PC3 PROSTATE CANCER CELLS ANALYSIS

Cell Group

Pascal King'oku, Jing Huang, Hazrat Ali, Xiao Su

Dated: May 2020

Course Responsible: Daniel Einarson

Dawit Mengistu

Ola Johansson

Table of contents

List of Figures

List of Tables

INTRODUCTION

Background

In our daily life, cancer is becoming an important issue in the whole world. Cancer affects everyone, like young and old, rich and poor, men, women, and children. It is a huge burden on patients, families, and society. Therefore treatment of Cancer is a thorny question now. Prostate cancer is a common malignant tumor of the urinary system. In the United States, The incidence rate of Prostate cancer has beyond the incidence rate of lung cancer, ranked first, In 2018, There are 164 690 new cases in cancer, and every 5 newly diagnosed cancers involved in one prostate cancer.

However, many deaths caused by cancer can be avoided. By choosing a healthy lifestyle, such as staying away from tobacco and public health measures such as planned immunization to prevent the occurrence of carcinogenic infections, 30% -50% of cancers can be prevented. Other cancers can be discovered, treated, and cured early. Even in the advanced stages of cancer, good palliative treatment can be used to alleviate the suffering of patients.

Figure 1. Sustainable development goals 3 and 11- sustainabledevelopment.un.org

Today, the development of data mining and related technologies brings changes to the cancer field. Data analysis is used in research to detect the presence of cancer cells and different treatment options and find correlations with carcinogens to help fight cancer. Integrate data analysis into epidemiology provides predictions and programs for disease distribution and control[19].

In this project, we are focusing on the analysis of prostate cancer cells. This research area is related to the UN's Sustainable Development Goals 3 and 11. To realize sustainable cities and communities according to the UN's agenda of 2030, health services will definitely play a crucial part.

Figure 2. FCM process showing how events are recorded - source lanl.gov
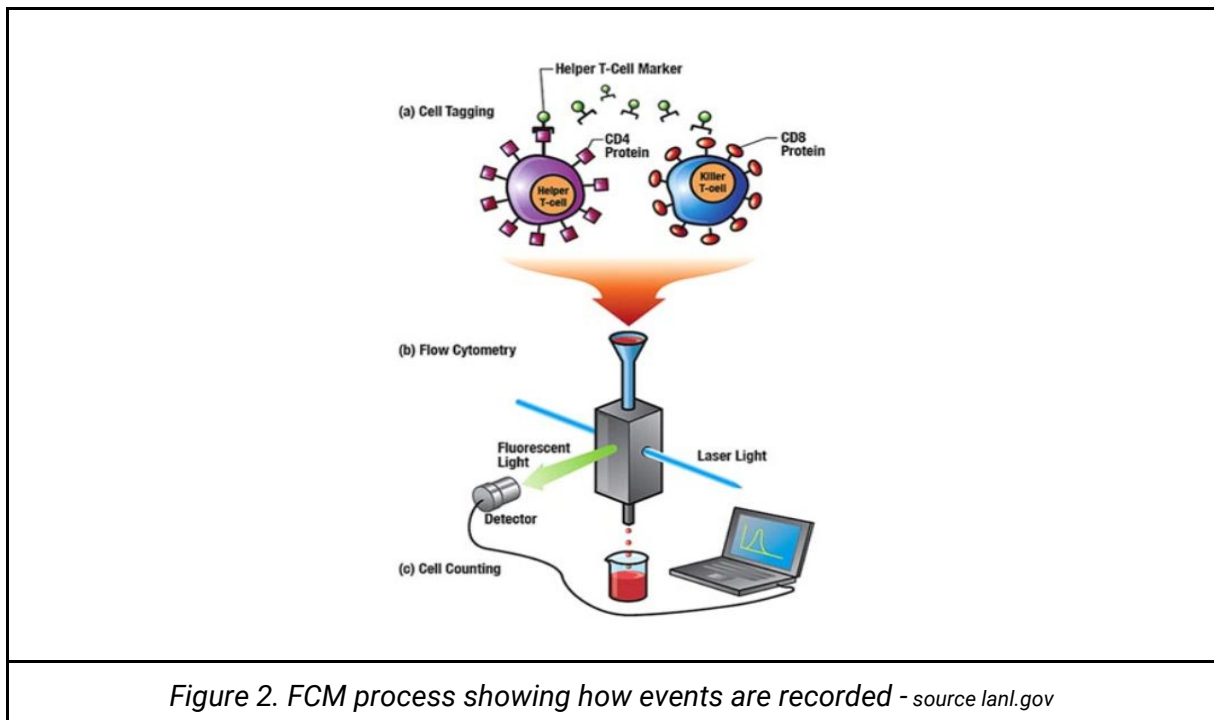
We have used the Dating Mining method to analyze and train several samples that are given by the supervisor. There are some steps In Dating Mining Method, like select data from FCS files, gating, cluster, and so on, and meanwhile, we also need to contribute the website and database for support to show those things.

Purpose

This project is intended to help researchers and medical staff in the prostate cancer field to analyze their datasets and prostate cancer field to analyze their datasets and get high precise results from the system. This application can also be extended to analyze cell data to support research in other medical fields.

For the prostate cancer cells analysis group, on the one hand, we will manage the knowledge of data mining, on the other hand, we will learn how to achieve projects on group work.

Motivation

Cancer is a group of diseases involving abnormal cell growth with the potential to invade or spread to other parts of the body. These contrast with benign tumors, which do not spread. Possible signs and symptoms include a lump, abnormal bleeding, prolonged cough, unexplained weight loss, and a change in bowel movements. There will be approximately 18.1 million new cancer cases and 9.6 million cancer deaths worldwide in 2018.

Based on the information above, it is necessary to develop techniques in the cancer field, and till now Data mining has achieved many results in the research of cancer. The early warning of cancer diseases often requires more accurate real-time and early warning tools, and the establishment of early warning models of cancer-based on data mining can help improve the early diagnosis of cancer, as while as Early warning and monitoring also help cancer institutions to take preventive and control measures to reduce the occurrence of cancer deterioration and complications.

For early warning of disease, first, collection the disease-related index data or risk factors, and then establish a model to discover the relationship between the pathogenesis and the hidden condition the data

An assisted cancer diagnosis is not only large in data but also intricate and interrelated. Analysis of a large number of cancer data, mining valuable diagnostic rules, will provide a reference for cancer diagnosis.

Cancer data mining has shown good development and the huge value of the application and will bring more support to many fields such as disease research, clinical and management decision-making, personalization of cancer services, and image recognition.

At the beginning of the course, we have learned some knowledge about sustainable development and also introduced the situation in the cancer fields and data mining fields, furthermore, we have researched this project, like a cancer cell, data clustering, training model. and so on. For the group work, we have built the group from the supervisor and make the plan for the project and allocate tasks as early as we can, have two meetings twice a week so that we can contribute the project in a good communication way.

This project was created as a collaboration between the computer science department and the biology department(Biomedicinska analytikerprogrammet), both in Kristianstad university, to find ways to use machine learning to ease the process of analyzing prostate cancer cells to enable the quick and easy detection of prostate cancer in collected samples.

Vision

This document describes an approach to detect the presence of PC3 prostate cancer cells by employing machine learning to analyze cell cycle stages. It is based on research at Kristianstad University. *[1]*

In the documentation, there are several components that have contributed to it, in the first part which is mainly to describe the background, purpose, motivation of the project, it will give the whole perspective to the reader. For the planning part, Analysis, WBS, SWOT analysis, and time plan involved before implementing the system, the Requirement is a necessary part of the documentation which is described by all staff what should be done. a more relevant component is supplementary requirements, risk, test, design. The reference part and Appendix part will appendix in the end.

We have contributed the artifact not only to follow templates from Canvas, But also it has been done in some new ideas which we think is important.

Literature Review

In current medical and biological problems, PC3 cancer cells are a very important issue, the PC3 is a human prostate cancer cell line used in prostate cancer research and drug development.[12] PC3 cells are useful in investigating biochemical changes in advanced prostate cancer cells and in assessing their response to chemotherapeutic agents. PC3 cells are also used to study viral infection in mammalian cells that exhibit an immune response. There is a lot of literature on how to detect and analyze PC3 cancer cells.

Flow cytometry (FCM) is a widely used tool in the field of immunology because it can measure cell properties at cell level resolution in a high throughput manner[1].

The process of measuring the event in flow cytometry is shown in figure 2. The antibodies are first dyed with fluorochromes that emit specific wavelengths of light when exposed to laser light. These cells are then passed through a capillary tube where they may be excited by laser light resulting in scattering of light in different wavelengths depending on their size, granularity, and fluorochromes. These wavelengths are then measured by using a detector that records the data called events and feeds it to a computer system for further storage and processing.

Some research work [13] shows literature about pure biology and medicine, but only a little statistics on the evolution time and generation area of cancer cells are not accurately analyzed by computer science algorithms. In addition, literature [14] and [15], Compared with [13], the description is more detailed, and the physical picture of the microscope is also used to verify. However, They still do not use computer science-related technologies and analytical algorithms to accurately analyze, they only rely on the results of certain samples for discussion. Because their samples are cultivated in the most suitable medium for survival, they are not universal, and they are not suitable for humans. Moreover, the data they sampled is relatively small, which does not meet the standards of big data statistics, and they rely on manual statistics and are not accurate. Therefore, machine learning algorithms should be used like we do. Substituting big data files directly can get accurate analysis results, which is both accurate and convenient.

These research articles use machine learning algorithms. such as [16] and [17], [16] which uses the flow cytometry profile to represent Annexin V-FITC staining in x-axis and PI in the y-axis, in addition. They also use microscope physical maps and a lot of medical knowledge to verify the conclusions, But these are only theoretical basis, and there is not mch experimental data, which cannot fully verify the accuracy of the conclusion. Their article focuses more on the research of medicine and biology. For machine learning algorithms and models, it only briefly explained the method and did not introduce it specifically. They need to do a lot of experiments and get a lot of data, and then visualize these data to find a specific classification. This requires the use of intelligent machine learning algorithms, and details the principles and operations of these algorithm models in order to provide more experts to see to improve and modify. Although these articles use some machine learning models, they are not intelligent, and the results obtained do not explain the problem, and do not explain the

principles of these machine learning models, so that we do not know how they are analyzed in the experimental content. [17] is also strong literature for medicine and biology. They introduced the knowledge of protein cell arrangement in medicine very deeply. But in the same way, in the field of computer science and technology, they also simply used the "clustering" algorithm. It does not mention too much algorithm content and a computer environment.

In addition, There are also several very strong research papers, for example [18] and [19], the [19] is a The introduction and description of FCS analysis methods that have detailed maps and related mathematical formulas. Very helpful for our algorithm model. [18] It is a perfect article, not only using knowledge in the fields of medicine and biology but also using machine learning algorithm models of computer science. And a lot of charts and graphs are calculated by the algorithm model. In addition, they use many machine learning algorithms for theoretical results and many experiments to verify them. It's Very worthy of our reference and learning. But the difference between us is their research is about normal cell culture, not about PC3 cancer cells. In addition, their shortcomings relative to us is that we have a visual front-end interface, ordinary users can easily and simply upload related files and get results. But they are just a complex algorithm model, requiring professionals to input data and get relevant results.

PLANNING

In this chapter we describe requirements, risks, tests, and then prepare a time plan and work breakdown structure to achieve these requirements.

Requirements

ITEMS	PRIORITY
R1. Allow users to securely log in	Medium
R2. Upload FCS files to database	High
R3. Notify user using visual feedback	Low
R4. Filter FCS files using metadata	High
R5. Analyze uploaded FCS files	High
R6. Allow user to select gates	Medium
R7. Display results in the webpage	Medium
R8. The FCS files should be selected from database	Medium
R9. Allow selection of columns to be plotted and transformation function	Medium
R10. Implement visual charts	Medium
R11. Train the prediction model	Medium
R12. Create separate Upload and Analysis pages	Low

Table 1: Requirements of the project

R1. Only registered users with valid login credentials should be allowed to access the application. Users need to type their username and password on the login page.

R2. The application should allow users to upload FCS files they want to analyze. Duplicate uploads should be prevented/overwritten.

R3. Users should be notified about the status of tasks. E.g. If files uploaded successfully, network problems, incorrect login credentials etc.

R4. Users should be allowed to select specific FCS files that match certain criteria e.g. location, specific date etc.

R5. The application should allow uploaded FCS files to be analyzed. File analysis should be sequential to ensure ease-of-use.

R6. Users will be able to enter coordinates of the specific gates they want to focus on for their analysis.

R7. All analysis results should be displayed sequentially on the analysis webpage to avoid confusion.

R8. Users should be able to select FCS files available for analysis from a dropdown box

R9. The application should be flexible to allow the selection of columns to be plotted on the X and Y axes. It should also allow for the selection of a desired transformation function (Hlog, Glog)

R10. Data from the FCS file should be plotted as a scatter plot

R11. The system should be able to predict the classification of an uploaded file. Machine learning will be used to determine how to classify samples using the trained model.

R12 The application should allow separation of functions such that different users can upload files and different users can analyze files.

Supplementary Requirements

Supplementary requirements	Priority
SR1. Set up a group meeting schedule	High
SR2. Set up a project management tool	Medium
SR3. Set up a group communication channel	Medium
SR4. Setup a repository	Medium
SR5. Setup online hosting	Medium
SR6. Hardware - high-performance CPU needed	High

Table 2: Supplementary Requirements involved in the project

SR1: Create a group meeting schedule to ensure consistent progress and accountability.
    We used teams, zoom, hangouts, google calendar for meetings.

SR2: Set up a project management tool to allocate and monitors the progress of tasks on an individual level
    We used trello online scrum/kanban board.

SR3: Create a group for the team to use to communicate about all matters regarding the project.

    We use Slack online collaboration tool.

SR4: Create an online repository to allow the project team to work on the same source code and to backup it up.

    We use github to host the code and for version control we use git.

SR5: The application should be hosted online so that people can access it from any location.

    We use docker base solution which is the most favourable architecture and we can easily host it on Digital ocean and AWS.

SR6: In the project, the analysis FCS files is needed, it will be an exhaust CPU.

    We used high computing CPS from AWS.

Non-Functional Requirements

Non-Functional requirements	Priority
NFR1. Stability	High
NFR2. Sustainability	High
NFR3. Fault tolerance	High
NFR4. Disaster recovery	Medium
NFR5. Resource consumption under a given load	Low

Table 2: Non-Functional Requirements involved in the project

NFR1: Refers to the probability that the software will make mistakes during the continuous operation time, for example, it will make mistakes once or several times a day.

NFR2: Does the project follow a few of the 17 points in the SDG. And try not to violate these 17 points

NFR3: It is the property that enables a system to continue operating properly in the event of the failure of (or one or more faults within) some of its components. If its operating quality decreases at all, the decrease is proportional to the severity of the failure, as compared to a naively designed system, in which even a small failure can cause total breakdown. Fault tolerance is particularly sought after in high-availability or life-critical systems. The ability of maintaining functionality when portions of a system break down is referred to as graceful degradation.

NFR4: When encountering man-made or some naturally occurring failures, can the project be restored to its own appearance.

NFR5: When processing a large number of files, resources will not be completely consumed and no stuck will appear

Risks

Risk items	Priority
RK1. Unclear requirements of the project	High
RK2. Project delays	Low
RK3. Inadequate subject-matter experience	Low
RK4. Limited technical expertise	Low
RK5. Unknown risks	Low
RK6. Compromising in design	Low
RK7. Communication breakdown	Low

Table 3: Risks in the project

RK1. Some requirements may be unclear and require further clarification. This will be indicated by our inability to clearly describe what we're supposed to achieve. The solution will be to discuss with group and teachers is needed

RK2. A time schedule has been created to track the progress of the project and each task should be completed before its deadline. The solutions to this risk are to implement as fast as possible and to assign enough time to deal with the problems that arise.

RK3. The team has varying levels of experience in different fields, like Flow cytometry and prostate cancer working in groups, testing and so on. The solution is consistent communication with project stakeholders and further research on the topic.

RK4. This is in regards to some of the technologies involved in the project, like react, python, Postgres. Some requirements will be challenging to achieve because technical expertise of developers vary. The solution is to try to learn more and help each other in the group.

RK5. This caters for the unknown risks that cannot be controlled or estimated. As the field of software development becomes more and more complex, the risks associated with it have intensified. There may also be external factors that may affect the project negatively. The solution is to quickly plan to get back on track as fast as possible.

RK6. In order to get started with the 'real' tasks as soon as possible, developers tend to give up the design process. This is a waste of programming hours, as designing is the most critical part of software development. The solution is to come up with clear designs before moving on to start development.

RK7. Communication is most important in the whole project, the people convey information through Communication, describe correctly and nice comprehension is necessary. The solution is to establish clear communication channels and to come up with better ways to describe tasks/challenges if the team doesn't understand.

Design

Design items	Priority
D1. Login Page	Desirable
D2. Register Page	Optional
D3. FCS Upload Page	Essential
D4. FCS Analysis Page	Essential

Table 4: Design in the project

D1. Two input fields will be filled in, one is "username", and the other one is "password".

D2. Three input fields will be used to create a user account: Name, Username & Password.

D3. Three input fields and a file dropzone area will be used to accept manually entered metadata and selected files.

D4. This will contain a number of sections. The first section will allow users to select files from a dropdown box. The FCS files displayed in this box should allow filtration by date, location & type. The second section should allow users to select the columns they want to plot on the X & Y axes and which function they want to apply to the values. Third section will contain the resulting plot and have input fields for the user to select the gate coordinates they want to focus on.

Tests

Tests	Passed/Failed
T1. Whether the webpage works properly	pass
T2. Can the FCS file be uploaded normally?	pass
T3. Can FCS files be sent to the database of the background server?	pass
T4. Does the application allow the input of gating coordinates?	pass
T5. Can the FCS file be analyzed？	pass
T6.Whether to transfer the analysis results (visualize plot) to the front-end web page	pass

Table 5: Tests in the project

T1. User login interface and It should upload multiple files, show the user if the upload was successful as well as the select address (location),   and visualize preliminary results.

T2. The FSC file should be uploaded to the backend without any data loss. And upload multiple files.

**T3.** FCS files should be sent to MongoDB and saved.

T4. the data of the FCS file is not needed at all. the backend should filter some data and keep only one of them.

T5. The background needs to analyze how many cancer cells and normal cells are in the FCS, and present a visual plot.

T6. The analysis results in the backend should be sent to the frontend and show it on the interface.
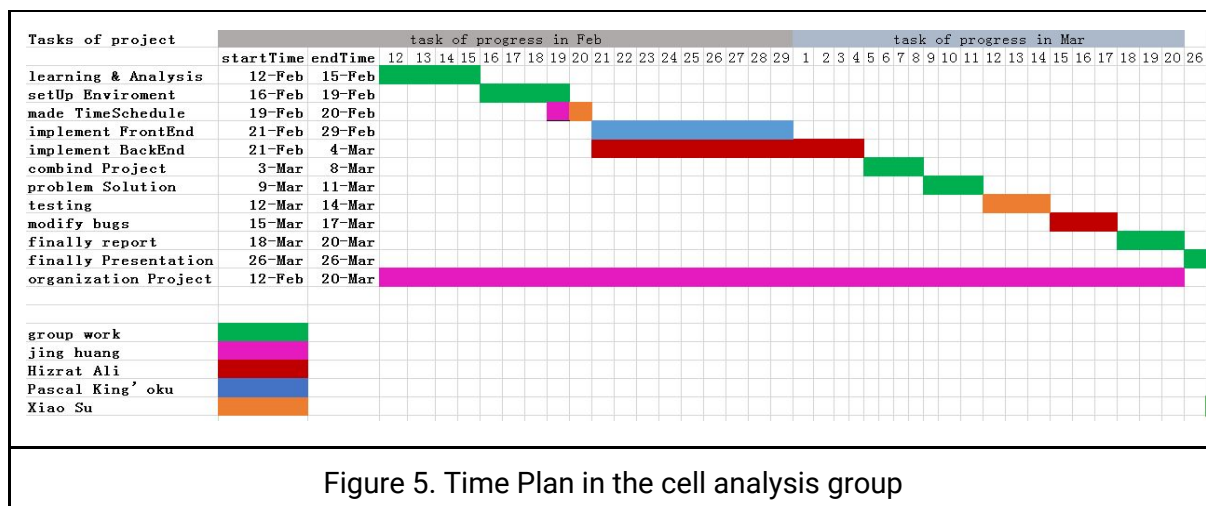
Work Breakdown structure

A work breakdown structure (WBS) in project management and systems engineering, is a deliverable-oriented breakdown of a project into smaller components[7]. Based on the requirement of the project, we have designed the work breakdown structure for the project, following the WBS, we can see the whole phase of the project, there are five main steps to contribute, those study on the project, project analysis and planning, front-end development, back-end development, and Testing and Modify, as while as we also can see the whole procedure for coding implement between front-end and back-end.

Figure 4. Work Breakdown Structure of the cell analysis group
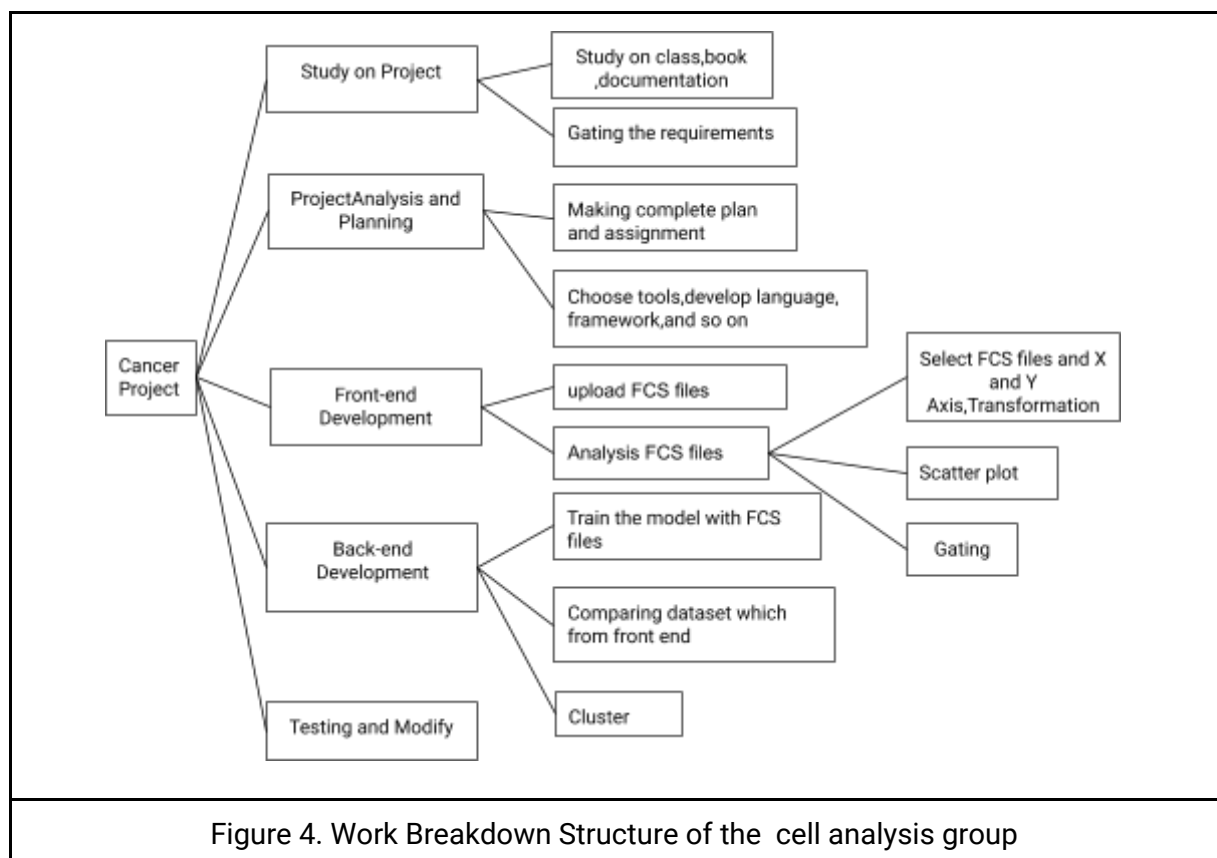
Time plan

The initial of the project, after making the plan, we have designed the time plan for the process of development, the season to make the time plan, in the one hand, The increasing importance of timely completion[8], on the other hand, it is more clear about how many tasks involved in the plan, tasks for everyone should be clear.

Tasks of project	startTime	endTime	task of progress in Feb	task of progress in Mar
learning & Analysis	12-Feb	15-Feb		
setUp Enviroment	16-Feb	19-Feb		
made TimeSchedule	19-Feb	20-Feb		
implement FrontEnd	21-Feb	29-Feb		
implement BackEnd	21-Feb	4-Mar		
combind Project	3-Mar	8-Mar		
problem Solution	9-Mar	11-Mar		
testing	12-Mar	14-Mar		
modify bugs	15-Mar	17-Mar		
finally report	18-Mar	20-Mar		
finally Presentation	26-Mar	26-Mar		
organization Project	12-Feb	20-Mar		
group work				
jing huang				
Hizrat Ali				
Pascal King'oku				
Xiao Su				

Figure 5. Time Plan in the cell analysis group

System Analysis

Based on our Data Mining background, The data mining process includes six steps: data selection, data cleaning, data assignment, data conversion, data mining, and result evaluation[20].
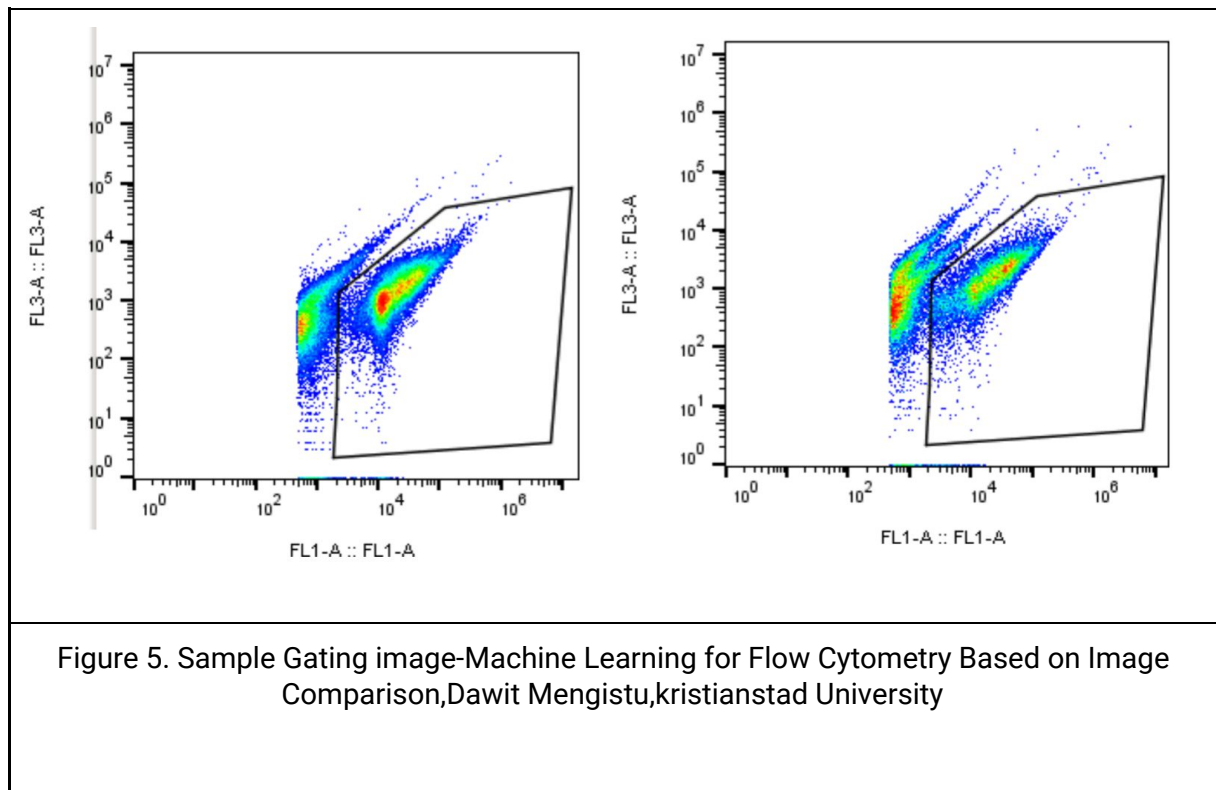
Items	Description
Data selection	Data selection includes the selection of data sources, data types, characteristic variables, etc. Among them, the selection of characteristic variables is crucial, and many analytical modeling explorations often start with hundreds or even more variables，for data clearing,
Data cleaning	Data cleaning is to solve data quality problems, and the other is to make data more suitable for mining
Data conversion	Data conversion is to convert or merge data to data processing form a description
Data mining	Data mining is the extraction from large amounts of data
Result evaluation	result evaluation is evaluated a result from the module.

Table 6: Data mining process - common

algorithms for data mining in medical big data research

For each FCS file, the raw data file is in .fcs format, and the raw data contain the information for each cancer sample, in each FCS file, each column(attributes) that could give information about the index, if intend comparing FCS file sample, selecting several columns in all attributes is necessary, but what comparison Approach should contribute here? following the project, if only two attributes are considered, it will be possible to produce a 2D image for each sample on the selected attributes, and then these 2D images are to be compared.[5] and then, we will Gate the image. figure-3 is a Gating image shown below，the Gating means

to filter out the data which is not needed. and then define a sample based on the training module. The last step results, The results gave the expert new insights into the necessary features and the classification knowledge and show the feasibility of an automated inspection system.[4].In the system, we have classification by K-means clustering, K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.[6]

Figure 5. Sample Gating image-Machine Learning for Flow Cytometry Based on Image Comparison,Dawit Mengistu,kristianstad University

DESIGN

Software Development Life Cycle

Software development deals consist of a lot of activities, for example, business requirements, deployment, and performance, etc. In particular software development, the life cycle consists of different phases which are shown in Figure 6 below.



Figure 6. Software Development Life Cycle
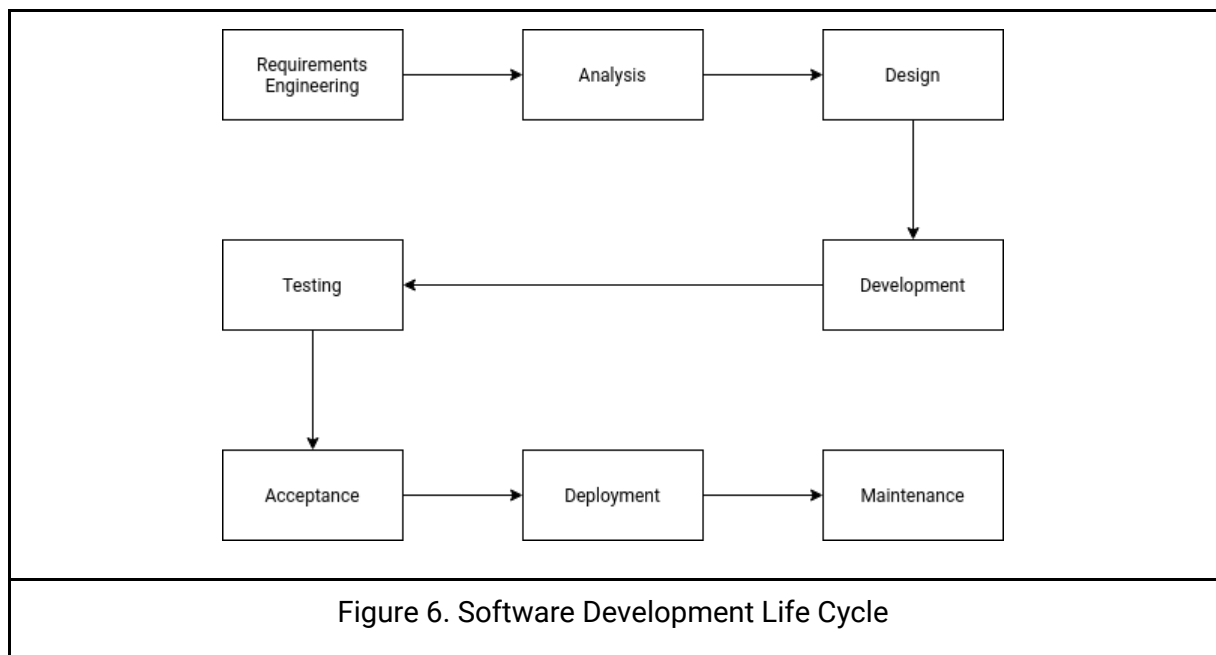
In the requirement phase, we take the business information/ requirements from the stakeholder and this process is also termed as requirements elicitation. In the Analysis phase, we convert the requirements from the previous phase into a more meaningful form that software engineers understand. In the design phase, the software architecture is designed and documented.

The software architecture defines Architectural patterns, Quality attributes, messaging mechanisms and APIs for the software system. Therefore it is more than just the high-level structure of the overall system. The architecture is the input to the development phase which the developers use to implement the system.

Software Architecture

It is the high-level structure of the software system. It can be better defined as how the different components are organized and assembled, how they communicate with each other and defining the constraints that govern the whole system.
From the definition above we can see that there are three major parts as shown in Figure 7 below[11].

Figure 7.Software Architecture
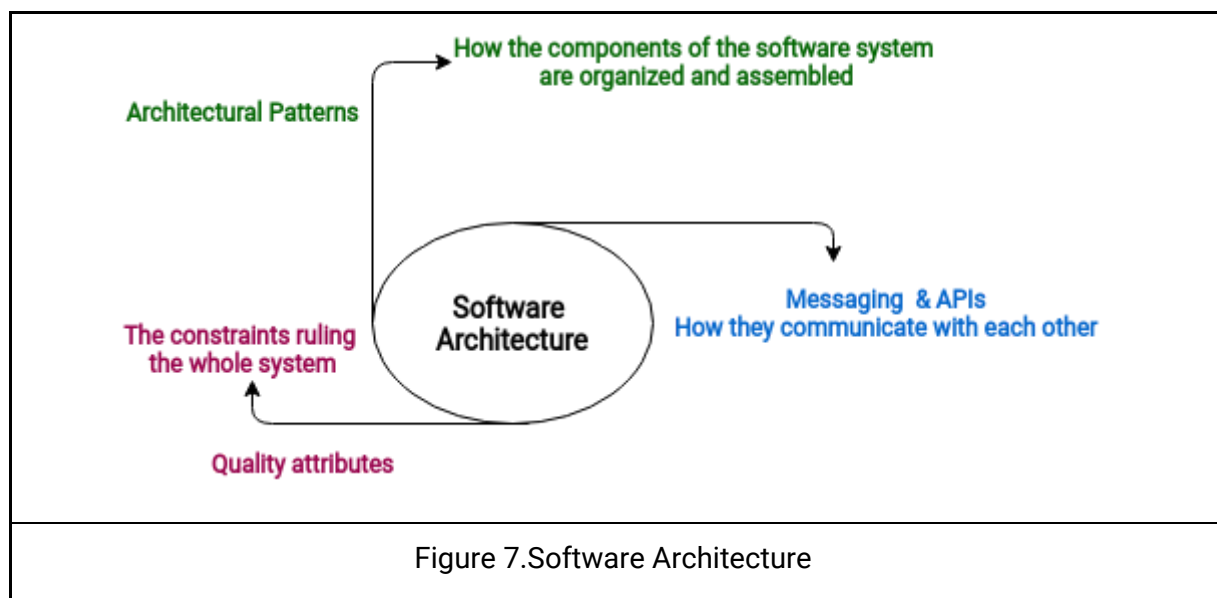
The architecture pattern is the overall structure upon which the whole system is developed. Examples are Layered architecture, microservice architecture, service-oriented architecture, event-driven architecture, etc.
Architectural patterns can be classified into two levels.
Level 1: Monolithic, Service-based, Distributed
Level 2: Layered, Microservices, Event-Driven

Layered architecture is the most commonly used and it contains several layers based on the separation of concerns. Different layers represent the major components of the system.

Microservice architecture is characterized by independently deployable units of components that make the whole system by mutual coordination (docker containers).

An event-driven architecture pattern is a kind of distributed architecture and is highly scalable. It consists of multiple distributed computing that coordinates with each other using Mediator or Broker topology.

For this software system, we have chosen a microservice architecture pattern as it provides more flexibility and scalability for small teams.

System Architecture

The system has been divided into multiple microservices. The front Gateway is implemented in Django Python Framework. Authentication Server has been written in Node JS and Express.

The layers of the system architecture are given in Figure 8.


Figure 8: System Architecture
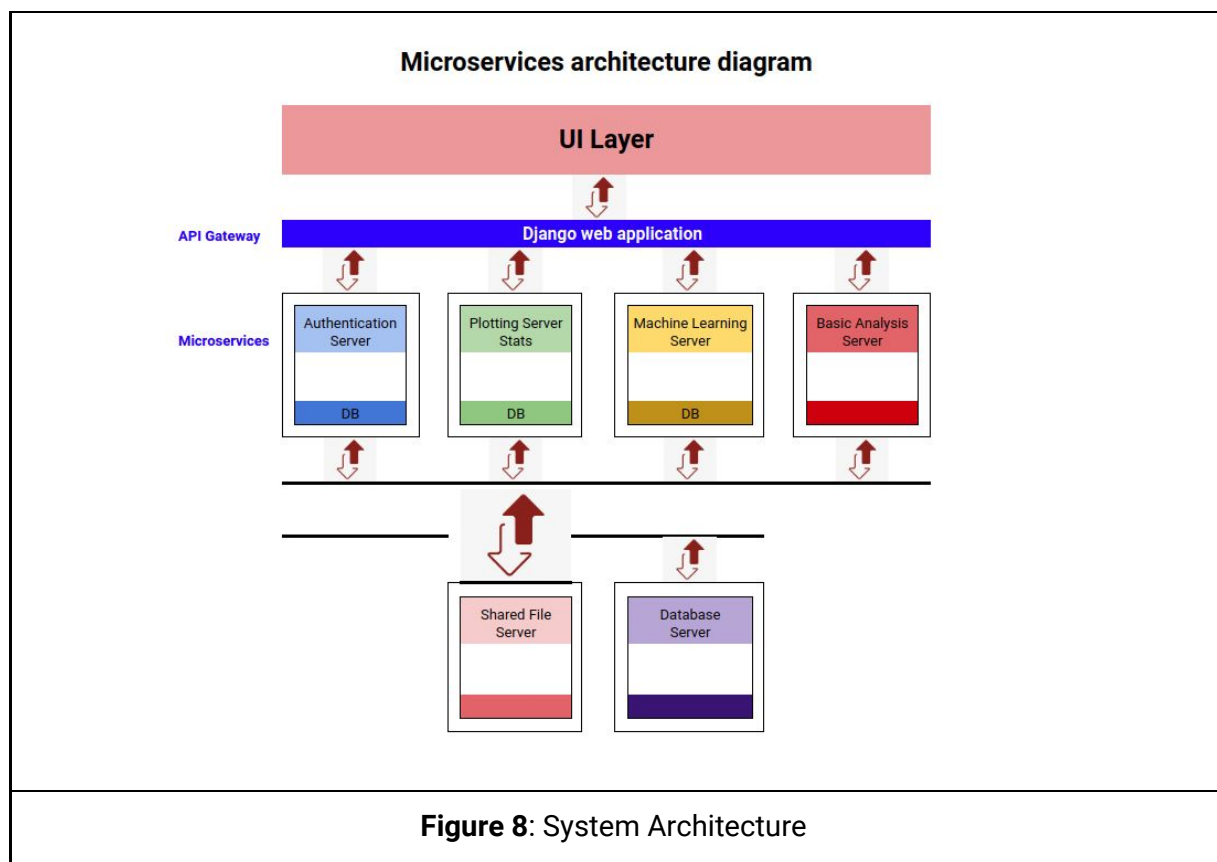
Authentication and Authorization

Authentication specifies who you are while authorization specifies what you can do. Although Microservices architecture has provided many advantages but at the same it is also challenging how to implement a secure, flexible and efficient authentication and authorization system.

We studied different approaches for different architectures to find out the best one which is more suitable for microservices based architectures. For example in case of a monolithic application the whole application is one single process and therefore session based authentication works fine. However in case of microservices based architecture the system is distributed among different microservices. The challenge is how to determine if the user is authorized for a service.

Token based authentication with API Gateway

When the user logins to the system the server sends a token which is stored in a cookie on the browser side.

Figure 9: Token based authentication system with Gateway

Source: https://panlw.github.io/15403881797785.html
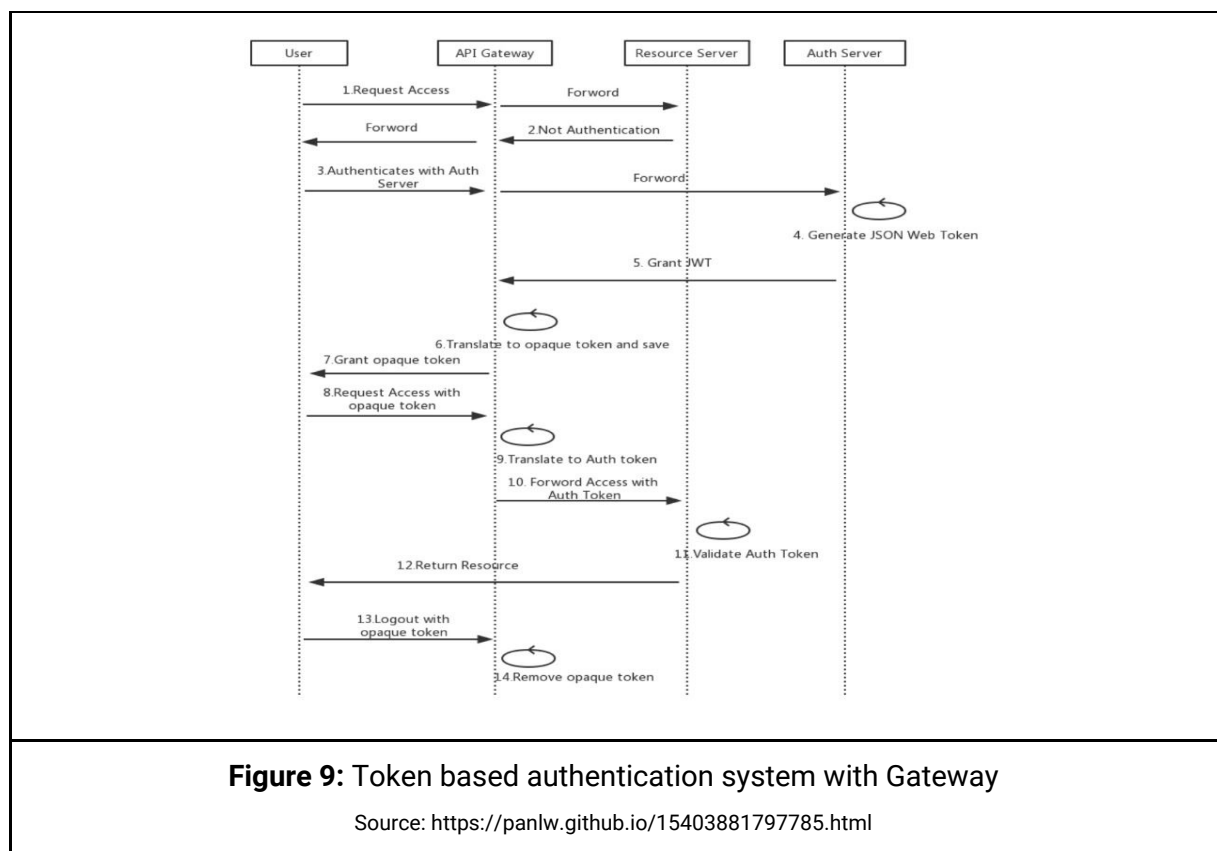
The subsequent requests to the server uses this token sent with each request. This makes it more secure than sending a user password over the wire each time.

To use the token based authentication, we used Json Web Token (JWT) which has a specific format of header, payload and signature and specified in RFC 7519.

We have used a Gateway which handles requests and gets services from other microservices. This way the Gateway is hiding other services in the network which makes it more secure. To authenticate the user, the Gateway takes credentials from the user and sends it to the authentication server (node server), which sends a token(on success) to the Gateway. The Gateway in turn sends the token to the browser which stores it in the cookie and sends it to Gateway for further requests. The Gateway will authenticate the token with the authentication server for each service that requires authentication. This way the authentication system is handled centralized. This process is better explained in figure 9.

IMPLEMENTATION

In this chapter we describe the implementation of the application in two sections i.e Frontend and Backend.

Frontend Development

Access

The web application can be accessed via the URL below:

1. http://hkrfcs.herokuapp.com
2. http://pc3.servebeer.com/
3. http://my-cell.ddns.net/

Log In

This page allows the user to input an email address and password to access the application.
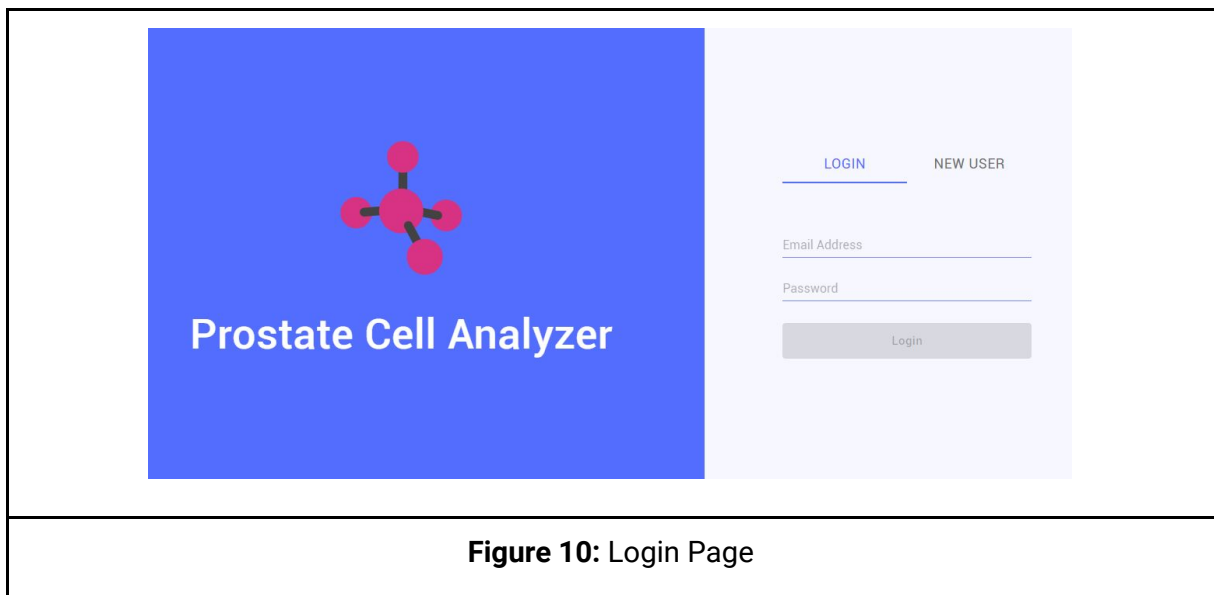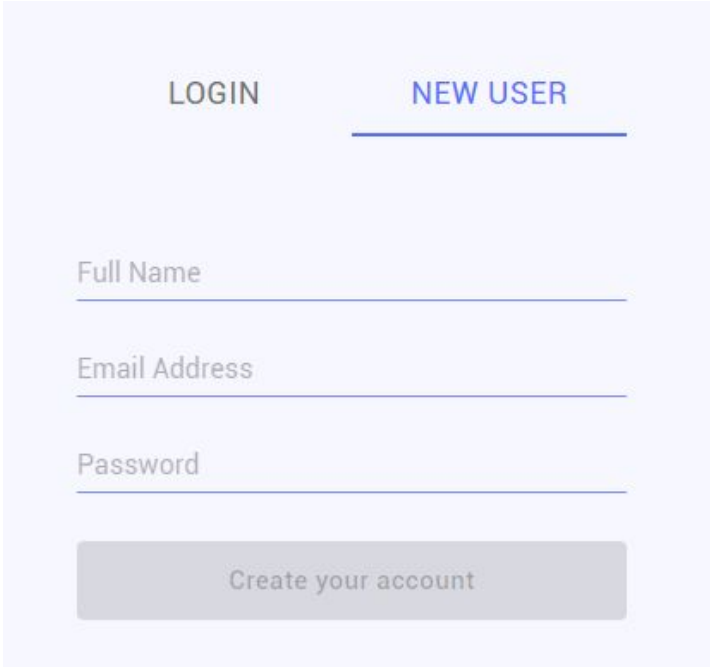

Figure 10: Login Page

Registration

Users who do not have accounts can access the application by creating a user account. They need to provide their name, an email address and a password they'll use.

Figure 11: Register Page

File Upload Screen

The file upload screen allows the user to upload FCS files that will be analyzed. A user can upload as many files as required simultaneously.

Figure 12: File upload screen

File Analysis Screen

This screen allows the user to select an FCS file to analyze. It has 4 steps which are structured sequentially to ensure the user doesn't miss a step. The steps are:

Step 1: Select FCS File

This section contains the list of all files uploaded to the database. All Available files are loaded by default to a dropdown.

The files to be loaded can be filtered by date, type or location. If filters are selected, the button "apply" should be clicked to return only FCS files that meet the specified criteria. The button "Reset" removes all filters applied.

Figure 13: FCS file selection & Filters

Step 2: Select columns to use and transformation function

This section is opened immediately after a user selects an FCS file to analyze. The columns to use for X and Y axes are selected. It is also possible to select a transformation function to be applied to the values if needed.

After desired options are selected, the "Plot" button is clicked to generate a scatter plot.

Figure 14: Column & Transformation selection
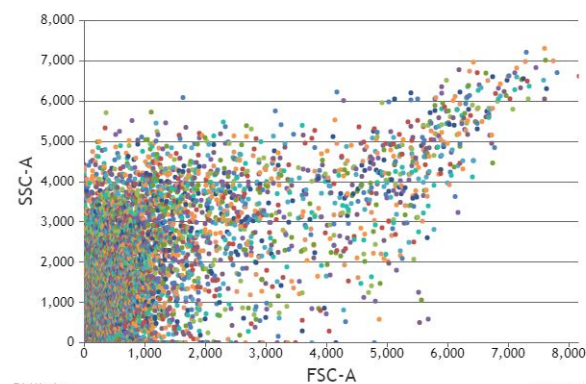
Step 3: A scatter plot is generated from the selected parameters

After the user selects the columns they want to plot, the data is displayed in a scatter plot as shown below:

Figure 15: Scatter Plot

It is possible for the user to zoom in on specific sections of the scatter plot by clicking on the graph and dragging.

Figure 16: Zoomed scatter plot
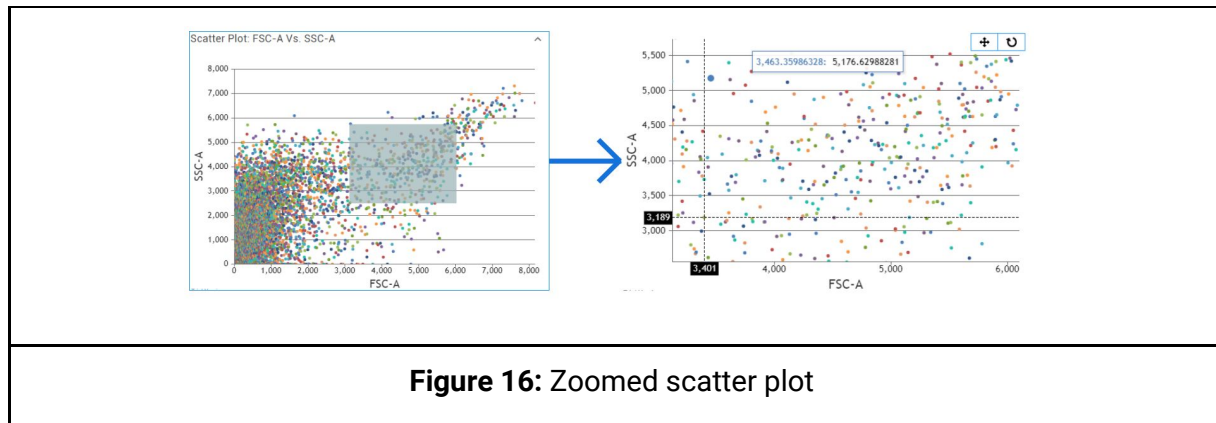
Step 4: Generate Clusters

The user then selects the gate coordinates and binwidth to be used after which they click on "Generate Clusters". Four charts are then presented:

1. A heatmap of the gated region is then presented to the user.
2. A line graph derived of all selected FCS files for comparison with the selected FCS file having a significantly thicker line
3. A scatter plot with dimensions reduced using SVD
4. A scatter plot with dimensions reduced using NMDS.

Figure 17: Heatmap from FCS sample & Line Chart comparison of selected FCS Files

After the gates have been created for all the files, Multidimensional Scaling and Singular Value Decomposition are applied to the data to allow plotting on 2-dimens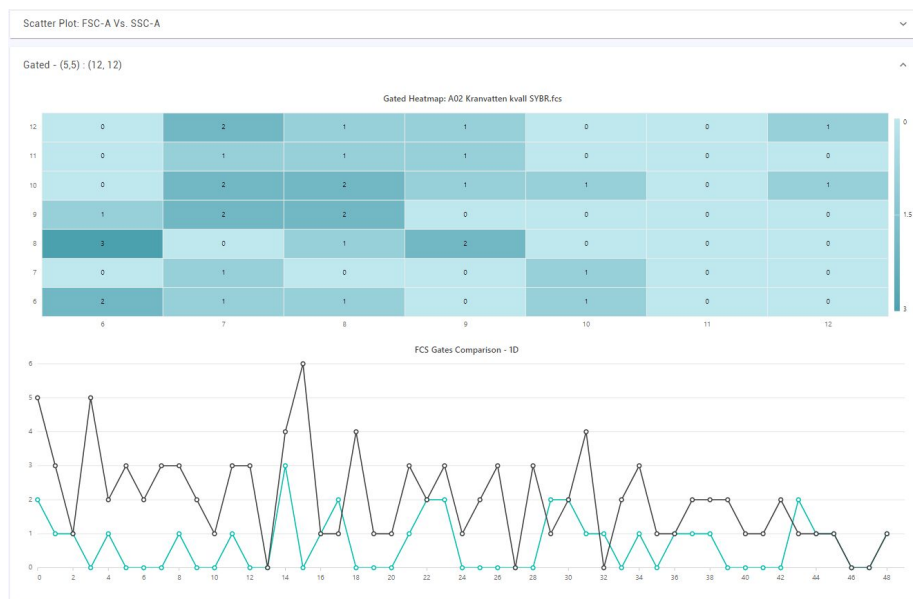ions. The clusters are generated using K-Means The application allows the researcher to recluster based on how many clusters they require.
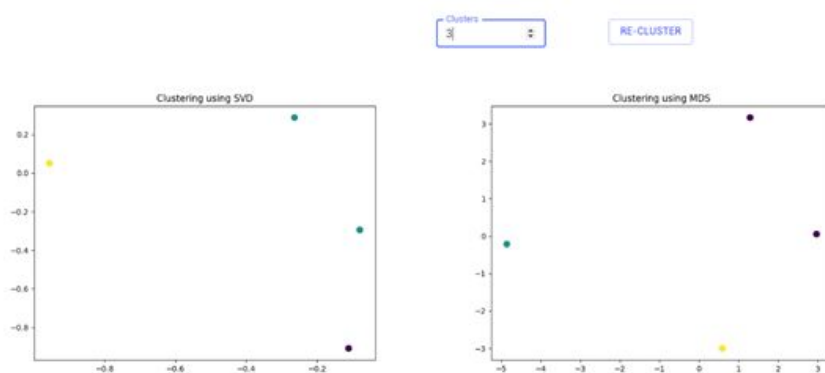

Figure 18: Clustering using SVD and NMDS

Backend Development

We have used docker based microservices. Each docker container is responsible for a specific functionality (SOLID principle) in the project. The api is developed in Django which connects to other microservices like PostgreSQL, Node Authentication Server, Basic Analysis, Plotting Server, Machine Learning Server.

To make the deployment really easy and transparent we used docker-compose tool which holds the information about all images for the services.

REST APi Endpoints

The endpoints of the Django API server are given below.

1. **/api/login : POST**

   This endpoint is used for login in the user to the authentication server. The Gateway will serve as a proxy in this case.

2. **/api/logout : GET**

   This is used to logout the user.

3. **/api/register : POST**

   This is used to register the users. The data is saved in the PostgreSQL server.

4. **/api/forgotpassword : POST**

   Used to reset password. (not implemented)

5. **/api/upload : POST**

   This endpoint is used to upload the FCS file by the logged in user.

6. **/api/basic : GET**

   It returns a json file showing the basic analysis about the FCS File.

7. **/api/plotting?ch1=channel1&ch2=channel2 : GET**

   It is used to get a json file of plot images.

8. **/api/machine-learning?ch1=channel1&ch2=channel2 : GET**

   This endpoint provides a json file about the plotted images from the machine learning algorithms.
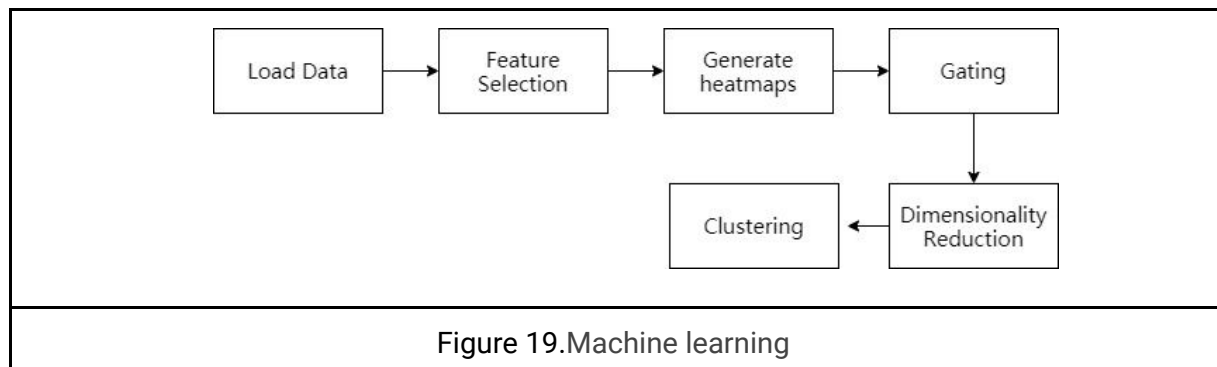
9. **/api/index.html : GET**

   This is the main entry point to the web application frontend.

10. **/api/files?location=loc?date-from=d1&date-to=d2&type=fcs : GET**

    Returns a list of all available FCS files and allows user to filter by date, location & type

Machine Learning Process
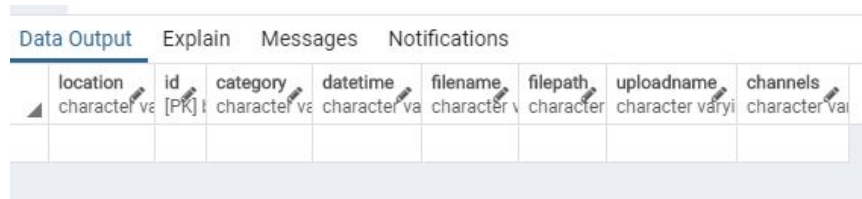

Figure 19.Machine learning

From a machine learning perspective, the steps are as follows:

1. **Load Data:** FCS files and accompanying meta data are uploaded from the frontend.
2. **Feature Selection:** Only the relevant columns are selected. In our case it was two columns. A scatter plot is then generated to allow the data to be represented visually. The same two columns are also selected in all other fcs files that match the data filters selected by the user.
3. **Generate heatmaps:** All the individual instances(tuples) are counted and represented in a matrix. Data range and binwidth determine the number of bins into which the data is aggregated. Smaller binwidth results in more bins which require more computing power and storage space.
4. **Gating:** The instances inside the gates are extracted and stored in a smaller file to await further processing. The gate coordinates determine the size of the gate.
5. **Dimensionality Reduction:** The gated files have multiple features and therefore need to be transformed to allow them to be represented in 2 dimensions. Single Value decomposition and Non-Metric Multidimensional Scaling are used in our case.
6. **Clustering:** Clustering algorithms are applied to the gated data whose dimensions have been reduced. By representing the data in 2 dimensions, experts can then determine which clusters correspond to which cells.

Data Storage Implementation

Our database of choice was Postgres due to the fact that it's open source, easy-to-use, scalable and has great performance. The application utilised two tables:

- **tbusers:** This helduser login information
- **tbmeta:** Stored metadata, FCS file names and extracted column names from the FCS files

Figure 20.Machine learning

Uploaded FCS files were stored in folders with a structure as shown below:

```
data/
├── raw/
├── transformed/
├── gated/
├── heatmap/
└── diff/
```

Figure 21. FCS folder file structure

Newly uploaded files were uploaded to the raw folder to await further analysis.

Analysis

Histogram 1d

Figure 22 below shows the histogram of the two channels. It shows the frequency distribution of points for a channel.

Figure 22: Histogram 1D
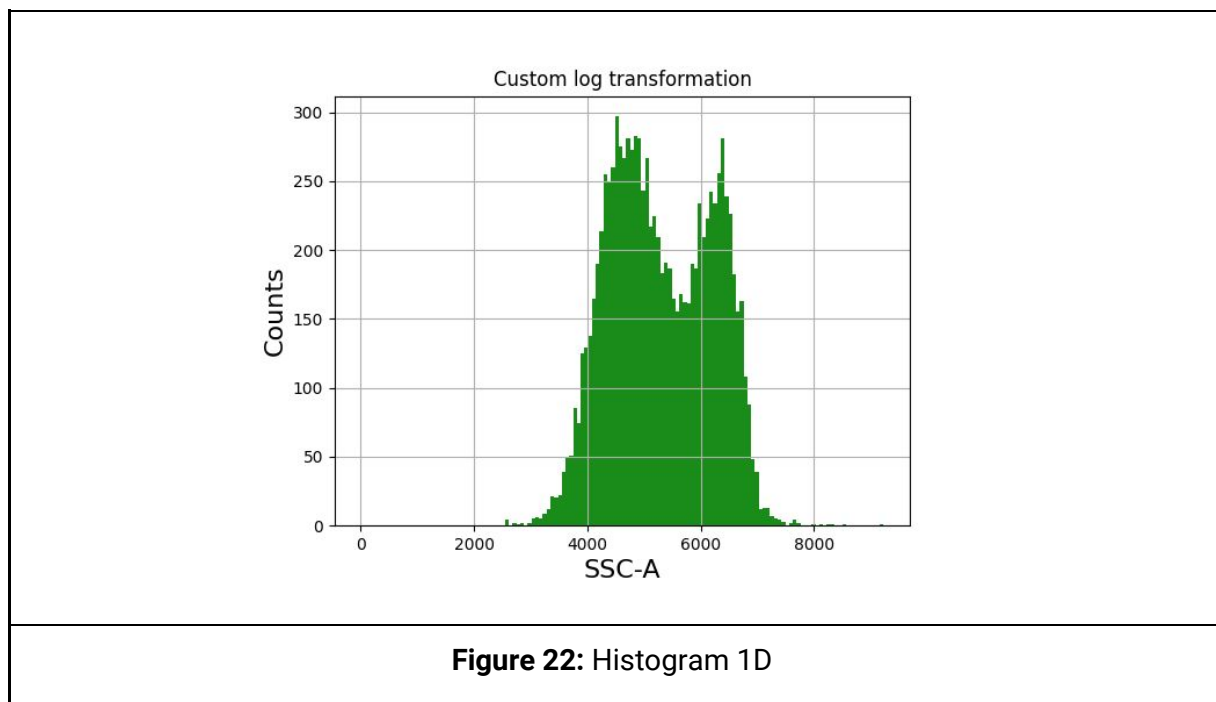
Histogram 2d

This is the histogram in two dimensions for the FCS file. It is plotting using the provided two channels from the select lists, as shown in Figure 23.
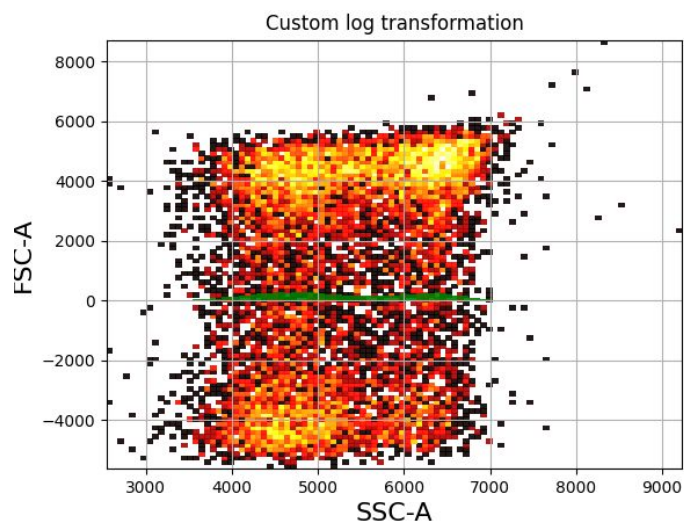
Figure 23: Histogram 2D

Scatter Plot

A scatter plot shows the relationship between two channels. Figure 24 shows the relationship between the two selected channels SSC-A and FSC-W.
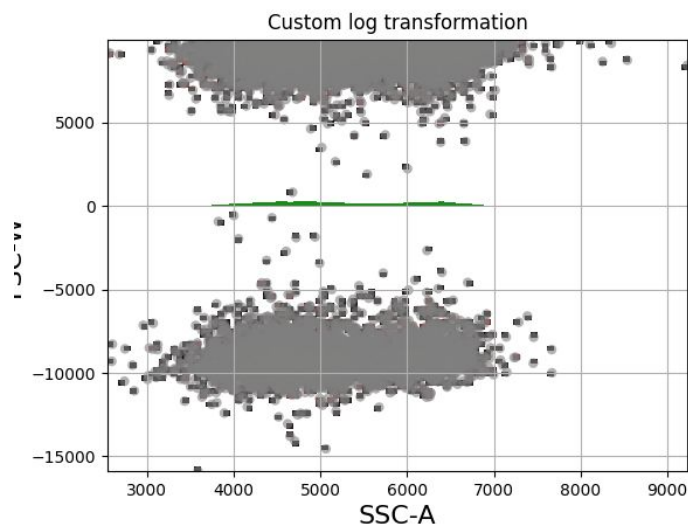
Figure 24: Scatter Plot (SSC-A, FSC-W)

Gates

Gating is a way for filtering channel data. There are several gates available in flow cytometry e.g Threshold, IntervalGate, QuadGate, PolyGate. The Figure 25 shows threshold gate which filters x-axis for value 1000. The left plot shows original while the right one shows the gated sample.

Figure 25: Threshold gate
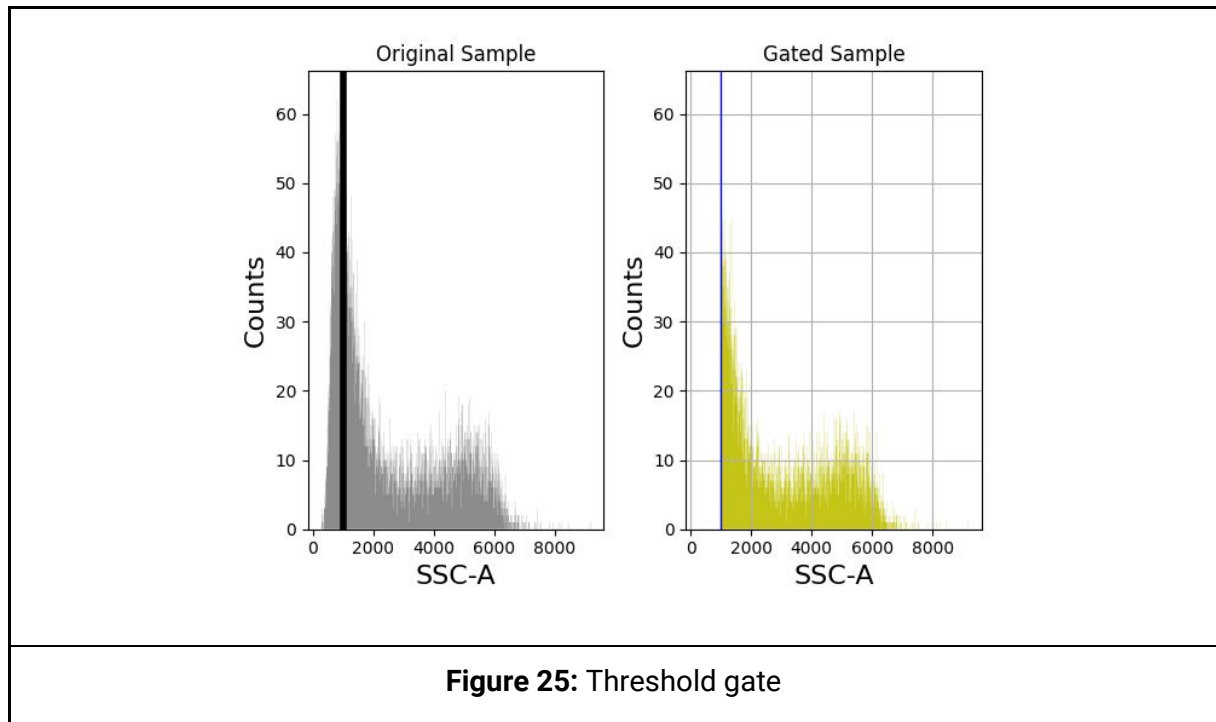
Compensation Plot

Flow cytometry is an optical technology that measures the properties using a laser beam. The material is dyed with different substances and when exposed to laser beam, it shows different fluorescence and light refractions. Each color goes to a separate channel, However there may happen spectral overlap. Compensation is a mathematical method to change channels data to rectify this problem. This is shown in Figure 26.
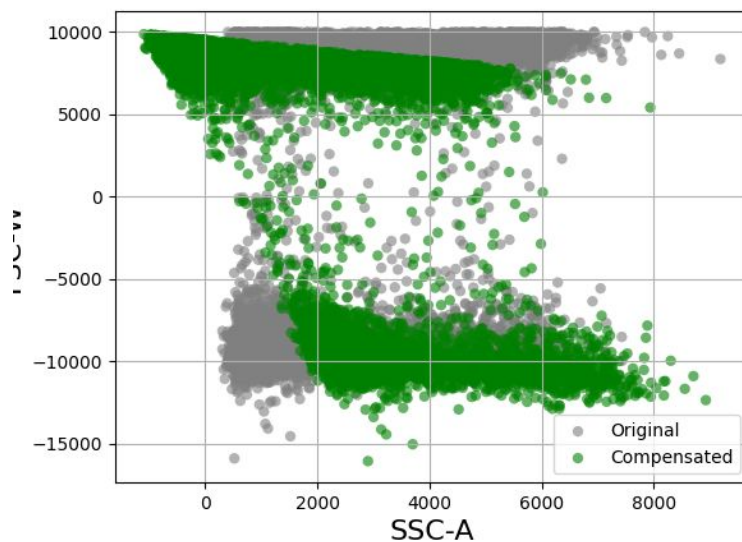
Figure 26: Compensation Plot

Gaussian Plot

The Gaussian Machine learning model is used to separate the data into two/more populations. It uses a default estimation method for parameters as shown in Figure 27.
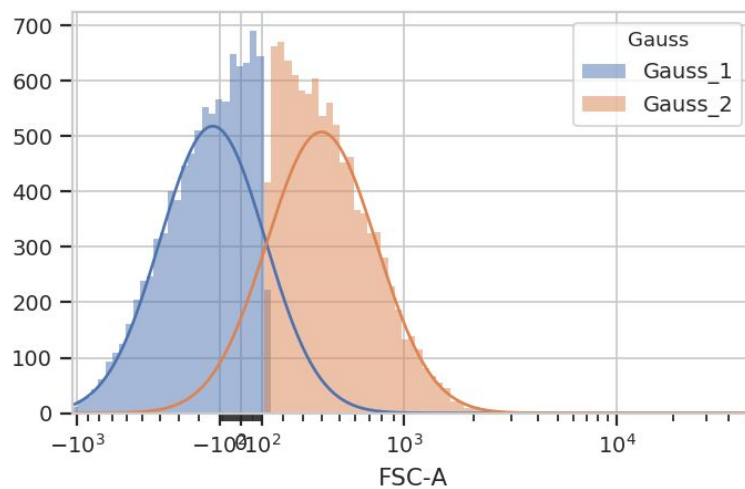
Figure 27: Gaussian Plot

Gausian mixture model two channels

The GaussianMixtureOp can work with multidimensions of channels. This is a plot of two channels workflow. This is shown in Figure 28.

Figure 28: Gaussian Mixture Model Plot
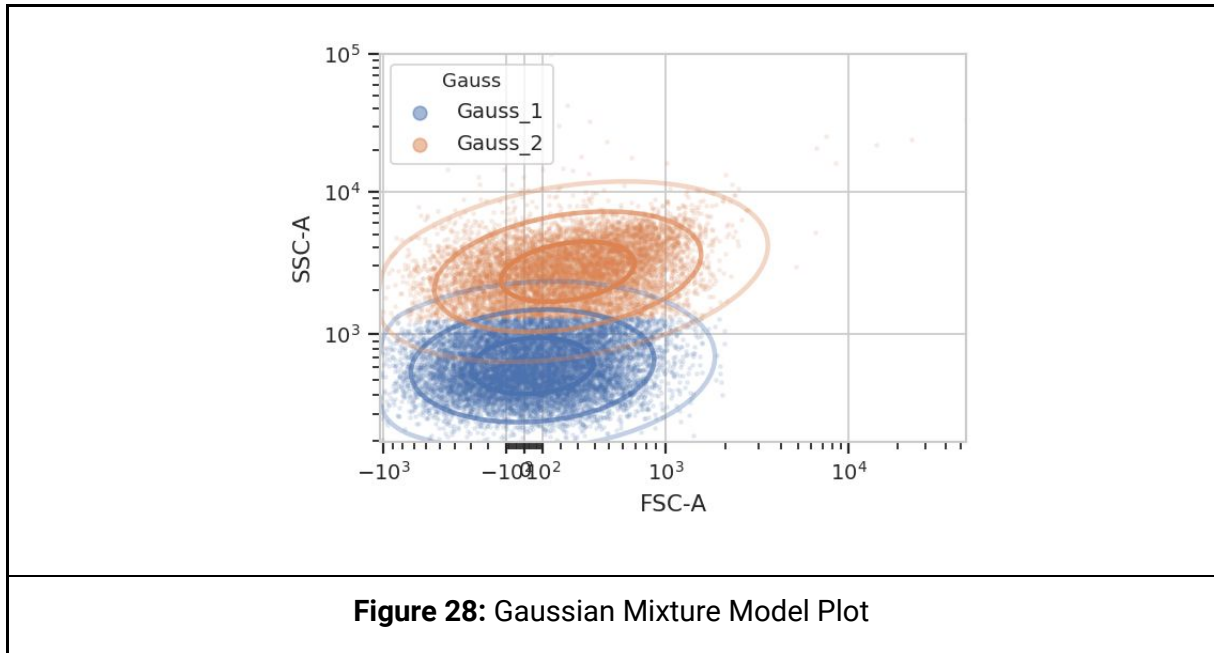
K-Means Clusters

Figure 19 below shows two clusters using K Means algorithm for K = 2. The centroid of the two clusters are marked with a star symbol. The channels used are FSC-A (Forward scatter) and SSC-A (Side scatter).

Figure 29: K Means clustering, K=2
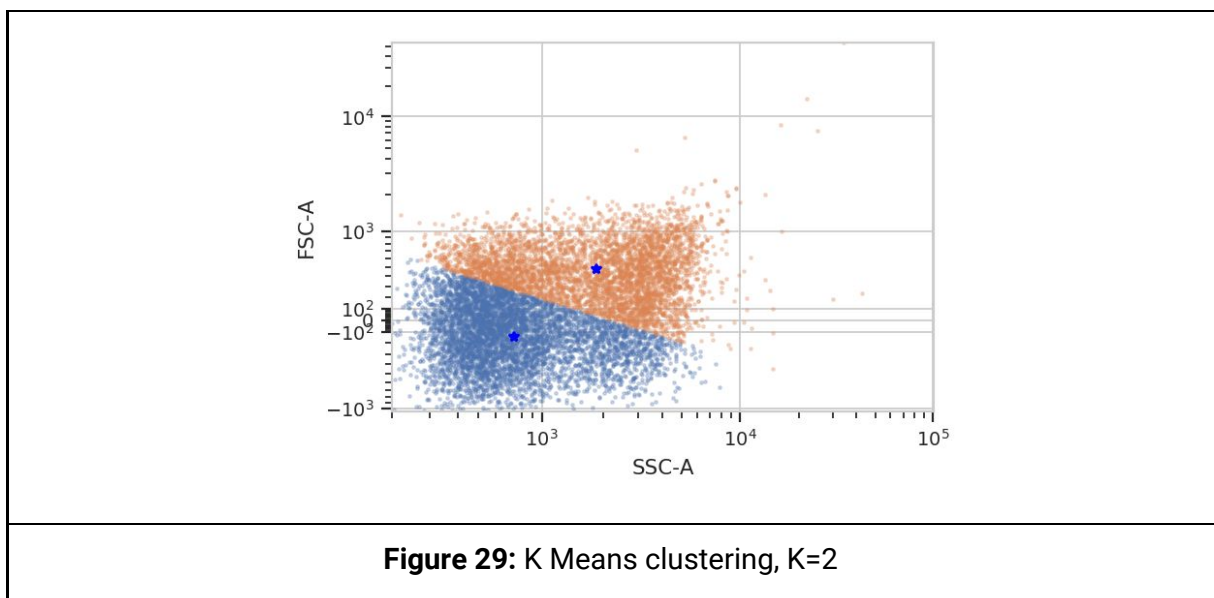
Similarly Figure 20 shows three clusters for the same channels but using K = 3.

Figure 30: K Means clustering, K=3
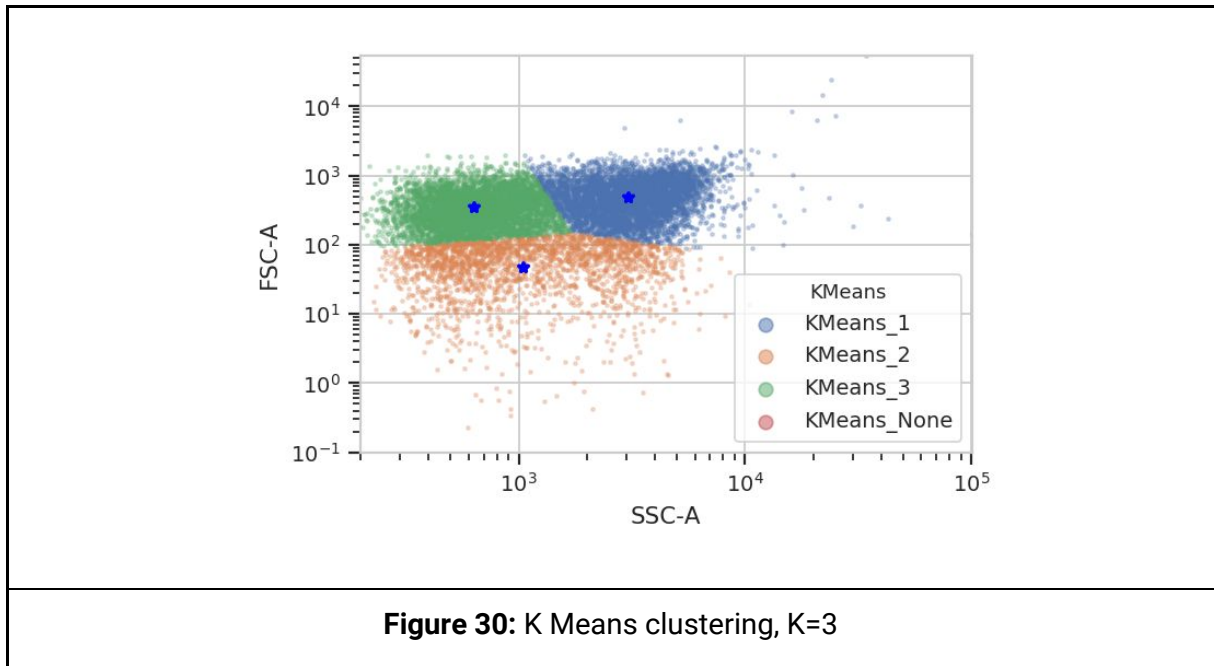
Classification

"In machine learning, classification refers to a predictive modeling problem where a class label is predicted for a given example of input data.[20]"

Binary classification is the type of classification which determines the two class labels for example sample. An easy example is to classify emails as either 'spam' or 'not spam'.

In this project we were given FCS files to be classified for either 'cancer' or 'no cancer'.

Popular algorithms that can be used for binary classification are:

- Logistic regression
- Decision Tree
- Random Forest
- K-Nearest Neighbors
- Naive Bayes
- Support vector machine

In our implementation we however used logistic regression, decision tree, random forest and then compared their accuracy. Logistic and SVM only supports binary classification.

In the classification we assigned a column diagnosis which shows classes 'cancer 1' or 'no cancer 0' for each cell.

RFP_Well_A3.fcs

Test Score

Decision Tree : 0.929
Logistic Regression : 0.929
Random Forest : 0.929

Number of Malignant and Benign cells

Decision Tree	Logistic Regression	Random Forest
Benign : 3056	Benign : 5458	Benign : 3056
Malignant : 6944	Malignant : 4542	Malignant : 6944

RFP_Well_A6.fcs

Test Score

Decision Tree : 0.929
Logistic Regression : 0.929
Random Forest : 0.929

Number of Malignant and Benign cells

Decision Tree	Logistic Regression	Random Forest
Benign : 2826	Benign : 5001	Benign : 2826
Malignant : 7174	Malignant : 4999	Malignant : 7174

RFP_Well_B3.fcs

Test Score

Decision Tree : 0.929
Logistic Regression : 0.929
Random Forest : 0.929

Number of Malignant and Benign cells

Decision Tree	Logistic Regression	Random Forest
Benign : 3252	Benign : 5271	Benign : 3252
Malignant : 6748	Malignant : 4729	Malignant : 6748

a1_24h_noc200.fcs

Test Score

Decision Tree : 0.929
Logistic Regression : 0.929
Random Forest : 0.929

Number of Malignant and Benign cells

Decision Tree	Logistic Regression	Random Forest
Benign : 20443	Benign : 21785	Benign : 20443
Malignant : 18675	Malignant : 17333	Malignant : 18675

Figure 31: Comparison - Logistic Regression, Decision Tree, Random Forest

We used a small function for this which could be tailored as per need. We then divided a sample dataset into a 70% train and 30% test data set. We evaluated the accuracy of our

three classification models i.e Logistic Regression, Decision Tree, Random Forest and found the results shown in Figure 31.

Evaluation of Results

R1. Allow users to securely log in

In the project, the login web page which we have developed is used for checking the authorized user, user need type username and password on the page, and then click the login button.

R2. Upload FCS files to database

The upload allows the upload of FCS files from the local computer to the database on the cloud.

R3. Notify user using visual feedback

Toast notifications are displayed in the system in all pages, when using login in the webpage successful or failed, upload FCS files successful or failed, and so on.

R4. Filter FCS files using metadata

In the upload webpage, Location and category are entered before uploading new FCS files. These will later on be used to filter the data on the analysis page.

R5. Analyze uploaded FCS files

The application has to allow uploaded FCS files to be analyzed. File analysis had been sequential to ensure ease-of-use.

R6. Allow user to select gates

On the FCS files analysis web page there are two X-Axis, and two Y-Axis must be select

R7. Display results in the webpage

After X-Axis and Y-Axis are selected, the results of the plot for FCS file analysis will show on the webpage.

R8. The FCS files should be selected from the database

The database has been designed to deploy on the cloud, on the upload page, the FCS files are saved on the database from the local computer, then the analysis Fsc file should be selected from the database.

R9.Allow selection of columns to be plotted and transformation function

In the system, on FCS files analysis webpage, there is an option part that must be selected, X-Axis, Y-Axis, Transformation. The options of X-Axis and Y-Axis is all parameters from the FCS file, the Transformation has HLOG and VLOG two options

R10. Implement visual charts

The user then selects the gate coordinates to be used after which they click on" Generate Clusters". Two charts will be shown on the screen.

R11. Train the prediction model

The system has not predicted the model by uploaded files.

R12. Create separate Upload and Analysis pages

Two different pages were created which allowed for the separation of the 'upload' and 'analysis' functions. Users can log in to perform only a single function which greatly improves security.

Ethics

Effects on the environment: Data mining and machine learning algorithms consume a lot of resources in terms of computing power and storage. As much as the benefits may be

considered to outweigh the disadvantages, this should not be an excuse to inefficiently manage computing resources

Data Privacy: Such data is very valuable. Medical companies need data for their research. It is important that such data is kept from the wrong hands. User authentication and ability to select user functionality helps to ensure data integrity and safety. Uploads may be accepted but downloads may not be as essential to research work.

CONCLUSION

This project presents a tool to store and analyze the cell data in order to make it easier to identify patterns/markers that allow for the creation of a model to predict presence of prostate cancer. We have explored the use of various data mining and machine learning techniques in extracting useful information to inform decisions and potentially even influence them in the future.

A lot of interesting discussions on approaches to the problem at hand were raised, the best of which are included and described in this document. Well-formed requirements at the early stages of a project remain to be one of the most important things to ensure goals are met and to minimize resource wastage by narrowing the focus.

Millions of people die every year due to the late detection of cancer. The importance of this project in relation to sustainable development goals is immense. It touches heavily on good health and well-being, and by extension, all the other SDGs since they are all related. The dimensionality of medical data remains to be a problem that may lead to greater energy consumption as the numbers are crunched but the tradeoff in this case is negotiable. Health as we know it is on the brink of a revolution.

We successfully analysed the data, clustered it and made classification models to help identify the different cancer cells. We developed tools which help visualize the data with different plots and enabled the user to tailor several parameters of the dataset for a more accurate evaluation.

FUTURE WORK

1. Supporting efficient concurrent analysis by multiple users. Users analyzing the same set of files should be able to work on the same dataset without affecting each others work

2. Include support for multiple clustering algorithms and dimensionality reduction techniques that can be applied dynamically.

3. In the system architecture the gateway is a central point for interaction between the end users and the services and it manipulates the data of authentication and can be considered as the man in the middle. This can be used for message sniffing and crafting for further security analysis.

4. We have developed a limited web application for analysis in the small scope of the course. We however laid down a great foundation using the techniques and technologies. Therefore both the frontend and backend can be extended further for adding more user interaction and features.

5. We laid down a basic foundation for application scaling using Docker for microservices, a future work would be to add scalability further.

REFERENCES

[1] "Sustainable_Project_In_Multi_disiplinary_Context.pdf: DT586B VT20 Sustainable Projects in Multidisciplinary Contexts."
https://hkr.instructure.com/courses/2444/files/357453?module_item_id=100692 (accessed May 21, 2020).

[2] "2017 China Cancer Status and Trends Big Data."
http://www.360doc.com/content/18/0104/14/47491966_718986607.shtml (accessed May 21, 2020).

[3] dpi campaigns, "About the Sustainable Development Goals," *United Nations Sustainable Development*.
https://www.un.org/sustainabledevelopment/sustainable-development-goals/ (accessed May 21, 2020).

[4] "Sustainable Projects In multi disciplinary context," HKR.

[5] "Files." https://hkr.instructure.com/courses/2444/files (accessed May 21, 2020).

[6] "Understanding K-means Clustering in Machine Learning."
https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1 (accessed May 21, 2020).

[7] "Work breakdown structure - Wikipedia."
https://en.wikipedia.org/wiki/Work_breakdown_structure (accessed May 21, 2020).

[8] "KSU Faculty - Time and planning." https://fac.ksu.edu.sa/ (accessed May 21, 2020).

[9] "Getting started with Django | Django." https://www.djangoproject.com/start/ (accessed May 21, 2020).

[10] "Fcm - A python library for flow cytometry | Semantic Scholar."
https://www.semanticscholar.org/paper/Fcm-A-python-library-for-flow-cytometry-Frelinger-Richards/b1b1f5d6ec41a5d56ed4218fe255c7b446c4b2ff (accessed May 21, 2020).

[11] "Software Architecture in Practice - Len Bass - inbunden (9780321815736) | Adlibris Bokhandel."
https://www.adlibris.com/se/bok/software-architecture-in-practice-9780321815736 (accessed May 21, 2020).

[12] "Robust Kinetics of RNA Virus: Transcription Rates Are Set by Genome Levels."
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5653219/ (accessed May 21, 2020).

[13] S. Mahoney, F. Arfuso, M. Millward, and A. Dharmarajan, "The effects of phenoxodiol on the cell cycle of prostate cancer cell lines," *Cancer Cell Int.*, vol. 14, p. 110, 2014, doi: 10.1186/s12935-014-0110-z.

[14] K. Zhang and D. Waxman, "PC3 prostate tumor-initiating cells with molecular profile FAM65B/MFI2/LEF1increase tumor angiogenesis," *Mol. Cancer*, vol. 9, p. 319, 2010, doi: 10.1186/1476-4598-9-319.

[15] "PC3 Is a Cell Line Characteristic of Prostatic Small Cell Carcinoma."

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3426349/ (accessed May 21, 2020).

[16] A. Rasul *et al.*, "Reactive Oxygen Species Mediate Isoalantolactone-Induced Apoptosis in Human Prostate Cancer Cells," *Molecules*, vol. 18, pp. 9382–9396, 2013, doi: 10.3390/molecules18089382.

[17] S. Zhang *et al.*, "Proteomic analysis of human prostate cancer PC-3M-1E8 cells and PC-3M-2B4 cells of same origin but with different metastatic potential," *PLOS ONE*, vol. 13, no. 10, pp. 1–25, 2018, doi: 10.1371/journal.pone.0206139.

[18] "Optimized processing and analysis of conventional confocal microscopy generated scanning FCS data - ScienceDirect." https://www.sciencedirect.com/science/article/pii/S1046202317302554 (accessed May 21, 2020).

[19] "Fluorescence Correlation Spectroscopy (FCS) - Core Facilities, University of Gothenburg, Sweden." https://cf.gu.se/english/centre_for_cellular_imaging/light-microscopy/lm_techniques/Fluorescence+Correlation+Spectroscopy+%28FCS%29 (accessed May 21, 2020).

[20] "4 Types of Classification Tasks in Machine Learning." https://machinelearningmastery.com/types-of-classification-in-machine-learning/ (accessed May 24, 2020).

APPENDIX

A. Project Plan

Tasks of project	startTime	endTime	task of progress in Feb																	task of progress in Mar																			
			12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	26
learning & Analysis	12-Feb	15-Feb																																								
setUp Enviroment	16-Feb	19-Feb																																								
made TimeSchedule	19-Feb	20-Feb																																								
implement FrontEnd	21-Feb	29-Feb																																								
implement BackEnd	21-Feb	4-Mar																																								
combind Project	3-Mar	8-Mar																																								
problem Solution	9-Mar	11-Mar																																								
testing	12-Mar	14-Mar																																								
modify bugs	15-Mar	17-Mar																																								
finally report	18-Mar	20-Mar																																								
finally Presentation	26-Mar	26-Mar																																								
organization Project	12-Feb	20-Mar																																								

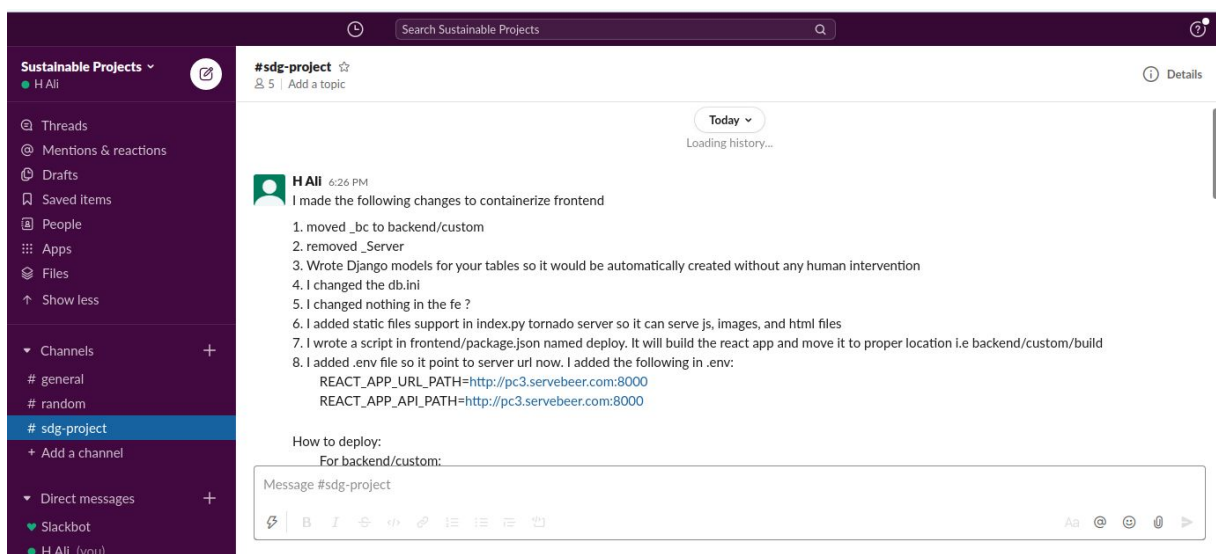|	
group work	
jing huang	
Hizrat Ali	
Pascal King'oku	
Xiao Su	

B. Project Management Tool

Trello
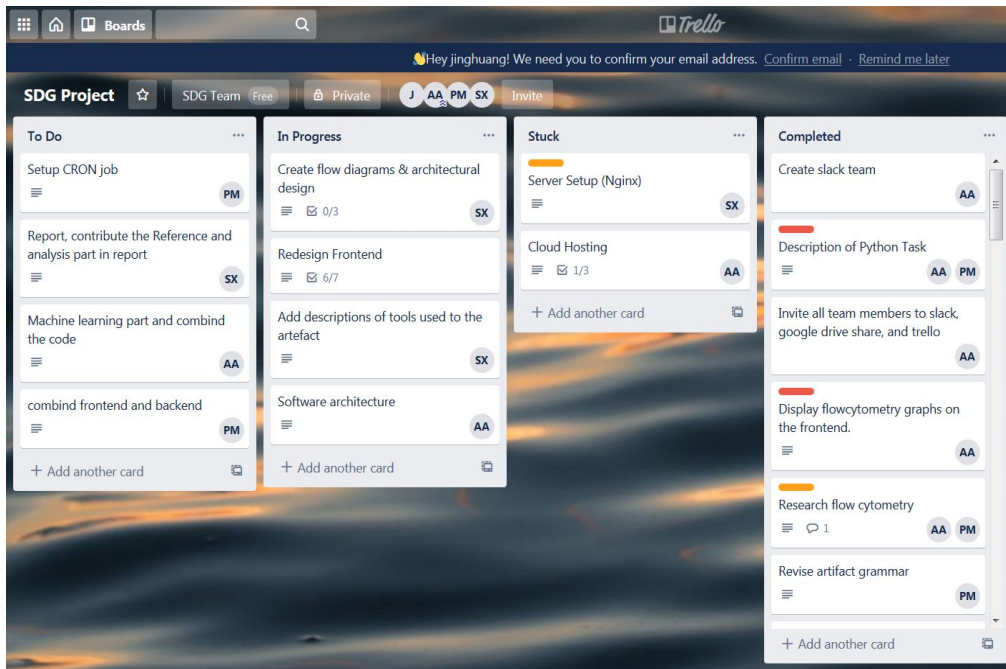
Slack

C. Design - Tools & Technologies

Draw.io

draw.io is the most tightly integrated Google Drive diagramming application available. draw.io. Works with draw.io is a completely free online diagram editor built around Google Drive(TM), that enables you to create flowcharts, UML, entity relation, network diagrams, mockups, and more.

Code2Flow

code2flow is pseudo code to flowchart generator. Use our smart syntax to generate optimal, beautiful, and readable diagrams automatically. You focus on the process, we do the rest. It takes just a few seconds — whether it is a complex technical algorithm, a business flow or anything in between.

D. Frontend - Tools & Technologies

React JS

React JS is a JavaScript library for building user interfaces. It is maintained by Facebook and a community of individual developers and companies.
React can be used as a base in the development of single-page or mobile applications. However, React is only concerned with rendering data to the DOM, and so creating React applications usually requires the use of additional libraries for state management and routing. Redux and React Router are respective examples of such libraries.

Canvas JS

This is a popular javascript and html5 data visualization library. It creates visually appealing and responsive charts .

Heroku

This is a popular platform as a service(PaaS) that allows users to host applications and databases in the cloud. We used it to host the frontend part of the application and the database.

NPMS.IO

This is a better open source alternative to npm.js that organizes node packages that takes into consideration attributes like download count, publish date, number of forks and stars on github etc. It sorts packages by quality, maintenance, popularity and personalities. This creates a better ranking system and saves developers time by making it easy to find the right packages for their project.

E. Backend - Tools & Technologies

Python

Python is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including procedural, object-oriented, and functional programming.

Virtualenv

It is simply an isolated copy of python packages related to a specific project. It avoids conflicts in different versions of packages for different projects.

If could be installed as :

sudo pip3 install virtualenv

Google cloud

Google Cloud Platform (GCP), offered by Google, is a suite of cloud computing services that runs on the same infrastructure that Google uses internally for its end-user products, such as Google Search and YouTube. Alongside a set of management tools, it provides a series of modular cloud services including computing, data storage, data analytics, and machine learning. Registration requires a credit card or bank account details.

Google Cloud Platform provides infrastructure as a service, platform as a service, and serverless computing environments.

Numpy

An open source library that adds support for large, multidimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays

Pandas

A BSD licensed, an open source library that provides high performance, easy-to-use data structures and data analysis tools

Sklearn

An open source library that provides tools for data mining and data analysis.

Matplotlib

An open source plotting library for the Python programming
language and its numerical mathematics extension NumPy

MongoDB

MongoDB is a cross-platform document-oriented database program. Classified as a NoSQL database program, MongoDB uses JSON-like documents with the schema. MongoDB is developed by MongoDB Inc. and licensed under the Server Side Public License (SSPL).

particularity in MongoDB.*[9]*

1. Memcached, Redis belongs to the key/value database
2. The difference between MongoDB and the above is that it belongs to the document database and stores documents (Bison (when modifying the JSON string based on

JSON, the position of the data behind this JSON string does not change, introduce space)-> binary or JSON)

3. The internal engine uses a JS interpreter to store the document into a bison structure. In the query, it is converted to a JS object and can be operated through the familiar JS syntax.

4. compared with traditional data - Traditional database: structured data, after the table structure is set, the content of each row must conform to the table structure, that is, the type of the number of columns is the same.

   Mongo document database, each document under the table can have its own unique structure (movie reviews, one table manages all reviews and responses, at least 4 tables in a traditional database, the correlation is very complicated)

5. mongo command line can write for loop and other scripts

6. Index:
   a. Indexes improve query speed and reduce write speed, so you must weigh commonly used query fields without having to build indexes on too many columns.
   b. In mongo, the index can be created in ascending and descending order of the field (1 takes effect, -1 descending order) to create an index for easy sorting.
   c. By default, Btree is used to organize the index file (wearing the index will generate the index file and also occupy space memory). The hash index is also allowed to be created after the 2.4 version.
   d. db.find (query) .explain (), if the output shows | "cursor": BasicCursor indicates that no index works.
   e. Index classification: ordinary index (single-column index, multi-column index), unique index (the indexed key value is unique, otherwise an error, you need to set unique: true), sparse index (if the value of the indexed field is empty In contrast, the index will not be established. The ordinary index will consider the value of the field column of the document to be null and index regardless of whether the value of the indexed field is empty.

Nginx

It is web server with the following features:
1. config file
2. upstart script
3. nginx sites available

F. Testing & Network Communication Tools

Postman

A postman is a powerful tool for debugging the HTTP interface created by Postdoc Technologies.

Postman's main features are an easy-to-use graphical user interface, which saves the history of API requests, unlimited use of collections, environment variables, running tests, and shared collections. You can use the collection runner to automate testing, flexible API monitoring, runtime, performance And accurate, simulated server, supports split-stack development.

In this project, we mainly receive the information sent by post in reactjs to check whether the coding is correct

SSH

Secure Shell (SSH) is a cryptographic network protocol for operating network services securely over an unsecured network. Typical applications include remote command-line, login, and remote command execution, but any network service can be secured with SSH.

SSH provides a secure channel over an unsecured network in a client-server architecture, connecting an SSH client application with an SSH server. The protocol specification distinguishes between two major versions, referred to as SSH-1 and SSH-2. The standard TCP port for SSH is 22. SSH is generally used to access Unix-like operating systems, but it can also be used on Microsoft Windows. Windows 10 uses OpenSSH as its default SSH client.

FTP Client

An FTP client is a software that uses the FTP protocol to transfer files to and from a remote computer.

FTP is the most widespread transfer protocol used to transfer files from one host to another over the Internet.

CURL

It is a command line browser. To upload a file use it like:

curl -F 'fcs=@/home/alex/projects/sustainable_projects/project1/README.md'

http://project1.loc/app1/upload

G. Revision History

Date	Version	Description	Author
2020 - 02 - 10	1.0	Initial Project meeting to specify requirements Added initial introduction initial Requirements Initial Supplementary requirements	Pascal King'oku Hazrat Ali Xiao Jing
2020 - 02 - 10	2.0	Added intro and fig about FCM process Initial Test	Hazrat Ali Xiao
2020 - 02 - 26	3.0	Chapter 2 - Software Architecture Design modifications & document edits Appendices	Hazrat Ali Pascal King'oku Xiao
2020- 03 - 17	4.0	Updates to design	Pascal King'oku
2020- 04 - 06	5.0	Planing	Jing Huang
2020- 04 - 11	6.0	Backend completed Backend deployed online	Hazrat Ali
2020- 04 - 12	7.0	Formatted sections into chapters Added authentication and SW architecture sections Added Backend section	Hazrat Ali
2020- 05 - 17	8.0	Added classification model to the backend Added installation instructions to Readme	Hazrat Ali

2020- 05 - 21	9.0	Fixed references - Added to Zotero Reference management system

Chapter titles i.e H1 capitalized

Fixed several formatting errors, line numbers, indentation, heading levels, misplaced contents | Hazrat Ali |
| 2020- 05 - 21 | 10.0 | Integration backend & frontend

Installation on live cloud server

Formated images and tables | Hazrat Ali |
| 2020- 05 - 24 | 11.0 | Added Analysis - 8 pages in report

Corrected Figure numbering

Fixed appendices

Fixed casing and headings levels

Addition to future work | Hazrat Ali |

H. Links

- Repo https://github.com/iloveyii/sdg-project
- Live Link 1 http://pc3.servebeer.com/
- Live Link 2 http://pc3.servebeer.com:8000/
- Docker hub https://hub.docker.com/u/alikth
- University https://www.hkr.se/