktops, and laptops. Offline assistance Once loaded, HTML pages can be accessed locally on the machine without an internet connection. HTML5 now includes native support for rich experiences, making it beneficial in the game creation industry. (HTM, 2023)

5.3.2 CSS

Cascading Style Sheets CSS is a style sheet language used to describe the presentation of a document written in a markup language such as HTML. CSS allows you to apply styles to web pages and allows developers and designers to define how elements behave and appear in browsers. It also specifies how a web page should look, including its colors, fonts, spacing, and other details. CSS is a rule-based language, which means that you define the rules by specifying groups of styles to be applied to specific elements or groups of elements on a web page. (CSS, 2023)

Other tools that can be implemented and called via CSS and HTML that we used with CSS were:

- **Google Fonts**

  Google Fonts is a Google-owned computer font and web font service. It includes font families that are free and open-source, an interactive web directory for browsing the library, and APIs for using the fonts with CSS. It also includes a collection of web fonts that you can freely use in your HTML web pages. We chose one font family to apply to all elements in the pages on our tool's website, and it was the font-family: Poppins that is Designed by Indian Type Foundry, Jonny Pinhorn (Goo, 2023)

- **Bootstrap**

  Bootstrap is a free and open-source CSS framework for front-end web development that is responsive. Bootstrap includes HTML, CSS, and JavaScript design templates and it focuses on making the creation of informative web pages easier.

  We used Bootstrap to download and use Font Awesome icons and features. (Boo, 2023)

- **Font Awesome**

  Font Awesome is a CSS and Less-based icon library and toolkit that was made by Dave Gandy for use with Bootstrap. This tool is highly used by millions of designers, developers, and content creators. (Fon, 2023)

5.3.3 MySQL

The most widely used open source database worldwide is My Structured Query Language MySQL, it is the second-most popular database, after Oracle Database. (MYS, 2023)

5.3.4 PHP

Hypertext Preprocessor PHP is a server programming language, it is a strong tool for creating interactive Web sites.

5.3.5 RStudio

R is a free software environment and programming language for statistical graphics and computing. For creating statistical software and conducting data analysis, statisticians and data miners frequently use the R programming language.

Rstudio was the first step in the project, Rstudio was used to normalize and annotate the data to organize them. lastly, the r script was converted to cvs for python to analyze.

5.3.6 Python

Python is one of the most widely used languages for AI applications, it was used to implement the AI model of the Pathway Genie. the cvs file was imported in python, we extracted the data, split them, applied machine learning algorithm and calculated the AI model accuracy.

5.3.7 JavaScript

JavaScript is a programming language that, along with HTML and CSS, is one of the core technologies of the World Wide Web. JavaScript is a scripting or programming language that allows complex features to be implemented on web pages. It allows you to create dynamically updated content, control multimedia, and animated images, among other things. (web, 2023)

5.3.8 Visual Studio

You may edit and develop using the Visual Studio as your creative integrated development environmentIDE. Visual studio was used in our project to program in various languages. (Vis, 2023)

5.3.9 Formspree

Formspree is an online service that allows you to create functional HTML forms without using PHP or JavaScript. The Formspree API makes it simple and flexible to add forms to a web page. To prevent the latest attack vectors, it employs an industry-leading Web Application Firewall with automatic updates. Formspree was used on our website via the Contact Us form.

5.4 Front-end Implementation

- **Home Page And Login/Sign up Page:**

  The user is first welcomed by the home page which is the first page that appears before logging in, as shown in figure 5.1, and when the user clicks on the GET STARTED or the LOGIN/SIGN UP button, the login page first appears, which is a simple page where users only enter their email and password to log in, as shown in figure 5.2, and if users do not have an account, they can register using the Signup page, where they will enter their email, username, and password as shown in figure 5.3.

  When the user logs in to their account, they see a new home page that illustrates the analysis pipeline, displays the user name, and gives the user access to all the pages in the website as shown in figure 5.4.

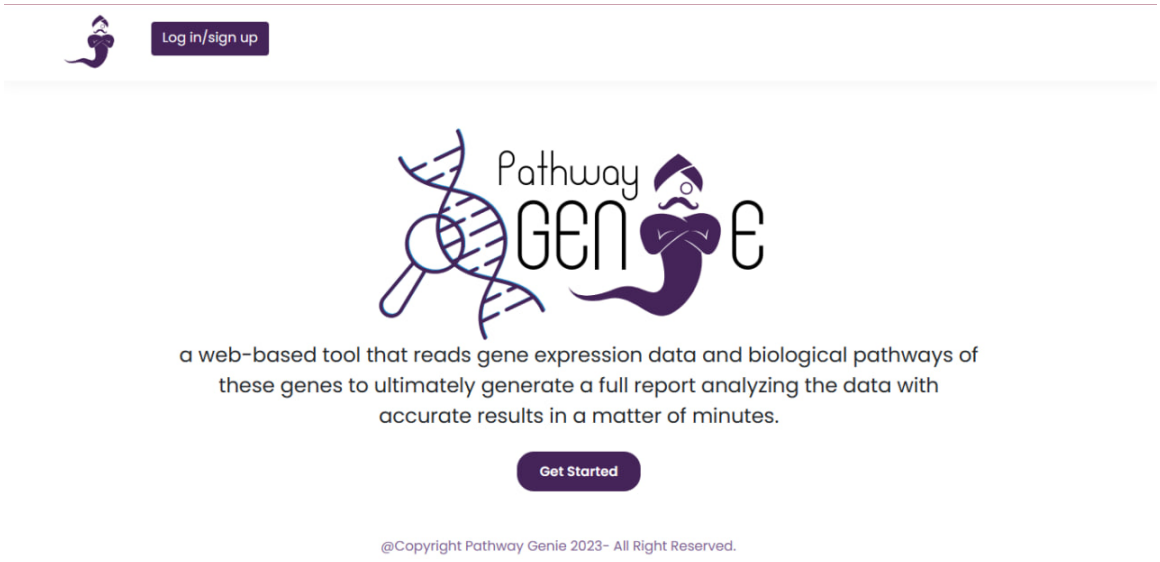


Figure 5.1: Home Page

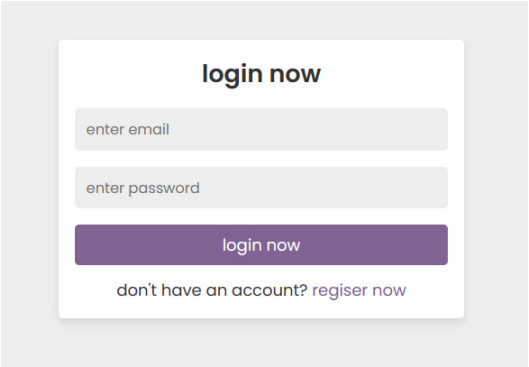


Figure 5.2: Login Page

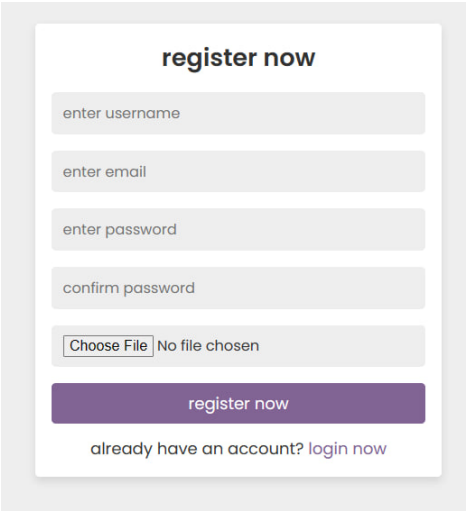


Figure 5.3: Register Page

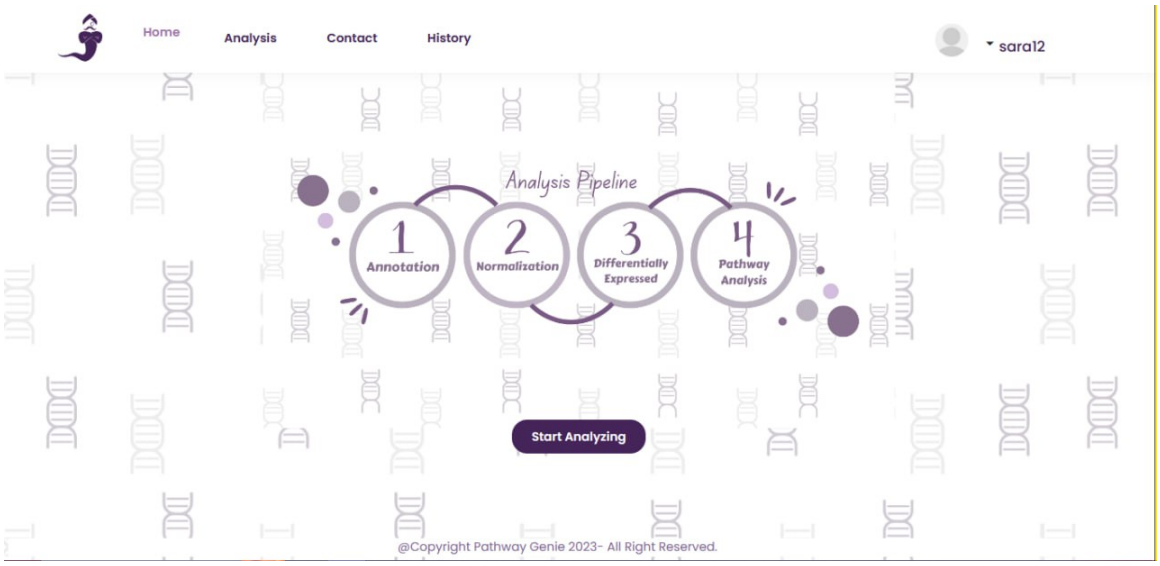


Figure 5.4: Home Page when logged in

When users do not comply with the terms and conditions for logging in or signing up, the website does not allow them to enter and one of two messages appears:

- Empty Fields Error:

  The user did not enter the email or password, or both either in the login or register page, the message will present as shown in figures 5.5 and 5.6.

- Incorrect Information Error:

  If the user enters incorrect information in the email or password or both either on the login or register page, the message will present as shown in figure 5.7.

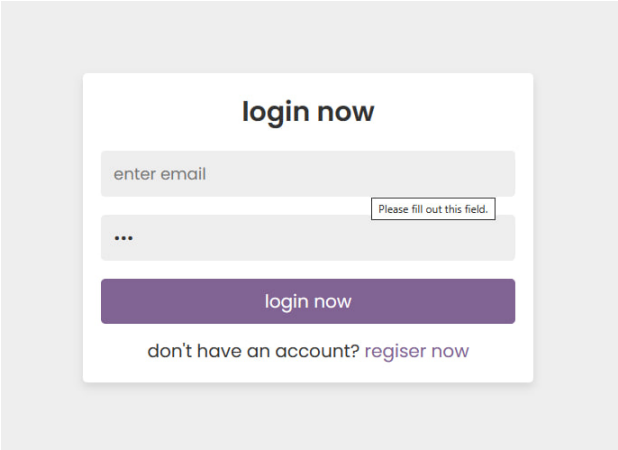


Figure 5.5: Login Empty Email

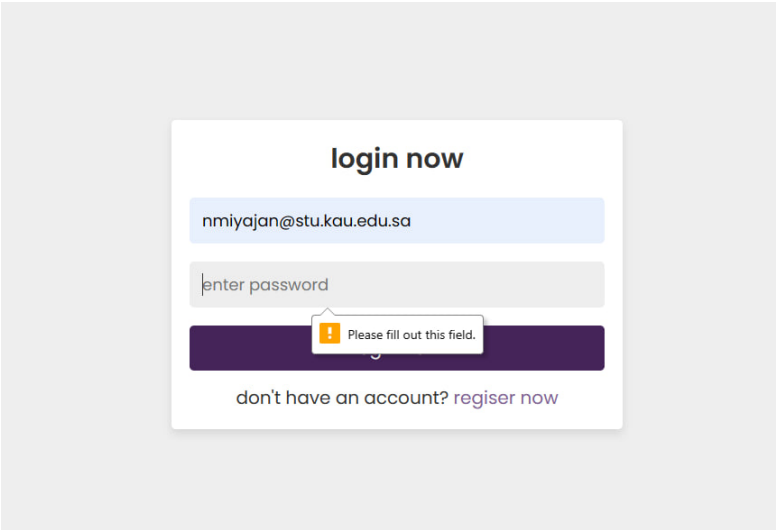


Figure 5.6: Login Empty Password

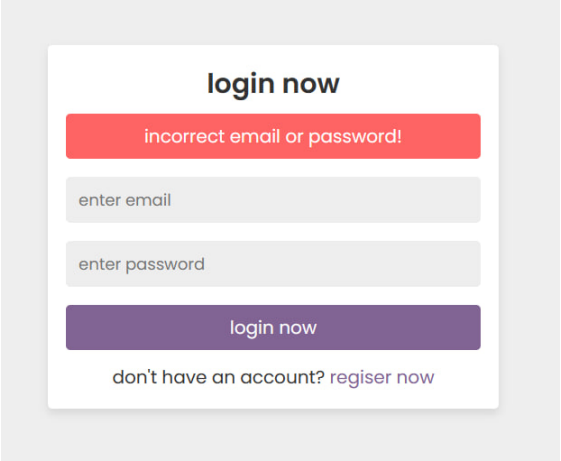


Figure 5.7: Wrong Password or Email

- **Contact Us Page:**

  Users can reach out to us with inquiries from our contact us page as shown in figure 5.8.

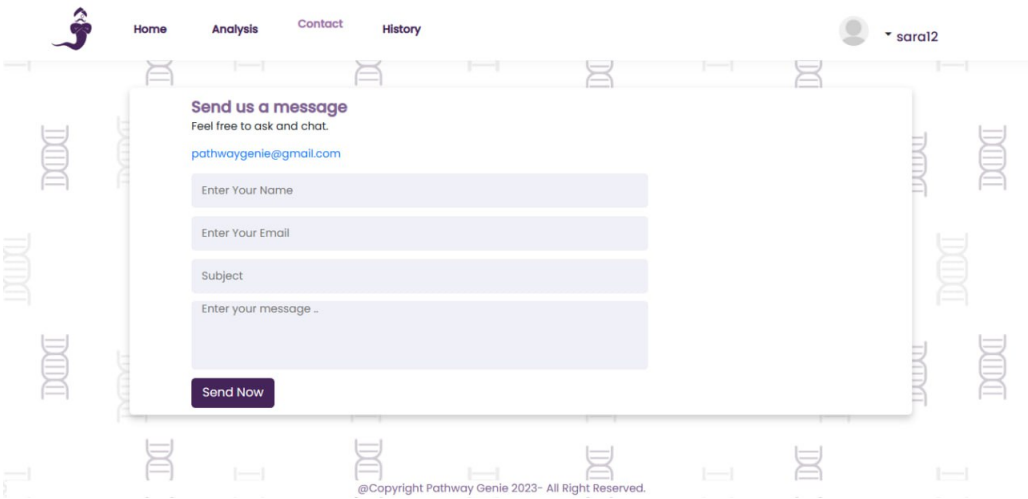


Figure 5.8: Contact Us Page

- **Account Page:**

  The user can get to this page from the drop-down list by the user name as shown in figure 5.9.

  From this page, the user can edit their personal information and update their password and profile avatar. while also being able to log out or go back to the home page as shown in figures 5.10 and 5.11.

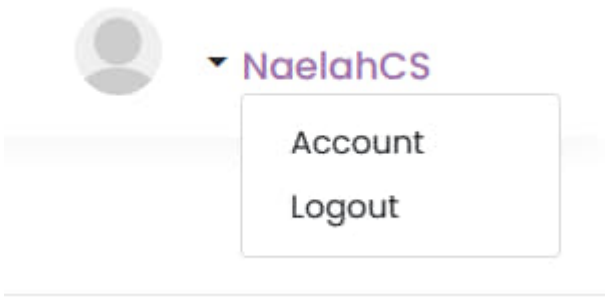


Figure 5.9: Username drop-down list

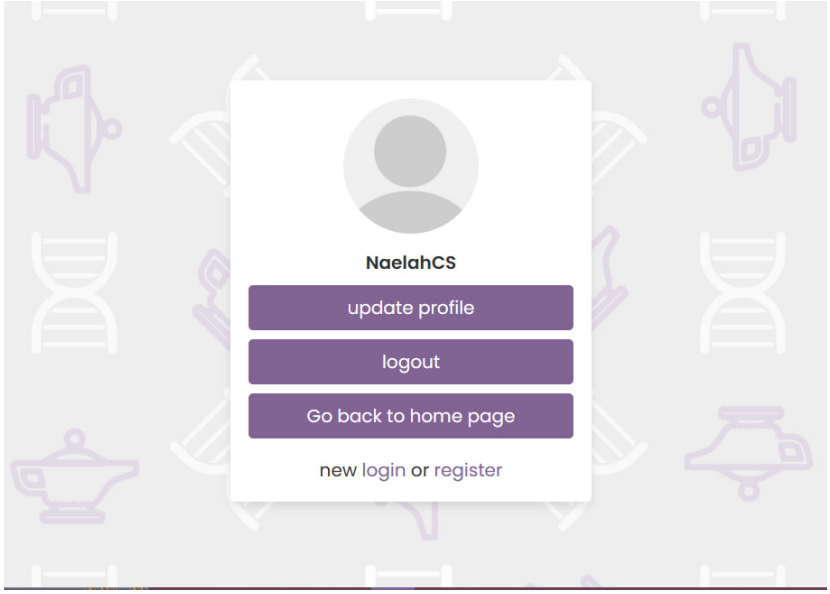


Figure 5.10: Account page

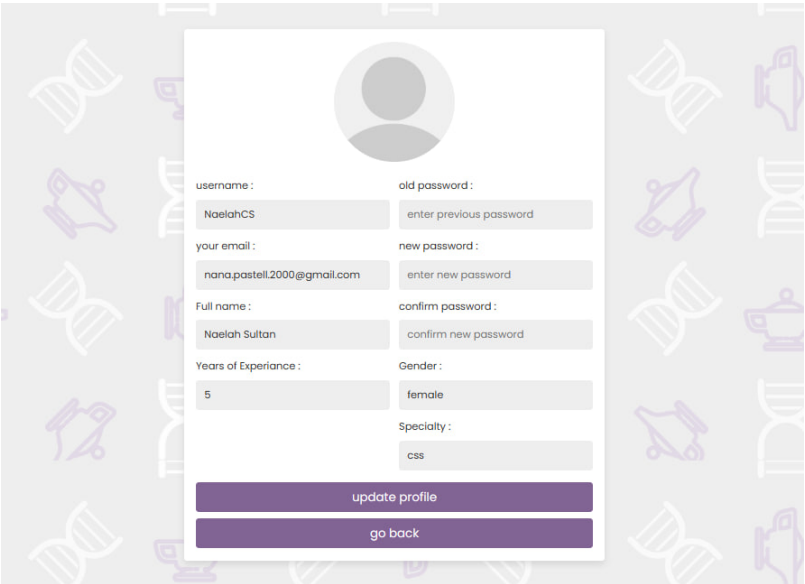


Figure 5.11: Update Account

- **Analysis Page:**

  The analysis page is where the user's data set will be uploaded as a CSV file, it will also give the user options to apply to the chosen algorithms that are listed as shown in figures 5.12 and 5.13.

  If the file uploaded is not in a CSV format the user will receive error messages as shown in figure 5.14.

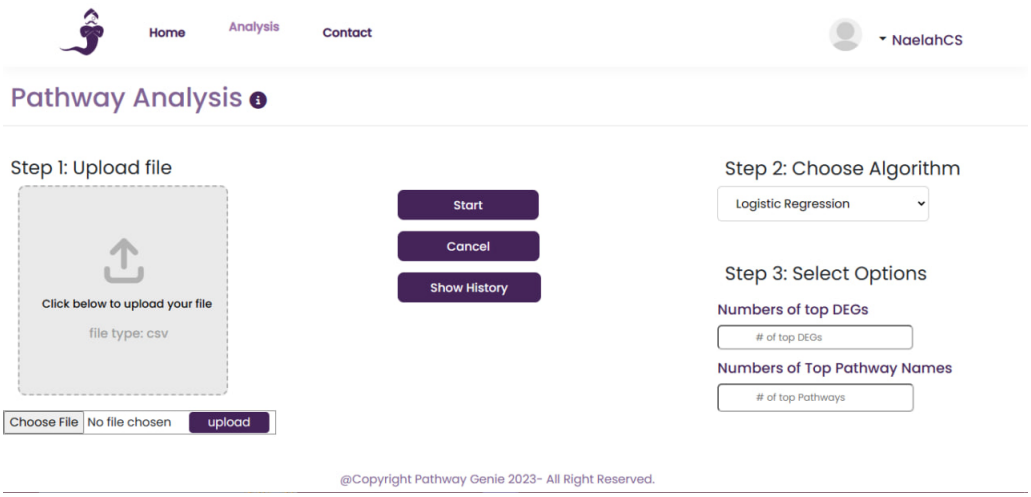


Figure 5.12: Analysis Page

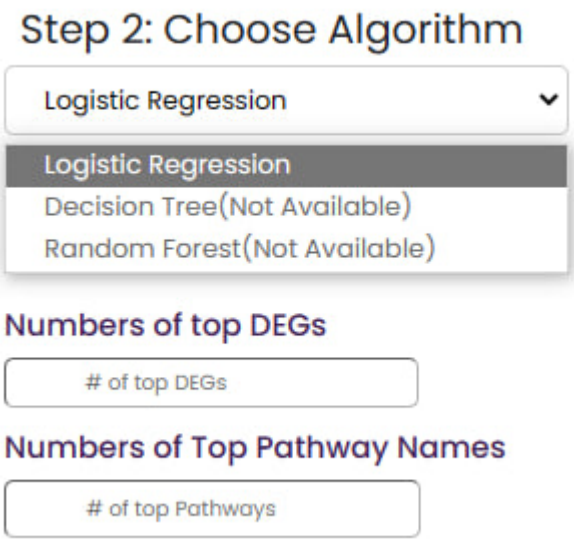


Figure 5.13: Listed Algorithms Choices

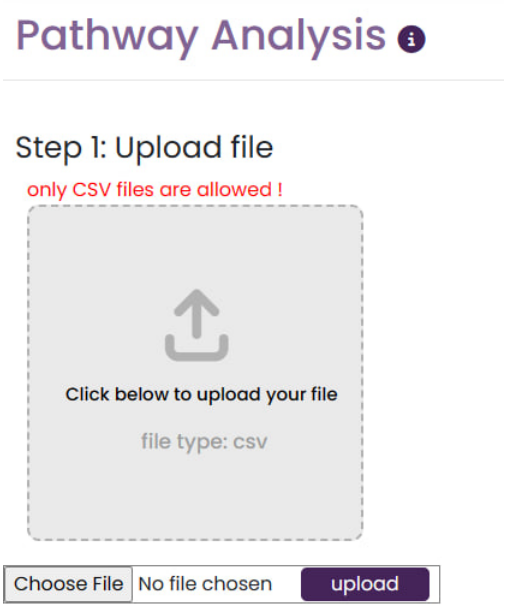


Figure 5.14: Only CSV files are allowed error

– Algorithms info

To get to this page the user clicks the 'i' icon on the analysis page. This page is simply to give the user an explanation of the differences between the algorithms that are listed in the analysis page as shown in figure 5.15. (Ran, 2023) (Dec, 2023) (log, 2023)

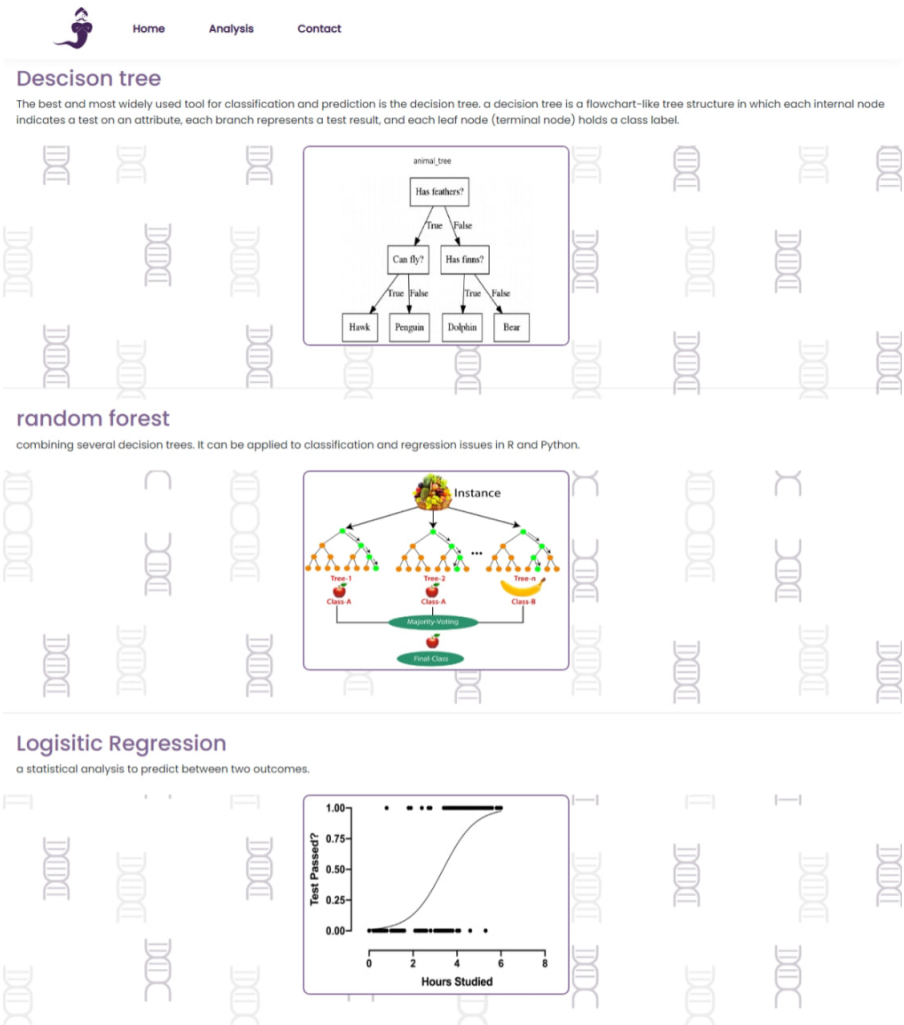
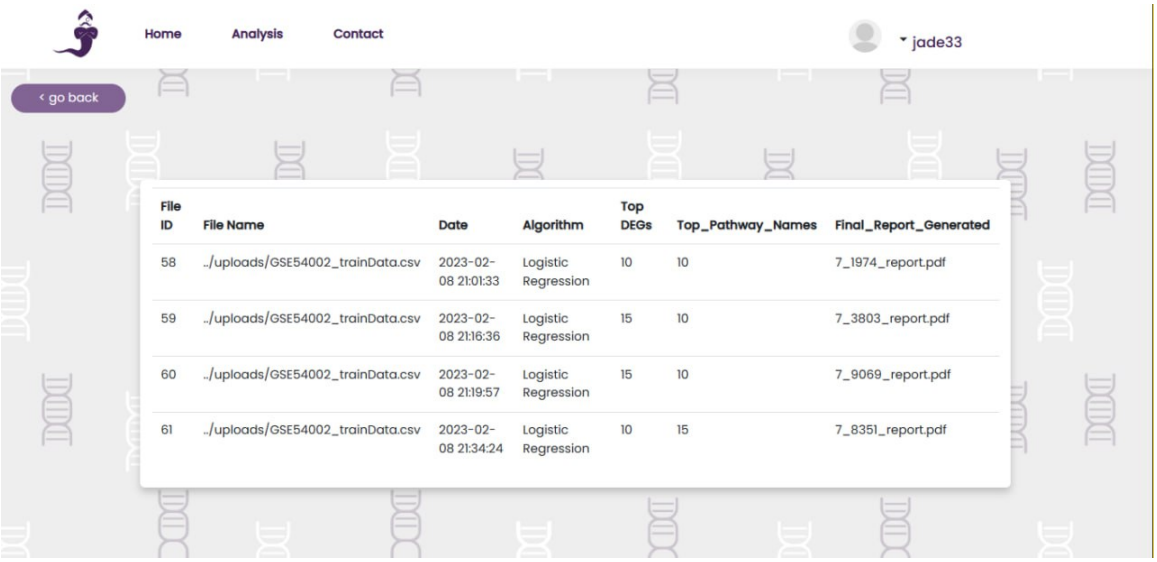


Figure 5.15: Algorithms information

– History

The History page will display all files that were uploaded to the database, such as the final reports generated and all the data files that the user uploaded, while also displaying the date of when the file was uploaded and the entered Numbers of top DEGs, Numbers of Top Pathway Names and the selected algorithm as shown in figure 5.16.



File ID	File Name	Date	Algorithm	Top DEGs	Top_Pathway_Names	Final_Report_Generated
58	../uploads/GSE54002_trainData.csv	2023-02-08 21:01:33	Logistic Regression	10	10	7_1974_report.pdf
59	../uploads/GSE54002_trainData.csv	2023-02-08 21:16:36	Logistic Regression	15	10	7_3803_report.pdf
60	../uploads/GSE54002_trainData.csv	2023-02-08 21:19:57	Logistic Regression	15	10	7_9069_report.pdf
61	../uploads/GSE54002_trainData.csv	2023-02-08 21:34:24	Logistic Regression	10	15	7_8351_report.pdf

Figure 5.16: History Page

– **Final Report Page:**

The report generated by the algorithm analysis will be displayed on the final report page where the user can view and print it as shown in figure 5.17.

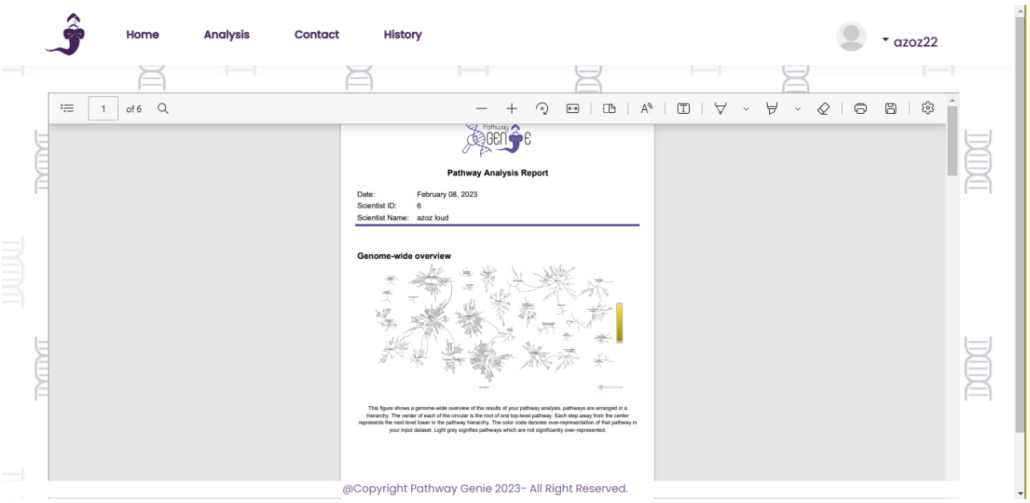


Figure 5.17: Final Report page

– **Logout:**

When users log out, a display message that reads "Thank you for using Pathway Genie!" appears before they are taken back to the initial home page as shown in figure 5.18.

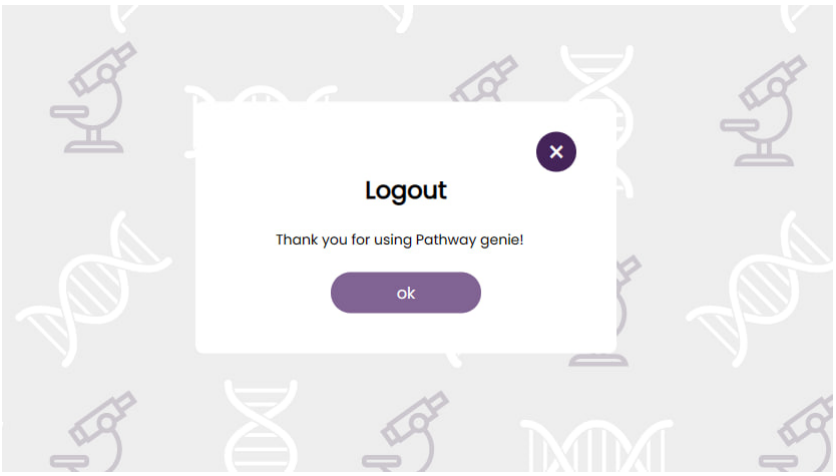


Figure 5.18: Log out page

5.5 Database Implementation

A database is a structured collection of data that is digitally stored and accessed. It contains critical information that the system will save and retrieve. It is significant because it enables users to manage and access data more easily.

SQL was used to build Pathway Genie's database. Structured Query Language (SQL) is a programming language designed for data management. It is especially useful when dealing with structured data, which includes relationships between entities and variables.

It first introduced the concept of accessing multiple records with a single command. Second, it removes the need to specify how to access a record, such as with or without an index.

SQL server only supports one storage engine, whereas MySQL supports multiple storage engines, each with its own set of specifications.

MySQL Workbench includes data modeling, SQL development, and extensive administration tools for server configuration, user management, backup, and much more. It also includes a visual console for easily administering MySQL environments and improving database visibility.

In Pathway Genie there are 2 tables that have been built "file" and "user" as shown below in figures 5.19 and 5.20.

Result Grid | Filter Rows: | Edit: | Export/Import: | Wrap Cell Content:

file_id	user_id	file_name	date	algorithm	num_DEGs	num_pathways	report
58	7	../uploads/GSE54002_trainData.csv	2023-02-08 21:01:33	Logistic Regression	10	10	7_1974_report.pdf
59	7	../uploads/GSE54002_trainData.csv	2023-02-08 21:16:36	Logistic Regression	15	10	7_3803_report.pdf
60	7	../uploads/GSE54002_trainData.csv	2023-02-08 21:19:57	Logistic Regression	15	10	7_9069_report.pdf
61	7	../uploads/GSE54002_trainData.csv	2023-02-08 21:34:24	Logistic Regression	10	15	7_8351_report.pdf
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL

Figure 5.19: Pathway Genie File Table

Result Grid | Filter Rows: | Edit: | Export/Import: | Wrap Cell Content:

id	fullname	user_name	user_email	gender	specialty	years	password	pic	date
6	azoz loud	azoz22	azozzz200@gmail.com	NULL	NULL	NULL	202cb962ac59075b964b07152d234b70	default-pic.png	2023-02-07 20:26:46
7	jade ali	jade33	jj@gmail.com	male	html	4	dc468c70fb574ebd07287b38d0d0676d	default-pic.png	2023-02-08 17:58:26
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL

Figure 5.20: Pathway Genie user Table

5.6 AI Model Implementation

As mentioned before, the AI model is the core of the project. Before we build the model, we annotated the data .then our AI model takes our data set. Afterward, To train the model, we used logistic regression to classify the data into a tumor and non-tumor samples. By using a specific feature that will be discussed further in the classification section, the model will detect deferentially expressed genes DEGs. Finally, it calculates the accuracy of the prediction.

5.6.1 Data Annotation

The process of labeling data with appropriate tags to make it simpler for computers to understand and analyze is known as data annotation. Since the effectiveness and accuracy of supervised learning models depend on the quality and quantity of annotated data, annotated data is an essential component of these models. for our train-test data we have done the annotation process manually, we labeled 20 tumors and 17 non-tumor samples according to the dataset information in ncb (2022).

5.6.2 Classification

Logistic regression is utilized in predictive analytics and classification. It is used to predict a binary result based on a collection of independent variables. There are only two outcomes that can occur in a binary situation: either the event occurs (1) or it doesn't (0). Independent variables are elements that could have an impact on the result. (Swaminathan, 2018)

The learning algorithm was used in the pathway genie to estimate the differentially expressed genes DEGs from the microarray gene expression data,

then detect the pathways from the DEGs. To build the model first we split the data into a train (%25) and test (%75) sets. Then Initialize our classifier and fit the train data to train our model. Then make predictions using the test set and calculate Accuracy by comparing predictions with the test set's label.

After the model classified data into a tumor and non-tumor samples we used the logistic regression feature coefficient to detect the DEGs. The regression coefficient indicates whether a change in a predictor variable increases or decreases the likelihood of the occurrence. In general, positive coefficients increase the likelihood of the event, while negative coefficients decrease the likelihood of the event. (Min, 2022)

5.6.3 Results

The following evaluation metrics are used to measure the classifier's performance. This section provides the results of the testing set. The metrics are:

- **Classification Report**

  For any classification report, you will find precision, recall, an F1 score, and support. recall refers to how many of a particular class you find compared to the total number of specific class elements, How many are correctly classified among that class will be the precision, the f1-score represents the harmonic mean of recall and precision, and finally, support is the number of instances of the specified class in your dataset. lastly, it calculates the accuracy, macro average, and weighted average. (krishnan, 2018) As it is shown in figure 5.21, the model was trained on 28 samples, which were divided into 75% for the test set and 25% for the training set, the training process was completed in 2 iterations with a weighted average is 96%, precision 97%, recall 96%, F1-score, and the accuracy 96%.

```
print(classification_report(y_test, y_pred))
✓ 0.2s

              precision    recall  f1-score   support

           0       0.93      1.00      0.96        13
           1       1.00      0.93      0.97        15

    accuracy                           0.96        28
   macro avg       0.96      0.97      0.96        28
weighted avg       0.97      0.96      0.96        28
```

Figure 5.21: Classification Report

- **Confusion matrix**

  A table is used to measure the model's performance. The calculation for the confusion matrix in figure 5.22 is for the test set. The elements that are calculated to get the matrix are:

  - True Positive (TP): Number of correct predictions for positive class.
  - False Positive (FP): Number of false predictions for positive class.
  - True Negative (TN): Number of correct predictions for negative class.
  - False Negative (FN): Number of false predictions for negative class.

  The confusion matrix plot in figure 5.22 refers to the model predicting 27 correct predictions out of 28.

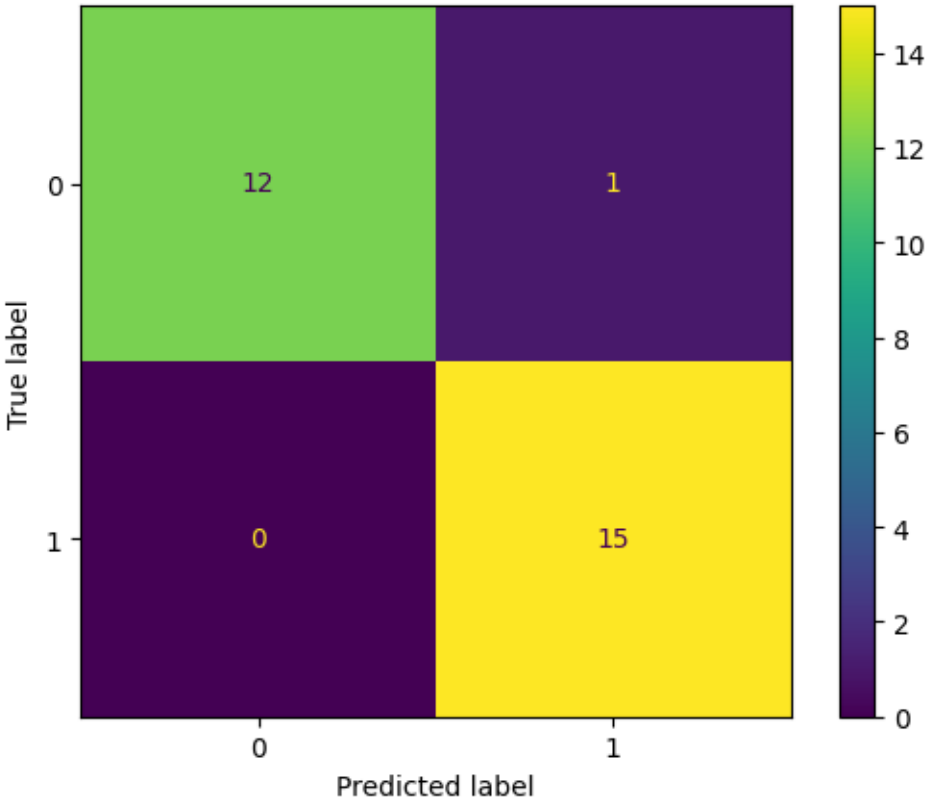


Figure 5.22: Confusion matrix plot

- **AUC - ROC Curve**

  performance measurements for classification problems. AUC score is a measurement of how much the model can distinguish between two classes and how far it is from guessing randomly. ROC curve is a graph showing the performance of a classification model at all classification thresholds goo (2022). The ROC curve is plotted with TPR against the FPR, where TPR is on the y-axis and FPR is on the x-axis Narkhede (2018). TPR is True Positive Rate or called Sensitivity and is calculated using:

$$TPR = \frac{TP + FN}{TP}$$

  FPR is False Positive Rate or Specificity and is calculated using:

$$FPR = \frac{FP + TN}{FP}$$

  Figure 5.23 shows the result of our classifier AUC – ROC curve. The AUC for the model turns out to be 0.9615. This confirms that the model does a great job of classifying data.

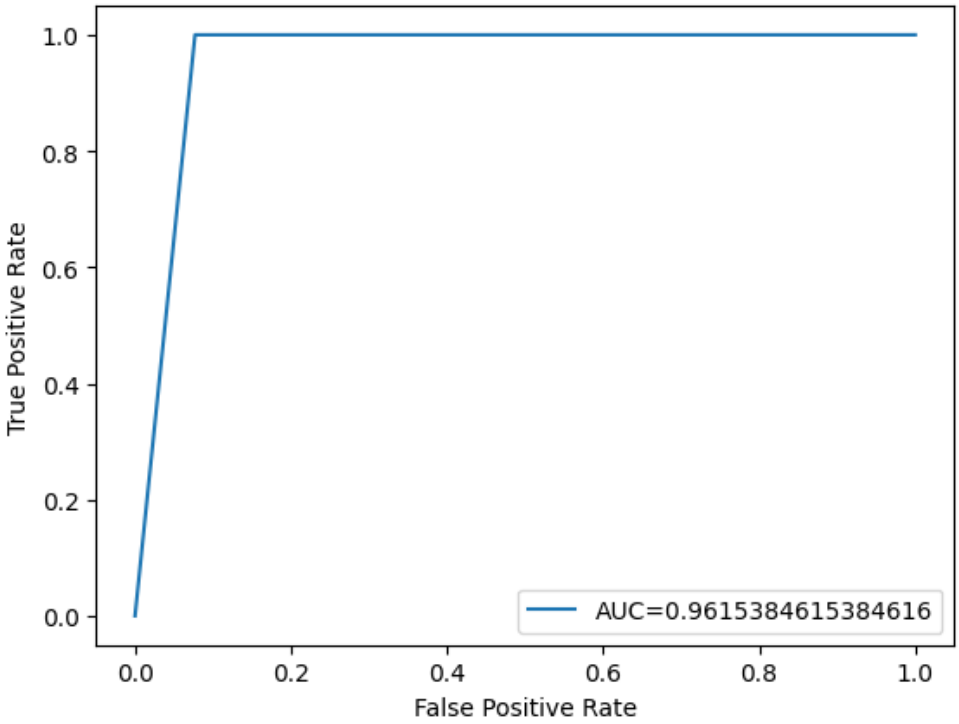


Figure 5.23: AUC - ROC Curve

5.7 Conclusion

In this chapter, we thoroughly discussed our front-end, and AI model. in addition, We described the tools and platforms we used to create Pathway Genie.

CHAPTER 6

TESTING

6.1 Introduction

In this chapter, The system is put through testing to make sure all of its parts are working well. The testing procedure also establishes whether or not the system's requirements have been fully met. In addition, testing methods will be discussed in this chapter. In section 6.2, the purpose, scope, and objective of the testing process were discussed, In section 6.3, elements of the testing were discussed Finally, each testing method is thoroughly described. Section 6.6 describes and differentiates between the two kinds of testing used in the project. Section 6.6 explains unit testing for each element, and section 6.6 describes usability testing with the target user and analyzes the testing results.

6.2 Purpose and Objectives

- **Purpose:**

  The purpose of this testing plan is to explain the testing techniques, approaches,procedures, and functionality that will be used to evaluate the developed system.

- **Objectives:**

  - Describe the testing scope, reference materials,definitions,and testing tools.
  - Specify the testing techniques to be applied.
  - Outline the testing protocols and the key requirements.
  - Describe how the testing environment is currently configured.

- Describe the testing methods applied to the application that was created.

6.3 System Components Testing

Major parts of Pathway genie's system's are the front-end , the AI model, and the back-end. The testing part makes sure that data flows correctly through these components by ensuring the following:

- Data flows between the components.
- Data presented to the scientist.
- Data fetching from the database.

6.4 Types of testing

The two testing methods used in this project are unit testing and usability testing.Software developers use unit testing to ensure that internal system functions are functioning correctly and in accordance with expectations. in addition, usability testing aims to evaluate how simple it is to use a function.The two testing techniques are each thoroughly covered in a separate section.

6.5 Unit Testing

The smallest testable components of an application, known as units, are individually and independently examined for proper operation as part of the software development process known as unit testing.The system's component parts underwent testing.the expected outcomes following each function are shown in table 6.4 after the integration process.

6.5.1 AI Model Unit Testing

As mentioned before, AI model has two processes first, we annotated the data, then AI model takes our data set. Afterward, To train the model, we used logistic regression to classify the data into a tumor and non-tumor samples.achieving an accuracy of 96% on our test set. Figure 5.23 above shows the model's performance on the test set from our dataset.

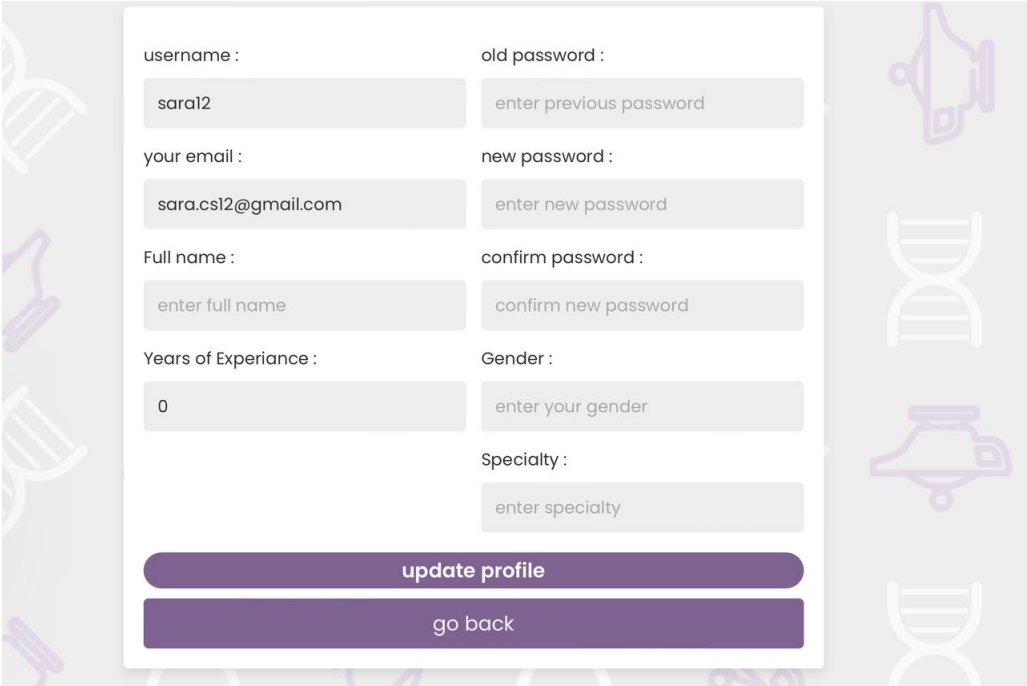
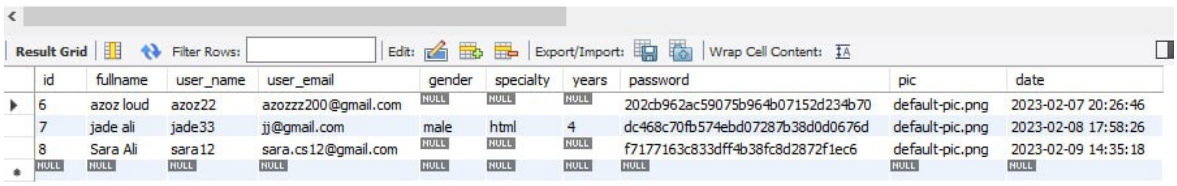


Figure 6.1: New profile account

6.5.2 Database Unit Testing

The database operations were tested using MySQL workbench, We experimented the database with creation, insertion, deletion, and update operations as shown in the figures below.

- **Creation:**
  - To see if it would appear in the database, a new profile account was made.



id	fullname	user_name	user_email	gender	specialty	years	password	pic	date
6	azoz loud	azoz22	azozzz200@gmail.com	NULL	NULL	NULL	202cb962ac59075b964b07152d234b70	default-pic.png	2023-02-07 20:26:46
7	jade ali	jade33	jj@gmail.com	male	html	4	dc468c70fb574ebd07287b38d0d0676d	default-pic.png	2023-02-08 17:58:26
8	Sara Ali	sara12	sara.cs12@gmail.com	NULL	NULL	NULL	f7177163c833dff4b38fc8d2872f1ec6	default-pic.png	2023-02-09 14:35:18
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL

Figure 6.2: The account was created in the database

- **Insertion:**
  - more information about the user was added To see if it will appear in the database
  - uploading the CSV file to check if the file will be uploaded in the database

as shown in figures below

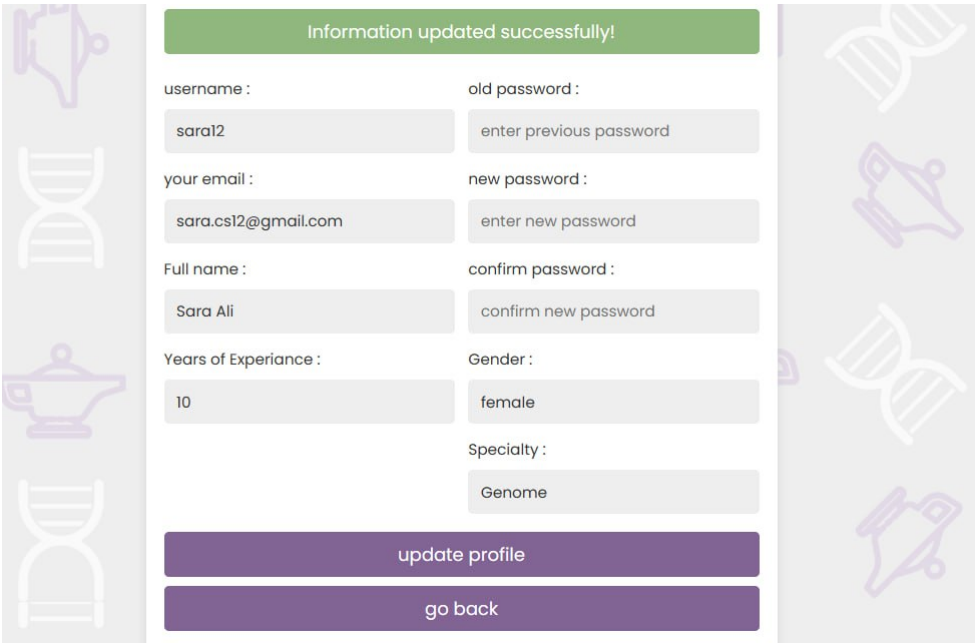


Figure 6.3: Inserting data

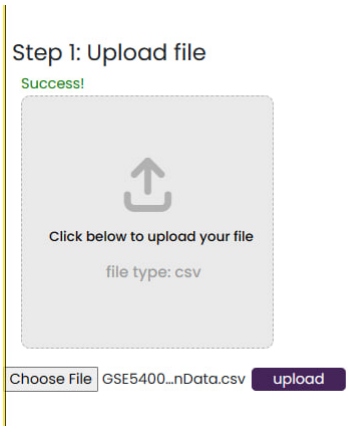


Figure 6.4: File Uploaded Successfully

Result Grid | Filter Rows: | Edit: | Export/Import: | Wrap Cell Content:

file_id	user_id	file_name	date	algorithm	num_DEGs	num_pathways	report
58	7	../uploads/GSE54002_trainData.csv	2023-02-08 21:01:33	Logistic Regression	10	10	7_1974_report.pdf
59	7	../uploads/GSE54002_trainData.csv	2023-02-08 21:16:36	Logistic Regression	15	10	7_3803_report.pdf
60	7	../uploads/GSE54002_trainData.csv	2023-02-08 21:19:57	Logistic Regression	15	10	7_9069_report.pdf
61	7	../uploads/GSE54002_trainData.csv	2023-02-08 21:34:24	Logistic Regression	10	15	7_8351_report.pdf
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL

Figure 6.5: The inserted data in the database

- **Update:**
  - password is updated to check if it will be updated in the database

as shown in figures below

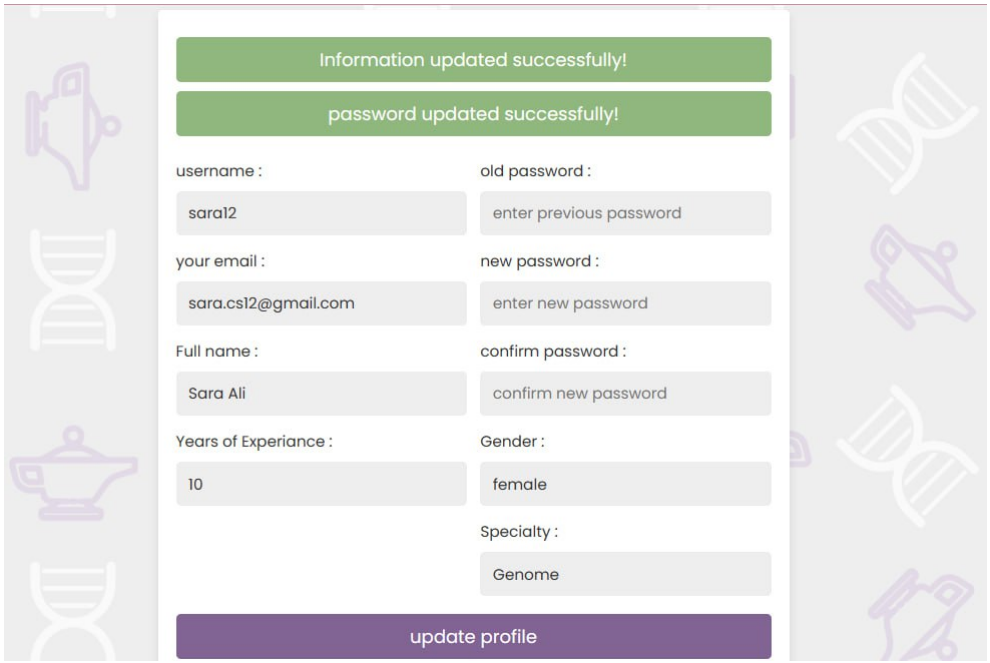


Figure 6.6: Password update

Result Grid | Filter Rows: | Edit: | Export/Import: | Wrap Cell Content:

id	fullname	user_name	user_email	gender	specialty	years	password	pic	date
6	azoz loud	azoz22	azozzz200@gmail.com	NULL	NULL	NULL	202cb962ac59075b964b07152d234b70	default-pic.png	2023-02-07 20:26:46
7	jade ali	jade33	jj@gmail.com	male	html	4	[illegible]	default-pic.png	2023-02-08 17:58:26
8	Sara Ali	sara12	sara.cs12@gmail.com	NULL	NULL	NULL	f7177163c833dff4b38fc8d2872f1ec6	default-pic.png	2023-02-09 14:35:18
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL

old password

Result Grid | Filter Rows: | Edit: | Export/Import: | Wrap Cell Content:

id	fullname	user_name	user_email	gender	specialty	years	password	pic	date
6	azoz loud	azoz22	azozzz200@gmail.com	NULL	NULL	NULL	202cb962ac59075b964b07152d234b70	default-pic.png	2023-02-07 20:26:46
7	jade ali	jade33	jj@gmail.com	male	html	4	[illegible]	default-pic.png	2023-02-08 17:58:26
8	Sara Ali	sara12	sara.cs12@gmail.com	female	Genome	10	6c8349cc7260ae62e3b1396831a8398f	default-pic.png	2023-02-09 14:35:18
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL

new password

Figure 6.7: Updated password in the database

6.6 Usability Testing

In order to assess how user-friendly and clear the interface is, as well as how simple it is for users to accomplish their goals. Usability tests are conducted with highly skilled technical users. The evaluation criteria of the usability evaluation are Participants, the environment of the test, evaluation tasks, and measure analysis.

6.6.1 Participants

pathway genie has been tested on a group of targeted users of AI and genomics professions. The purpose of the test is to judge how simple the tasks are to use and understand. Five females with ages ranging from 30 to 40 years old took the test.

6.6.2 Environment of the Test

The test was conducted online in team viewer, where participants took control over the laptop and started using the interface.

6.6.3 Evaluation Tasks

To confirm the usability of our application Pathway Genie, we evaluated 7 tasks which were: register, then log in to the application, start the analysis process and follow the steps listed on the page, go to the history page, go back and re-upload another file, cancel the analysis before starting, update personal information, then log out. Appendix 3 contains a list of evaluation questions.

1. **Register:** The user creates a new account by entering their email, username and password. User must correctly create an account to log in.

2. **log in:** The user must log in.

3. **Start the analysis process and follow the steps listed on the page:** The user must be able to follow the steps by successfully entering the number of top DEGs, top number of pathway names, selecting an available algorithm and uploading a file.

4. **Go to the history page:** The user must be able to go to the history page via the navbar and see files history.

5. **Go back and re-upload another file:** user must go back to the analysis page from the history page using the back button and then re-follow all steps on the page again.

6. **Cancel the analysis before starting:** Just before starting the analysis, they must be able to press the cancel button and then cancel to get to the home page.

7. **Update personal information:** The user must be able to find the profile page by clicking their username to get a dropdown list with the account page listed. Then they must be able to either update already submitted information or fill in empty fields to update their information.

8. **log out:**

   The user must be able to find the logout page by clicking the logout button on the profile page or via the dropdown list from the username.

6.6.4 Objective Measure Analysis

Three objective performance measures were used to evaluate how well end-users completed the usability test tasks are number of clicks, time required to complete the task, and degree of autonomy are questions that are asked after each test.

1. **Number of clicks** The first applied measurement was counting the number of clicks. Table 6.8 summarizes the five participants' results based on the number of clicks

Task	Evaluation Task	Participant's Number of Clicks	
		Min Number of Clicks	Max Number of Clicks
1	Register	8	10
2	Log in	4	6
3	Start the analysis process and follow the steps listed on the page	9	13
4	Go to the history page	1	2
5	Go back and re-upload another file	5	7
6	Cancel the analysis before starting	1	1
7	Update personal information	13	17
8	log out	3	4

Figure 6.8: Minimum and maximum of the participant's clicks

2. **Duration** The second evaluation metric of the system's usability is the duration taken to complete the task. Table 6.9 summarizes the duration the five participants took to complete the task.

Task	Evaluation Task	Participants duration	
		Min duration of time	Max duration of time
1	Register	22 seconds	28 seconds
2	Log in	11 seconds	13 seconds
3	Start the analysis process and follow the steps listed on the page	1 minute	2 minutes
4	Go to the history page	3 seconds	7 seconds
5	Go back and re-upload another file	14 seconds	25 seconds
6	Cancel the analysis before starting	3 seconds	3 seconds
7	Update personal information	31 seconds	33 seconds
8	log out	6 seconds	9 seconds

Figure 6.9: Minimum and maximum of the participant's duration

3. **Autonomy Level** Autonomy level expressed on a three-level ordinal scale, from 1 = "not able to achieve the task independently", 2 = 'can achieve the task independently, with help' to 3 = "can achieve the task independently, without help". Table 6.10 summarizes the results of the five participants.

	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8
Participant 1	3	3	3	3	3	3	3	3
Participant 2	3	3	2	3	2	3	2	3
Participant 3	3	3	3	3	3	3	1	3
Participant 4	3	3	2	3	2	3	1	3
Participant 5	3	3	2	3	3	3	1	3

Figure 6.10: Autonomy level of participants

6.6.5 Subjective Measure Analysis

Subjective measurement measures people's opinion. In order to apply subjective measurement, some questions at the end of the survey asked participants to rate their satisfaction, ease of use and write their feedback.

1. **Satisfaction** measures the end users' overall satisfaction with the system, based on the users were equally satisfied and moderately satisfied. Figure 6.11 shows the analysis of the participants' satisfaction.

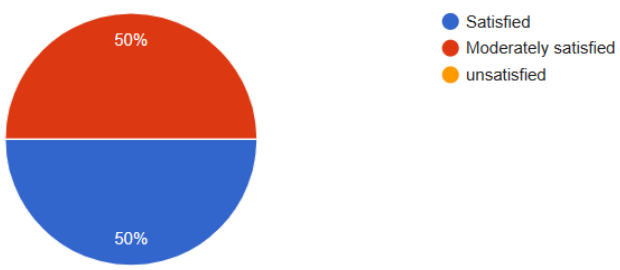


Figure 6.11: Result of the overall satisfaction of users

2. **Easy to use** One of the most important usability metrics is how simple and easy to use the system's interface. 75% of users reported that it is moderately easy since it takes time to analyze and generate report the other 25% of users reported that it was easy to use. Figure 6.12 shows the analysis of the how easy the website is.

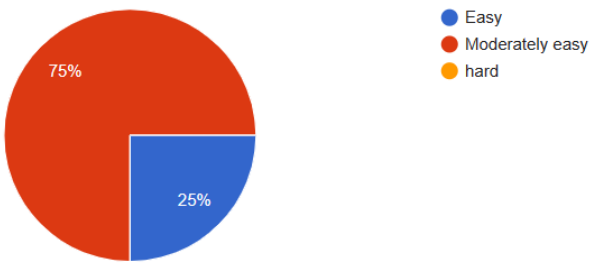


Figure 6.12: Result of how easy the website to use

3. **Feedback** Participants were asked about their opinions and if they had any comments about using the interface. here are some of the feedbacks:

   - The analysis was little bit slow in loading
   - Great work, analysis page needs more information.
   - loved the idea of the project, wish there was a back button in the history page.
   - Easy and clear to use, just the analysis needs to be a bit faster

Test ID	Test Name	End-user input	Expected Output	Actual Output
1	Register	User Full name, User name, User Email, User Password, Confirmed Password	Account created and user can log in	As expected
2	Log in	User Email and User Password	Enter the second home page	As expected
3	Start the analysis process and follow the steps listed on the page	Uploaded File, Number Of DEGS, Number Of Pathway Names, And Algorithm	After the analysis, the successful user enters the final report page.	two of the users had trouble accessing the final report due to network problem
4	Go to the history page	-	Enter History page.	As expected
5	Go back and re-upload another file	Uploaded File, Number Of DEGS, Number Of Pathway Names, And Algorithm	-	As expected
6	Cancel the analysis before starting	-	The user is moved to the home page.	As expected
7	Update personal information	Full Name, Username, Email, Age, Gender, Years Of Experience, Specialty, Old Password, New Password, Confirm Password	Get the message "information updated successfully" or Get the message "password updated successfully"	As expected
8	log out	-	Get Logout pop up	As expected
9	Contact us	Name, Email, Subject, Message	Get verification steps; and verification message.	As expected

Figure 6.13: Pathway genie Application System Testing

6.7 Conclusion

This chapter has given an overview of the various system testing components, including purpose and objectives. system's usability and functionality was tested to make sure the system is working.

CHAPTER 7

CONCLUSION AND FUTURE WORK

7.1 Introduction

This senior project report uses an iterative waterfall methodology and has seven chapters. Introduction, project description, Literature Review, Data collection and requirements, Architecture and design, implementation, and testing phases were all completed. This project includes applied concepts in this project which were from programming, software engineering, Bio-informatics, databases, Artificial Intelligence (AI), Precision medicine, and Data science. in the following sections, the technical and practical challenges learned skills, and future work is discussed.

7.2 Challenges

This project encountered numerous challenges as it went through its many stages of development. a list of the challenges is provided below.

- Since the team members had no experience with R programming language, it was a challenging task.
- It was challenging to balance our data set.
- Since we are computer science students, at first it was a little difficult to understand genomics and the concepts of genes.
- Large CSV files couldn't be uploaded to our website because the old server only supported files of 10 MB or less.

7.3 Learned skills

The gained skills of this project:

- Programming in R language
- Building an AI model
- Integrating AI model with the website
- Web development
- Problem solving

7.4 Future work

- The AI will also recommend medications when it generates a pathway.
- The user of the website will have access to view and download files that have been saved to the history page.
- adding random forest and decision tree algorithms to the website.
- more scientific report

References

(2015). Problems, challenges and promises: perspectives on precision medicine. Available at https://academic.oup.com/bib/article/17/3/494/1744384#35638155.

(2017). Human genomics projects and precision medicine. Available at https://www.nature.com/articles/gt201777#Abs1.

(2017). Precision medicine for all? challenges and opportunities for a precision medicine approach to critical illness - critical care. Available at https://link.springer.com/article/10.1186/s13054-017-1836-5#Sec15.

(2018). U.s. food and drug administration. Available at https://www.fda.gov/medical-devices/in-vitro-diagnostics/precision-medicine.

(2019). Difference between dna and rna. Available at https://byjus.com/biology/difference-between-dna-and-rna/.

(2020). Biological pathways fact sheet. Available at https://www.genome.gov/about-genomics/fact-sheets/Biological-Pathways-Fact-Sheet.

(2020). The why and how of distributed databases. Available at https://fauna.com/blog/the-why-and-how-of-distributed-databases.

(2021a). Database types explained. Available at https://phoenixnap.com/kb/database-types#ftoc-heading-8.

(2021). Transcription vs translation worksheet. Available at https://www.technologynetworks.com/genomics/articles/transcription-vs-translation-worksheet-323080.

(2021b). What is a distributed database? Available at https://phoenixnap.com/kb/distributed-database#ftoc-heading-4.

(2021). What is a gene? Available at https://medlineplus.gov/genetics/understanding/basics/gene/.

(2021). What is waterfall methodology? Available at https://www.ifaai.org/blog/what-is-waterfall-methodology/.

(2022). Classification: Roc curve and auc. Available at https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.

(2022). Coefficients and regression equation for fit binary logistic model. Available at https://support.minitab.com/en-us/minitab/21/help-and-how-to/statistical-modeling/regression/how-to/fit-binary-logistic-model/interpret-the-results/all-statistics-and-graphs/coefficients-and-regression-equation/.

(2022). Gene expression definition, process, regulation, mechanism. Available at https://microbenotes.com/gene-expression/.

(2022). Geo - ncbi. Available at https://www.ncbi.nlm.nih.gov/geo/.

(2022). Understanding precision medicine and its benefits. Available at https://www.softwareadvice.com/resources/precision-medicine/.

(2022). Waterfall model. Available at https://www.istockphoto.com/search/2/image?excludenudity=false&phrase=waterfall%20model%20vector&istockcollection=main%2Cvalue.

(2022). What is a database? Available at https://www.oracle.com/sa/database/what-is-database/.

(2022). What is bioinformatics? Available at https://www.yourgenome.org/facts/what-is-bioinformatics-and-how-do-we-use-it/.

(2022). What is breast cancer? Available at https://www.cdc.gov/cancer/breast/basic_info/what-is-breast-cancer.htm.

(2022). What is data mining? a beginner's guide (2022). Available at https://bootcamp.rutgers.edu/blog/what-is-data-mining/#what.

(2022). What is data science? Available at https://ischoolonline.berkeley.edu/data-science/what-is-data-science-2/.

(2023). Bootstrap · the most popular html, css, and js library in the world. Available at https://getbootstrap.com/.

(2023). Css tutorial. Available at https://www.w3schools.com/css/.

(2023). Decision trees explained with a practical. Available at https://towardsai.net/p/programming/decision-trees-explained-with-a-practical-example-fe47872d3b53.

(2023). Font awesome. Available at https://fontawesome.com/.

(2023). Google fonts. Available at https://fonts.google.com/.

(2023). How simple logistic regression differs from simple linear regression. Available at https://www.graphpad.com/guides/prism/latest/curve-fitting/reg_simple_logistic_and_linear_difference.htm.

(2023). Introduction to html. Available at https://www.w3schools.com/html/html_intro.asp.

(2023). Machine learning random forest algorithm - javatpoint. Available at https://www.javatpoint.com/machine-learning-random-forest-algorithm.

(2023). Mysql. Available at https://www.mysql.com/.

(2023). What is javascript? - learn web development: Mdn. Available at https://developer.mozilla.org/en-US/docs/Learn/JavaScript/First_steps/What_is_JavaScript.

(January 10, 2023). Ide and code editor for software developers and teams. Available at https://visualstudio.microsoft.com/.

Ahmed, H., Hamad, S., Shedeed, H. A., and Saad, A. (2022). Enhanced deep learning model for personalized cancer treatment. *IEEE Access*.

Brown, S. (2021). Machine learning, explained. Available at https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained.

Davis, S. and Meltzer, P. S. (2007). Geoquery: a bridge between the gene expression omnibus (geo) and bioconductor. *Bioinformatics*, 23(14):1846–1847.

Elmasri, R. and Navathe, S. (2011). *Fundamentals of Database Systems*. Addison-Wesley.

Gustriansyah, R., Suhandi, N., and Antony, F. (2019). The design of uml-based sales forecasting application. *Int. J. Recent Technol. Eng*, 7(6):1507–1511.

Hussein, N. A. K. and Al-Sarray, B. (2022). Deep learning and machine learning via a genetic algorithm to classify breast cancer dna data. *Iraqi Journal of Science*, pages 3153–3168.

Joshi, S. (2021). Data virtualization 101: Why businesses need to virtualize data. Available at https://www.g2.com/articles/data-virtualization.

Kamal, M. S., Northcote, A., Chowdhury, L., Dey, N., Crespo, R. G., and Herrera-Viedma, E. (2021). Alzheimer's patient analysis using image and gene expression data and explainable-ai to present associated genes. *IEEE Transactions on Instrumentation and Measurement*, 70:1–7.

Kelleher, J. and Tierney, B. (2018). *Data Science*. The MIT Press Essential Knowledge series. MIT Press.

Khalifa, N. E. M., Taha, M. H. N., Ali, D. E., Slowik, A., and Hassanien, A. E. (2020). Artificial intelligence technique for gene expression by tumor rna-seq data: a novel optimized deep learning approach. *IEEE Access*, 8:22874–22883.

krishnan, M. (2018). Understanding the classification report through sklearn. Available at https://muthu.co/understanding-the-classification-report-in-sklearn/.

Kung, D. (2013). *Object-Oriented Software Engineering: An Agile Unified Methodology*. McGraw-Hill Higher Education.

Maria Luisa Amaral, Galina A Erikson, M. N. S. (2018). Bart: bioinformatics array research tool. Available at https://pubmed.ncbi.nlm.nih.gov/30089462/.

Narkhede, S. (2018). Understanding auc - roc curve. Available at https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47.

Sumathi, S. and Sivanandam, S. (2006). *Introduction to Data Mining and its Applications*. Studies in Computational Intelligence. Springer Berlin Heidelberg.

Suza, W. and Lee, D. (2021). *Genetics, Agriculture, and Biotechnology*. The MIT Press Essential Knowledge series. Iowa State University Digital Press.

Swaminathan, S. (2018). Logistic regression — detailed overview. Available at https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc.

Yamamoto, M., Wakatsuki, T., Hada, A., and Ryo, A. (2001). Use of serial analysis of gene expression (sage) technology. *Journal of immunological methods*, 250(1-2):45–66.

APPENDIX 1

Interview Questions

The following were the questions of the Interview:

1. How long have you worked in the genetics industry?
2. Which tools do you use for gene expression?
3. What unique qualities does the tool have?
4. What difficulties do you usually encounter when performing gene analysis?
5. What more features would you like to see in upcoming products to make the gene analysis process easier?

The result of the interview:

1-How long have you worked in the genetics industry?

The last three years.

2-Which tools do you use for gene expression?

For gene expression, I use R packages such as DESeq and EdgeR.
KEGG: Kyoto Encyclopedia of Genes and Genomes https://www.genome.jp/kegg/
Enrichr https://maayanlab.cloud/Enrichr/help#basics

3-What unique qualities do the tool have?

KEGG is a database resource for understanding high-level functions and utilities of the biological system.

4-What difficulties do you usually encounter when performing gene analysis?

New pathways are unknown. Also, there is some inconsistency when comparing the output from different tools.

5-What features would you like to see in upcoming products to make the gene analysis process easier?

Updated pathways, maybe! And generating a full report.

All best wishes.

•••

Figure 1.1: Interview result

APPENDIX 2

Survey Questions

The Questionnaire questions were the following:

1. How many years have you been in the field of genomics?
2. Which tools do you use for pathway analysis?
3. Does it use Artificial Intelligence?
4. What feature do you use the most in this tool?
5. How long does it take you to analyze a pathway?
6. What additional features would you like to see in Pathway Genie?

The results of the Questionnaire:

Figure 2.1: Questionnaire capture 1

Figure 2.2: Questionnaire capture 2

APPENDIX 3

Usability Testing Survey

Task 1: Register

- A) How many clicks?
- B) How long did it take to finish the task?
- C) How many mistakes were done?
- D) What is the autonomy level?

Task 2: Log in

- A) How many clicks?
- B) How long did it take to finish the task?
- C) How many mistakes were done?
- D) What is the autonomy level?

Task 3: Start the analysis process and follow the steps listed on the page

- A) How many clicks?
- B) How long did it take to finish the task?
- C) How many mistakes were done?
- D) What is the autonomy level?

Task 4: Go to the history page

- A) How many clicks?
- B) How long did it take to finish the task?
- C) How many mistakes were done?
- D) What is the autonomy level?

Task 5: Go back and re-upload another file

- A) How many clicks?
- B) How long did it take to finish the task?
- C) How many mistakes were done?
- D) What is the autonomy level?

Task 6: Cancel the analysis before starting

- A) How many clicks?
- B) How long did it take to finish the task?
- C) How many mistakes were done?
- D) What is the autonomy level?

Task 7: Update personal information

- A) How many clicks?
- B) How long did it take to finish the task?
- C) How many mistakes were done?
- D) What is the autonomy level?

Task 8: log out

- A) How many clicks?
- B) How long did it take to finish the task?
- C) How many mistakes were done?

- D) What is the autonomy level?

  How easy the interface to use?

- A)Easy
- B)Moderately Easy
- C)Hard

  how satisfied are you ?

- A)satisfied
- B)Moderately satisfied
- C)unsatisfied

  Do you have any feedback?