FIT3163 - Data Science Project 1 - Assignment 6 - Project Proposal with Literature Review

Workshop: Friday, 2pm - 4pm

Group: FIT3163_CL_04

Members: Elaine Liong (ID: 29942357),

Jack Ooi (ID: 29037077),

Vionnie Tan (ID: 30092809)

Word Count: 7835 words

1.2. Table of Contents

2. Introduction

In Australia, cancer is one of the leading causes of death - accumulating to over 50,000 deaths in 2019 (Cancer Council Australia, n.d.). The Australian Institute of Health and Welfare (2020) predicts that in 2020, new cases of cancer are expected to rise to up to 150,000 cases. As a result, detecting early stages of cancer has been highly sought after by researchers alike. The growth of cancer predictability has also grown significantly since the introduction of Machine Learning and Artificial Intelligence. Despite this positive growth, cancer survivability within patients - especially on the five-year scale after first receiving treatment remains unsatisfactory as most cancer patients suffer from relapses and metastasis (Deng et al., 2015). This problem statement gives rise to the purpose of this proposal, where our team aims to build a successful predictive model that allows for earlier detection of gastrointestinal cancers through detecting effective biomarkers and using image processing methodologies whose final product will be distributed publicly as an online tool for health institutions to use. Detecting cancer tumours at an earlier stage could increase survivability within patients (Deng et al., 2015). In line with our primary aims for this project, other supplementary aims that we aim to achieve include developing the predictive model with an accuracy greater than 95% and fast delivery of results that would optimally be below 5 minutes. Success of the project could be stemmed from customer and stakeholder satisfaction in using our final product.

Included in this proposal is a brief description of the contextual meaning of cancer predictive modelling, an extensive record of our project management plans that involves both design and testing aspects of our project. As most of the project execution would begin next semester, touching up on the technical skills to build the predictive model such as learning on how image processing works is of utmost priority. Next, we would put these skills into deployment of the predictive model through rigorous testing and training from the given dataset - improving the accuracy and efficiency of the model as well as fixing bugs that arise from the model. Lastly, successful building of the predictive model leaves us to build the online tool through an application called WordPress. Once it is built, that would complete most of the preliminary requirements of the project. Supplementary documentations including final report, presentation and users manual would also be given by our team.

3. Literature Review

3.1. Introduction to Literature Review

Cancer is one of the main causes of death around the world at the present time. According to the World Health Organization (n.d.), “cancer is more likely to respond to effective treatment when identified early, resulting in a greater probability of surviving as well as less morbidity and less expensive treatment”. This suggests that proper diagnosis of cancer in its early stages can help save thousands or millions of lives and therefore, it is important for us to find suitable data mining techniques to detect cancer at the earliest possible stage using predictive modelling. This literature review will outline the background of the project and the rationale for the project. Additionally, it will also illustrate the discussion of predictive models for cancer and provide synthesis and evaluation of relevant published work. This review will focus on the history of cancer classification techniques, issues of medical data mining, classification techniques, image processing techniques, and feature extraction techniques.

3.2. Content

History of Cancer Classification Techniques

According to Munir et al. (2019), the typical cancer classification techniques used by doctors regularly include Asymmetry, Border, Color and Diameter (ABCD) method, seven-point detection method, Menzies method, and pattern analysis. These methods are applied to medical images such as ultrasound images, X-ray images, and computed tomographic (CT) images. These methods, however, are considered inefficient and doctors are demanding better methods for cancer diagnosis (Munir et al., 2019). Furthermore, investigation of medical images is always very challenging, especially in the case of histopathological imaging due to its complex nature (Nahid et al., 2018). Nowadays, “[a]rtificial intelligence and cancer diagnosis are gaining attention as a way to define better diagnostic tools" (Munir et al., 2019). Munir et al. (2019) also mentioned that deep neural networks (DNNs), in particular, provide the ability to successfully analyse images intelligently. These DNN models, that include Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), have made revolutionary improvements in the data analysis field (Munir et al., 2019).

Issues of Medical Data

There are many sources to obtain medical data such as images, interviews with the patients, physicians' notes and interpretation etc (Delen, 2009). Delen (2009) illustrates the ideas of a variety of data and ethical and social issues of health data. According to Delen (2009), medical data is hard to obtain as the data are large, complex, heterogeneous, hierarchical, time series and of varying quality. Kumar et al. (2014) supports the heterogeneity of data mentioned by saying raw health data is huge and heterogeneous. From the heterogeneity data point of view, Kumar et al. (2014) states that medical data is written in different grammatical structures to explain the relations among medical entities. While for the ethical and social issues, Delen (2009) says that medical data is related to humans so that there is a chance where the patient's confidentiality might be breached and the possibility of ensuing legal action. Overall, many studies indicate that there are some issues when the process of medical data mining is taking place.

Classification techniques

There are multiple algorithms of modelling available for data categorisation and training and testing such as Artificial Neural Networks (ANNs), Decision trees (DTs) and Support vector machines (SVM) (Gupta et al., 2011). According to Delen (2009), decision trees separate observations into branches and construct a tree for the purpose of improving the prediction accuracy. Kumar (2014) shows the idea of a decision tree is by classification and sorting them down the tree from the root. Additionally, after the process of learning, ANNs is able to forecast changes and events in the system (Kumar et al., 2014). In the learning phase of the networks of ANNs, the networks adjust the weights so that they are able to predict the correct class label of the input (Gupta et al., 2011). SVM works by separating a given set of binary labeled training data with a hyper-plane that is maximally distant from them (Furey et al., 2000). As stated by Kumar et al. (2014), DT is faster than ANNs, however, Gupta et al. (2011) do not agree and stated that methods have their own advantages and disadvantages. Literature suggests that different classification techniques serve different purposes.

Image processing techniques

According to Ansari et al. (2017), there are three steps to use image processing to detect cancer which are image preprocessing, image segmentation and feature extraction. Jain

et al. (2015) also emphasizes the same three ways of image processing techniques. Additionally, Ansary et al. (2017) provides more details on pre-processing skills such as grayscale conversion, noise removal and image enhancement. Likewise, Jain et al. (2015) provides different pre-processing skills such as image resizing, image contrast and brightness adjustment and gamma correction. Based on Ansari et al. (2017) and Jain et al. (2015), image segmentation can be done by using automatic thresholding and masking operation in R, G and B planes. Next, Ansaru et al. (2017) suggest that gray level co-occurrence matrix is effective on feature extraction. Jain et al. (2015) propose by extracting the geometric features of segmented skin lesions to extract features. Overall, different image processing techniques come with different pros and cons.

Feature extraction techniques

According to Alom et al. (2018), there are several deep learning algorithms for feature extraction including Convolutional Neural Networks (CNN), Alexnet, Resnet etc. According to Indolia et al.(2018), CNN is described as the concept of hierarchical feature detectors in a biologically inspired manner. Alom et al. (2018) says that Resnet is a design of ultra-deep networks that did not suffer from the vanishing gradient problem. Alom et al. (2018) also says that AlexNet is the most recent stage in the development of all traditional machine learning approaches. Additionally, Demir et al. (n.d) reported that the accuracy of using resnet for early detection of skin cancer has reached 84.09%. Spanhol et al. (2016) also reported that by using CNN, it gained an accuracy of 80.8% to 85.6% when predicting breast cancer. Titoriya (2019) has achieved an accuracy between 93.8% to 95.7% when using Alexnet when predicting breast cancer. Literature suggests that different Alexnet feature extraction techniques produce higher accuracy than others.

3.3. Conclusion

In conclusion, cancer predictive models implemented using different machine and deep learning methodologies have boosted the predictability of cancer occurrences within patients, allowing them to receive earlier treatment. This brief literature review has covered the history of cancer classification techniques such as Menzies method, followed by the recurring issues associated with ethical and social issues on mining data (Delen, 2018). Several classification, image processing and feature extraction techniques such as the DNN models that include Convolutional Neural Network (CNN) and Recurrent Neural Network

(RNN), have made revolutionary improvements in the data analysis field (Munir et al.,2019). According to Gupta et al. (2011), the ongoing growth of different classification techniques fosters their own advantages and disadvantages. There are also several ways to process images such as image preprocessing, image segmentation and feature extraction (Ansari et al, 2017). Last but not least, among all the feature extraction techniques such as CNN, Alexnet and Resnet, Alexnet has the highest accuracy when it comes to predicting followed closely by Resnet (Titoriya, 2019).

4. Project Management Plan

4.1. Project Overview

As mentioned in the introduction, our main project objective is to successfully build an online tool used to predict early stages of gastrointestinal cancer deployed as a website for health institutions to use. To do this, we would utilize machine learning and deep learning methodologies such as feature extraction from packages such as Convolutional Neural Network (CNN), Residual Neural Network (ResNet), and AlexNet as well as determining predictors that have the greatest impact on categorization and classification models. The project timeline spans across a year and is divided into two major parts, whereby the first part of the project is dedicated to project planning, management and preliminary designs. The second part of the project focuses on further development of the predictive model through coding and rigorous testing and training of the dataset. Some of the milestones we aim to achieve in this project is to create a model that is equally fast, efficient and accurate as well as creating a user-friendly interface for our customers to easily navigate and obtain their results.

4.2. Project Scope

4.2.1. Project Deliverables

The expected project deliverables from our online tool for cancer prediction could be segmented into two different parts. Firstly, project management-related deliverables such as developing a business case related to our project, inclusion of a weighted scoring model to further emphasize project decisions. With regards to our Project Scope, our team has developed a scope statement that succinctly describes product characteristics and requirements associated with the project deliverables. In addition to this, a requirement traceability matrix would be further developed - clearly depicting the functional and non-functional requirements of the project. All projects come with their associated risks. To keep this in check, our team has chosen to develop a risk register that states possible risks and challenges that could be faced by our team members throughout the project.

Secondly, Product-related deliverables of the project include a copy of our predictive model's software code as well as inclusion of a research report covering extensive depths of our cancer prediction modelling. Further deliverables include our online tool for cancer prediction in the form of a successful website where users can publicly access them and gain information regarding their probability of having cancer. As part of our online tool

development, we have also included our final design document and software code associated with it. Upon successful completion of the project, a brief users manual as well as a demo would be given to our users with extensive “How-to”s and general accessibility of our software.

To sum up, our project deliverables are as follows:
Project management-related deliverables:

1. Business Case
2. Weighted Scoring Model
3. Scope statement
4. Requirements Traceability Matrix
5. Risk Register
6. Final Project Presentation
7. Final Project Report

Product-related deliverables:

1. Predictive model software code
2. Research Report
3. Online tool for cancer prediction
4. Online tool design document
5. Online tool software code
6. Users Manual
7. Demo
8. Final Report and Presentation

4.2.2. Product Characteristics and Requirements

4.2.2.1. Functional and Non-functional Requirements

Our online tool for cancer prediction has its unique characteristics that could be divided into two major parts - Functional and Non-Functional Requirements. Functional Requirements in this case, describes what the system should do while Non-Functional Requirements describes how the system works. In line with our project requirements, functional requirements such as knowledge regarding the programming skills needed to successfully build an efficient predictive model, identifying important predictors for

successful cancer categorization, and finding relevant datasets coming from reputable sources such as Kaggle that could serve as additional datasets used to build our predictive model. Not only that, but choosing which software systems we would use in our project timeline such as developing a completely new system or building on pre-existing software systems makes part of our Functional requirements. On the other hand, Non-Functional Requirements such as user-friendliness of the UI Interface aims for ease of access when customers access the website. At the end of the project timeline, our team also aims to meet stakeholder expectations from our predictive modelling - that is, but not limited to, efficient and fast model output times.

4.2.2.2. Requirements Traceability Matrix

To ensure all the project requirements are addressed, a requirements traceability matrix has been created (see Appendix A for requirements traceability matrix). As mentioned above, the requirements are divided into functional and non-functional requirements. The functional requirements include image processing, finding important predictors, programming skills, as well as integrating the software system with the health institutions' API, whereas the non-functional requirements include developing the user interface and meeting stakeholders' expectations. The requirements traceability matrix will help our team members to check whether all mentioned requirements are met and fulfilled.

4.2.3. Product User Acceptance Criteria

To ensure that stakeholders' expectations are met, our team aims to draw from the preliminary requirements expected from our project and extend it into creating user stories that would lead to the building of a suitable acceptance criteria. With our end product being an online tool used for cancer prediction, a suitable user story would allow users to view and access their results anytime and anywhere through login functionality on the website. This means that the acceptance criteria for this user story would be the creation of an online tool for cancer prediction that allows membership modelling whereby new users could sign up and existing users could login to their respective accounts. The developed cancer predictive model also deals with sensitive data associated with their private user information, which means that an acceptance criteria regarding the integrity and handling of private information should be maintained. Another acceptance criteria that is associated with our project is to ensure that expectations with the user's are met.

4.3. Project Organisation

4.3.1. Process Model

The life cycle model to be used for this project is agile. Agile Methodology, which focuses on the flexibility of project scopes and adapting towards continuous changes throughout the project. It is often described as an iterative approach over traditional project management methodologies that takes into account the rigorous importance of initial planning and an unchanging scope (Serrador & Pinto, 2015). This approach is more preferred than the other approaches like predictive, adaptive, or incremental as this approach emphasizes on individuals, interactions, customer collaboration and responsiveness (Serrador and Pinto, 2015). Furthermore, according to Ungureanu and Ungureanu (2014), Agile approach is specific and applicable especially for collaborative teams. As agile methods are not strongly constrained by budget for materials, which is also applicable for us, as a project given by university, no budgets are given to us so that in this case agile method works better than predictive or waterfall methods. Buganová and Šimíčková (2019) mentions that agile methods have higher adaptability to change and the contracting authority which is us students can adapt to the project to the current needs of the customer or the feedback which is the feedback given from tutor. This will lead to higher customer satisfaction as it ensures that our product is according to the customer's requirements. This approach is also chosen since this approach is operated in short and rapid cycles and actively uses feedback which will allow our team to quickly adapt to changing environments and requirements. This will eliminate the likelihood of our project failing completely as testing is regularly done to ensure our final product quality is maintained.

4.3.2. Project Responsibilities

There are two major project functions related to our project - namely executing the predictive modelling and final website design. For the predictive model function of our project, we plan to delegate tasks equally between the three of us as neither of us has had experience in image processing. So, we plan to work together simultaneously using the best use of our current knowledge and try to successfully execute the predictive model. Our utmost priority is to build the predictive model successfully, then we would use further knowledge to try to make the model more efficient. To build the predictive model together, it is important for us to keep track of our current codes and executions which could be done

through our Github repository as well as online storage systems (Google Drive) to store processed images.

The final website design is our final product deliverable and it allows users to view their probability of contracting cancer through their medical images. As with predictive modelling, our team is expected to delegate equal parts of contribution into building a working website using an application called WordPress. Jack takes responsibility for building the accessibility features of the website's homepage which includes buttons that allow users to access their results.

As most of the coding work would take place after the semester break, each of us would take assumed responsibility into learning the basics and touching up on ideas that may seem foreign to us - image processing methodologies, coding capabilities and learning PHP frameworks. This guarantees that by the start of next semester, we would be able to start project coding immediately. It is also each of the members' responsibility to report any bugs and errors as well as providing constructive feedback and recommendations towards the project.

4.4. Project Management Process

4.4.1. Risk Management

Risks are an inevitable part associated with every project. It is therefore important for us to acknowledge the potential risks from the current project and try our best to minimize its chances of happening. To sum up the risks associated with our online tool for cancer prediction, we have attached a Risk Register (see Appendix E) that highlights descriptions, categories, triggers, root cause, etc. associated with each potential risk. The risks mentioned in the risk register have been grouped into different categories such as people risks, technical risks, resource risks, operational risks, etc. As our project deliverables can be sub-sectioned into 2 major parts - predictive model and website, the main risks associated with our project would mainly relate to unsuccessful building of the predictive model, website building and website hosting. Alongside each of these risks, we have determined their root causes and potential responses to solve them in case they arise.

Some of the notable risks mentioned in our risk register involve developing a low accuracy predictive model. Providing inaccurate results to our customers could halt their prevention attempts towards curing cancer and it leaves our team at an organisational risk as

we could get sued over customer's false results. This increases the team's responsibility to develop an accurate predictive model that could - to a certain extent - predict the likelihood of contracting cancer in a patient's cells to avoid mishaps involving lawsuits. Another prominent risk that could highly disrupt our team environment is when team members decide to leave the project as it leaves a waterfall effect on the remaining team members with concerns on whether or not they would be able to successfully complete the project. It greatly increases the remaining member's responsibility especially with pressures and unmet expectations from stakeholders that could further degrade the quality of the predictive model and website made. As a result, it is extremely important for each of us to minimize the chances of these risks happening and provide a backup plan in case the team comes across these risks.

4.4.2. Monitoring and Controlling Mechanisms

4.4.2.1. Stakeholder and Communication Plan

To ensure maximum productivity within each member of the team, it is important to have effective communication both between the team members and associated stakeholders to ensure awareness regarding the project trajectory and collaboration as a "single-entity" instead of multiple, weaker entities. To do this, our team plans to conduct 2 x 2-hour meetings per week to discuss objectives and expectations to be met during the week. In addition to this, we'd also include a weekly report consisting of the tasks undertaken, critical reflections, each team member's assigned roles and responsibilities, accomplishments as a team and plans to delegate in the upcoming weeks. This report would be kept as a Google Doc and during each meeting with our stakeholders each week, we would update them as well as ask for feedback depending on where we are at the project timeline. Communication within the team has been settled through Messenger and weekly group meetings are conducted using Zoom. If in particular weeks our team is unable to complete the task and require some extra help, it is our responsibility to attend extra meetings and consultations with the stakeholders to ensure that each mishap is solved before moving forward on the next part of the project. To further summarize, a communication table (see Appendix B) is used to explain reporting and communication plans for the project.

4.4.2.2. Review and Audit Mechanisms

The agile approach to our project gives us the opportunity to constantly review and monitor existing mechanisms in our project. It is therefore important for us to keep track of existing codes - regardless whether they're working or not, and constantly build on them to

ensure successful implementation. Quality assurance such as maintaining accuracy levels to be > 95% reduces the risk of providing inaccurate information to our customers and ensures greater customer satisfaction. With each version of the predictive model created, it is important to note each version's pros and cons including their accuracy levels to ensure that preceding models created are in better quality than previous versions. To do this, our team would create a specific GoogleDoc to document each bug encountered with the program with details such as the date encountered and date the program was cleared of the bug.

4.5. Schedule and Resource Requirements

4.5.1. Schedule

Effective schedule management is crucial to avoid schedule issues, avoid conflicts between team members as well as to ensure the success of our project. To ensure that all our team members completely understand all work required, we have specified activities that the team members must perform, document the relationships between the specified activities, as well as estimated activity duration needed to complete each activity. The project is divided into four major parts which are initiation, planning, execution, and close out. These parts are then broken down into smaller parts. To further explain what was mentioned, we have developed a Work Breakdown Structure (WBS) (see Appendix C) that lists all activities or tasks together with a Gantt Chart (see Appendix D) that displays our project activities including their corresponding start and end dates.

4.5.2. Resource Requirements

To reduce the chances of project failure, it is important to keep track of required resources as well as look into opportunities to better allocate them in cases of inefficiencies. As part of our deliverables, our main resources would be applied into the building of both the predictive model and website. Preliminary learning of deep learning associated with image processing methodologies could take up to 200hrs of online learning through websites such as Datacamp and stackoverflow as most of us have no prior knowledge on these methodologies. Successfully applying these foundational knowledge would allow us to code required infrastructures deemed necessary to create a successful predictive model. This also allows our team to allocate more time and resources onto enhancing the initial predictive modelling through intensive hours (approximately 100 hours) of testing and training on given datasets. Delaying these foundational learning could serve as a bottleneck towards the next

stages of the project - particularly in website development. In addition to this, our team's unfamiliarity with the Themosis framework fostered by our chosen application to website development - WordPress would also require additional learning hours (approximately 50 hours in total) in order for our team to be familiar with the application. Other minor software requirements such as finalising our choice on the different programming environments used - a debate on whether to use Python, R or MATLAB and deep learning methodologies such as Convolutional Neural Network (CNN) and its associated models such as alexnet, resnet18, and vgg19 could take up to 50 hours of discussion - identifying each methods pros and cons and choosing the best relating to our project specifications. Maintenance requirements such as preserving the integrity of our source code, detecting bugs and reducing downtime for both the predictive model and website helps to ensure that our deliverables are able to run smoothly. This is expected to take at least 200 hours of time.

Another important aspect of this project is the allocation of hardware resources required for computing, storage, and networking parts of the project. As our related dataset requires us to conduct the project using image processing methodologies, it is therefore necessary for team members to have a relatively high performing CPU and GPU in order to post fast training and testing times that could in turn enhance the prediction accuracy. Alongside this, discussion of storage and networking requirements have mostly taken place and it is expected that we do not take more than 10 hours into finalising our decisions on hardware specifications.

Last but not least, human resources requirements remain an important aspect in contributing to the success of the project as it relates to our personal hard and soft skills. For this project, we have a total of three members - each of us occupying an equal role contribution to the project, and a supervisor who will foresee our project timeline and report on us such as by providing recommendations and feedback on each minor project submission.

5. External Design

User Interface

- User Interface Design (Website Design)

The User Interface Diagram is created (see Appendix F) by using a website called “Figma" and it represents the overall design of our website. Figma’s user-friendly interface allows us to easily create high fidelity prototypes that can clearly depict what our final product would look like - ensuring it is within the scope of our project and expectations of both stakeholders and customers. The diagram is taken from our project design submission earlier. On the top right of the website, there are 4 buttons and each of them brings the users to different pages. The about button brings the user to a page where it explains everything about the website including what it can do and developer’s name and contact numbers. The help button serves as a manual to help users to understand how to use the website. The Login button allows users to login to their account, if they did not sign up yet, the sign up button allows them to sign up. In addition to the customer's upload of cancer medical images, there are 4 questions that patients have to fill which are the ages of the patient, history with cancer, disease symptoms and history of medical treatment. These questions greatly increase the accuracy of the prediction model and avoid miscalculation. As part of our requirements to make the website user-friendly, it is therefore vital that texts, buttons, font sizes, typefaces and colour scheme of the website is clearly maintained and accounted for users who may experience disabilities such as color blindness.

- Backend flowchart

A backend flowchart (see Appendix G), created with Lucidchart, allows us to visualise the back-processes and sequence of activities happening at our website visually. These processes are made in-line with our UI Interface design as the processes shown should have an implementation on the design. The diagram is also taken from our project design submission earlier. Users will be required to login to be able to use the cancer prediction functionality in the main page of our online tool for cancer prediction. Users will be able to sign up if they are not registered yet. Users will also be able to see what our online tool is about as well as ask for help if they need to. In using the predict functionality, the users will be required to fill in their details and questions shown in the user interface design, upload their medical image, and click the predict risk button which will then start the risk prediction

process. The risk prediction process includes image pre-processing, classifying or matching the pre-processed image to cancer or no cancer, and obtaining the prediction result. The result along with the user's pre-filled details will then be displayed to the users.

External Packages

To ensure that our website runs smoothly and efficiently, it is therefore important for us to integrate application programming interfaces (API) into our web application to ensure regulatory compliance. An API helps us to simplify integrations between different parts of softwares, allowing us to switch between different screens, views and commands whose processes make the application more user-friendly. Software APIs also provide the means for our web application to be integrated with other third party applications such as through emailing features with Gmail and removes the possibility of operating our web application as a standalone tool.

Some of the APIs that would be integrated into our current system of works are, but not limited to, an emailing tool like Gmail, a file management tool like Google Drive, an instant communication software like Messenger and Zoom. A task list management tool like Trello and coding software like Github would allow our team to better organise our work and reduce the redundancies taken in updating tasks in all of our project management tools. In addition to this, we have chosen Trello as our designated project management API as Trello has provided us the ability to integrate scrum into our project management as we are using an Agile approach in this project. Amazon API Gateway is another API that we have considered to use as it helps us to publish, maintain, monitor and secure REST, HTTP and WebSocket APIs at any scale. This, combined with the fact that we are using AWS to host our website makes it a good choice for us.

Performance

Both Website and Predictive Performance of the model are key factors in determining our customer satisfaction. This means that prioritizing the reduction of delays in both our predictive model performance and website load time is essential. Our customers would be more inclined to use our website if they are easy to use and provide fast response times. These would in turn increase incoming traffic at our website and raise visibility as opposed to competitors. To achieve this, there are several guidelines that we would be implementing to

decrease website load times. As reported by Google, most websites take around 7-9 seconds to fully load their website landing pages including pictures, spreadsheets, etc. We will be using this as a benchmark for our website load times. As mentioned above, we have chosen AWS services to host our website. In particular, we plan to have a dedicated server in AWS to ensure reliability and robustness in handling massive amounts of data and traffic. It is also important for us to minimize and combine smaller entities of data into a single large entity. In our web application, WordPress, it can be done simply through installation of an internal plug-in. This reduces clutter and minimizes redundant HTTP requests that may contribute to lag experienced in website loading times.

6. Methodology

Toolset and Approaches

Programming language and libraries

The advancements in deep learning associated with image processing methodologies have broadened our choice of programming environments - Python, R, MATLAB and its associated libraries such as neuralnet in R and scikit-learn in Python. For the purposes of this project, the programming language chosen for this project is Python. Our team has chosen this programming language due to the fact that all our team members are familiar with it and have sufficient background knowledge of the language. Python's readability and dense syntax allows for easy understanding of algorithms and it is a preferred, server-side programming language suitable for integration on websites. To create a deep learning algorithm to train on the given tumor images, our team will be utilizing a convolutional neural network (CNN) with deep residual learning obtained from the package resnet18. resnet18 is chosen as it has a shorter training time, provides incredibly accurate observations and reduces the risk of overfitting the model. Other libraries that we are going to use for the predictive model include, but not limited to, Matplotlib, Seaborn, Numpy, Keras, Scikit-learn, PyTorch, TensorFlow, etc. These libraries mentioned above serve as a foundational stepping stone that will enable our team to create interactive visualizations including informative statistical graphics for users, process data, training and testing our model that would eventually lead to successful completion of a deep learning algorithm.

Version Control System

As our project timeline spans across six months, there would be significant changes and updates to our initial source codes. To keep track of changes across different versions, our team has chosen GitHub. Due to the distance constraint that exists in our team, most of individual and group work would be done virtually so keeping our data safe and accessible between us is of utmost priority. Github allows for ease of collaboration through its pushing and pulling mechanism that can be easily accessed through each of our individual devices. Github's open source nature allows us to keep clear documentation of our codes and serves as a secure online repository to reduce the risk of losing source code. Another key aspect of Github is that it allows for integration between Amazon Services and Google Cloud who can provide detailed reports of customer feedback obtained from the website. However, a big

drawback of Github is its public nature, increasing breaches of integrity from opposing competitors who may plagiarise our source code without providing the required documentation.

Data Management

Efficient management of data should be optimised in order to ensure smooth operations in our deliverables. Included in our initial project brief is a dataset introduced by Kather, Jakob Nikolas. (2019) and available on (https://www.kaggle.com/joangibert/tcga_coad_msi_mss_jpg). These contain sample images of gastrointestinal cancer that serves as our main source of data utilised for both training and testing of the model. Our dedicated data management storage application would be Google Drive. As university students, we are given unlimited amounts of storage that would allow ease of image storage. Platforms like Google, Github and AWS allow for smooth integration to ensure data loss is minimized through their associated APIs. In addition to this, a database system - preferably MySQL infrastructure would allow for seamless storage of customer's privacy data including their login credentials that would be used to access their personal records. As customers would also upload their medical images of cancer tumour onto the website, this serves as a secondary dataset that could be used to further enhance the predictive accuracy and efficiency of the current predictive model. All of these would be integrated and encrypted into our dedicated Google Drive storage.

Algorithm Used

To create the predictive model using deep learning, our team will be utilizing convolutional neural network (CNNs) - in particular, making use of resnet18, a residual learning algorithm that can help classify MSI (microsatellite instability) versus microsatellite stability given in the dataset. Training of the model would be done on the normalized histological images provided in the given dataset. We would be utilizing transfer learning techniques to classify tumor versus normal tissues provided in each image.

High-Level Pseudocode

To sum up the algorithm explained above, from each image in the dataset, we would manually identify the tumour unless explicitly stated. Then, a neural network would be created to classify tumor cells versus normal cells. The regions of tumour cells would be

tessellated, color-normalized and sorted into MSI (Microsatellite Instability) and MSS (Microsatellite stability). At this point, we would train on each of the MSI and MSS sorted images using a residual learning algorithm in convolutional neural networks (CNNs). The resulting images would then be used to predict cancer occurrence within patients. To conclude, we will use resnet for feature extraction, after that, we will separate the data into a 70:30 training and testing set. Then, we will use either DT/SVM/ANN to train the model and evaluate using AOC.

Pre-processing Steps

The dataset used in this project that is introduced by Kather, Jacob Nikolas (2019) as mentioned in the previous section has been pre-processed as follows:

- Images have automatic detection of tumor
- Resized to 224 px x 224 px at a resolution of 0.5 μm/px
- Color normalization with the Macenko method
- Patients are assigned to either "MSS" (microsatellite stable) or "MSIMUT" (microsatellite instable or highly mutated)
- Reformatted to JPG format

Using the pre-processed images, feature extraction will then be performed. Then, the dataset will be divided into 70% training set and 30% testing set.

7. Test Planning

To ensure that our deliverables are up to the standards and satisfaction of both our stakeholders and customers, it is vital to conduct testing over a ranged period of time to ensure smooth operation of our deliverables. As mentioned above, our project mainly has two major parts - predictive model building and the final website. For our predictive model, following successful training on the MSI and MSS images of cancer tumour cells, the resulting model would be used to test on untouched images in the dataset to test its predictability on determining whether the specified images contain tumour or not. Our initial dataset is divided into 70% training and 30% testing to evaluate the model's performance. Some of the tools we are planning to use to measure the predictive accuracy of our predictive models are through receiver operating curves (ROC) to determine optimal cut off levels for each of the images. Developing ROC curves would also lead us to predict classifier performance using the model's area under the curve performance (AUC). In addition to this, a concordance index (c-index) could be used to further grade the predictive accuracy of the predictive model. Other various models including generating a confusion matrix is useful to measure the accuracy, precision, recall, specificity of the developed model. Throughout rigorous testing, our team aims to minimize the loss parameter involved in classification and ensure that all parameters are optimized with minimal error.

Our website would also have to undergo testing such as confirming the reliability with AWS's hosting system, that could handle mass traffic with reduced website downtime. Integrating the predictive model onto the website infrastructure is another challenge, as in both scenarios it should display and provide the same, accurate results. Ensuring customer's private information is kept within integrity and removing the risks of data breaches is another key aspect in testing the website.

8. Conclusion

To sum up, cancer is one of the leading causes of death around the world. Early diagnosis of cancer can help save a lot of lives since effective treatment can be done when cancer is identified early. This motivates us to be able to mitigate the risks of late detection of cancer especially since doctors are now demanding better methods for cancer diagnosis. The main objective of this project is to be able to successfully predict the early stages of gastrointestinal cancer using a predictive model and deploy this model as an online tool in a website for health institutions to use by utilizing machine learning and deep learning methodologies.

At the completion of this project, our project team expects to deliver a business case, a weighted scoring model, a scope statement, a requirements traceability matrix, a risk register, a final project report and presentation, as well as our predictive model software code, research report, online tool (website) for cancer prediction, online tool design document, online tool software code, and a user's manual.

Our cancer predictive model and online tool have their unique characteristics. These characteristics can be divided into functional requirements, which includes programming skill and finding suitable and relevant datasets from reputable sources to build our predictive model, as well as non-functional requirements, which includes meeting stakeholder expectations and creating a user-friendly user interface for our online tool. All of which are included in the requirements traceability matrix. Drawing on from these requirements, we can produce user stories that mimic what a customer would expect out of our final online tool, such as being able to obtain their results accurately and through relative ease. These user stories make up the acceptance criteria of this project.

Agile project management approach is chosen for this project since it is operated in rapid cycles and actively uses feedback that allows our team to adapt to changing requirements. This approach also emphasizes customer collaboration which will result in higher customer satisfaction. Using this approach will also reduce the risk of our project completely failing as testing is regularly done.

Regarding our project responsibilities, our team plans to delegate tasks equally between the team members for the predictive model by working together through GitHub. Our team also plans to distribute equal parts of the website building task. Since our team members have no experience in image processing and building a website, each of us also has the responsibility to learn the basics of image processing and PHP frameworks. Additionally, each team member is responsible for reporting bugs and errors during the development of both the predictive model and website.

To minimize chances of risks happening, it is beneficial for our team to recognize potential risks in our project. This includes people risks, technical risks, resource risks, management risks, operational risks, scope risks, etc. Therefore, we have created a risk register that notes all of the potential risks that may arise as we are developing our predictive model and online tool. Some notable risks include, but are not limited to, having low accuracy on the predictive model, not being able to fulfil customer's requirements and expectations, as well as failing the whole project due to lack of coding capabilities.

It is important for our project team to have good communication between each other and our stakeholders to gain optimum productivity. To ensure this happens, our team will be conducting 2 hour meetings twice a week and produce a weekly report that will highlight the tasks undertaken, roles and responsibilities, as well as plans for upcoming weeks. One meeting is allocated specifically for any discussions with stakeholders. Additionally, our team will also communicate via Messenger for any urgent concerns. Ensuring bugs are kept to a minimum and regularly checked, our team aims to document any bug encountered in a timely manner, keeping high accuracy from the predictive model as our main priority in audit mechanisms.

To avoid schedule issues and conflicts, and ensure success of this project, all team members must understand all the work required and its duration completely. To achieve this, our team has created a Work Breakdown Structure (WBS) to list all of the activities as well as a Gantt chart that shows the activities mentioned in the WBS together with their start and end dates.

Keeping track of required resources is also important for our team to reduce risks of project failure. This includes learning image processing methodologies which could take up

to 200 hours, improving the initial predictive model that may take around 100 hours, learning basics of the frameworks required for building our website which will take about 50 hours, around 50 hours of discussion on choosing different methodologies to apply, as well as at least 200 hours of maintenance to keep everything running smoothly.

In designing the final product of this project, our team has come up with 2 design interpretations. First, the user interface design that we created using Figma. This prototype exhibits what our end product would look like to the users. Secondly, we have constructed the backend flowchart to show the sequence of activities happening in the website. To ensure our website runs smoothly, we will integrate APIs into our website that will allow us to make our website more user friendly. As performance of our website is of utmost priority, we have chosen AWS services to host our website to achieve a highly responsive website.

For this project, our team has chosen Python, along with its libraries, as our programming language since the team members have sufficient knowledge of this language. We have also chosen GitHub as our version control system to be able to easily collaborate as well as keeping our data safe and accessible between the team members. As university students, we are granted unlimited storage space on Google Drive, so our team has picked this platform as our data management storage application to be able to store the medical images for our training and testing set as well as other data like our source code conveniently. In addition, we would need a database system, namely, MySQL, that would allow us to store our online tool user's login information. Furthermore, we will be utilizing CNN to classify our dataset as MSI or MSS. To pre-process our data further from the pre-processed dataset, we will perform feature extraction on the images then divide it into 70% training set and 30% testing set.

With regard to test planning, we have two components to test, that is, our cancer predictive model and our website which will be the platform in which the predictive model is deployed. Firstly, to test our model, we will be using the 30% testing set in order to evaluate our model's performance. This includes measuring the model's are under the curve (AUC) derived from its receiver operating curve (ROC), measuring concordance index (c-index) to further evaluate the accuracy of the model, as well as producing a confusion matrix as a means to measure the model's accuracy, precision, recall, and specificity. The testing will be done with the aim of optimizing all parameters used so our model has minimal error. Lastly,

testing our website will be done by verifying that our website is able to handle mass traffic with reduced downtime, ensure that we provide the same results in our website as the output of our predictive model alone, as well as ensuring user's private information are kept confidential.

9. Appendices

Appendix A

Requirements Traceability Matrix

REQUIREMENTS TRACEABILITY MATRIX					
Project Name:	**Data Mining Technique To Detect Cancer Using Predictive Modelling**				
Project Manager Name:	**Vionnie Tan**				
Project Description:	**Building a predictive model to determine early stages of cancer**				
ID	***Requirements (Functional or Non-Functional)***	***Assumption(s) and/or Customer Need(s)***	***Category***	***Source***	**Status**
001	Image Processing - Find relevant dataset that could be used for testing & training	Source of images are come from a legit source	Functional	Kaggle	In Progress
002	Programming Skills - Understanding topics regarding AI, Machine Learning, Deep Learning, Transfer Learning	Extensive knowledge of these programming skills increases the chance of efficiency of our predictive model	Functional	Online resources such as Stack Overflow, Leetcode, and Monash Units	Planning
003	User Interface - Allow login specific of the health institutions	Our stakeholders have to be able to easily access and understand the interface for them to use the predictive model	Non-functional	Project supervisor	Planning
004	Stakeholder expectations met	Accuracy of the Model has to be >	Non-functional	Stakeholders	Planning

		95%			
005	The software system should be integrated with health institutions's API	If we are building on an existing model, then our predictive model must be integrated with the current system.	Functional	Stakeholders	Planning
006	Identify important predictors that have significant impact on successful cancer categorization	Identifying important predictors will allow us to increase model accuracy	Functional	Project supervisor	Planning
Documentation: (Include any justification and assumptions made)					

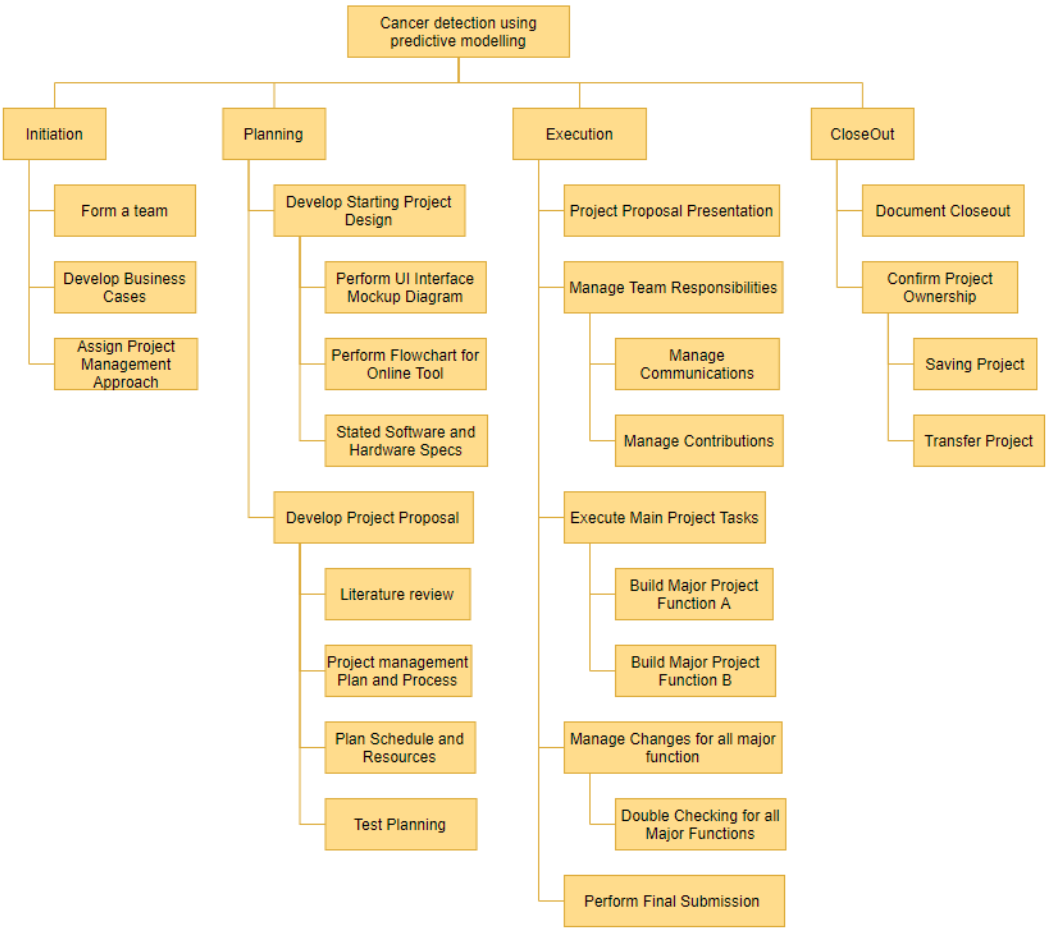
Appendix B

Communication Table

Stakeholders	Document Name	Document Format	Contact Person	Due
Project Supervisor	Progress report	Soft Copy via Google Doc and Zoom meetings	Afsaneh Koohestani	Weekly
Internal Management	Weekly status report	Soft copy and Zoom meetings	Jack Ooi, Elaine Liong, Vionnie Tan	End of week
Internal Management	Daily communication	Communication via Messenger	Jack Ooi, Elaine Liong, Vionnie Tan	Daily

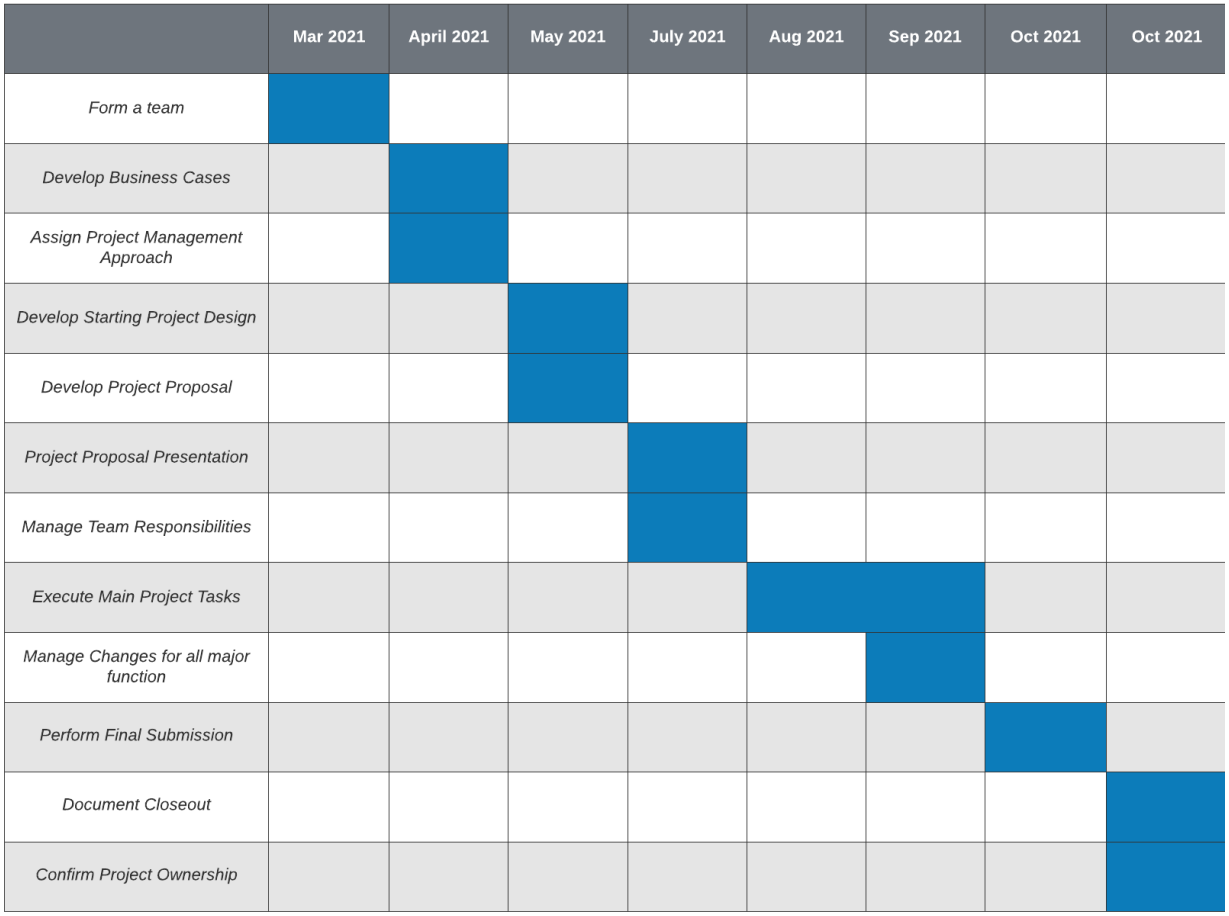
Appendix C

Work Breakdown Structure (WBS)



Appendix D

Gantt Chart



Appendix E

Risk Register

Risk Register

No.	Rank	Risk	Description	Category	Triggers	Root Cause	Potential Responses	Risk Owner	Probability	Impact	Status
1	6	Losing team members	Team members leaving the team	People risk	A team member decides to leave the team	Team member's personal issue	Consult project manager, redefine task responsibilities for each remaining team member	Team	5%	High	Potential
2	9	Team members unable to contribute	Team member not able to complete their task responsibilities	People risk	A team member encounters some issues that affects their work	Team member's personal issue	Consult project manager	Team	10%	Medium	Potential
3	8	Slow decision making / Project	Indecisive and not prioritising the success	Management risk	Clash within personal interests of team members and unsuccessful	Lack of open-mindedness and clarity within team members	Conduct internal meetings with team members and settle on a middle ground.	Team	15%	Medium	Potential

		Conflicts	of the project rather for personal gains. Unclear of project objectives and requirements.		understanding of given tasks.						
4	4	Delay in completion of earlier phases of project increasing failure of project completion	Not enough time to meet the schedule target to complete the project	Resource risk/ Schedule Risk	Team member does not deliver task responsibilities on time	Time estimated for a certain task is not enough	Revise schedule estimates	Team	20%	Medium	Potential
5	1	Predictive model has a low accuracy	The predictive model developed produces wrong	Technical risk / Performance Risk	When users inputs a medical image to the predictive model	Predictive model not properly developed	Reidentify important predictors for the model, do more research on algorithms used in developing predictive model	Team	30%	High	Potential

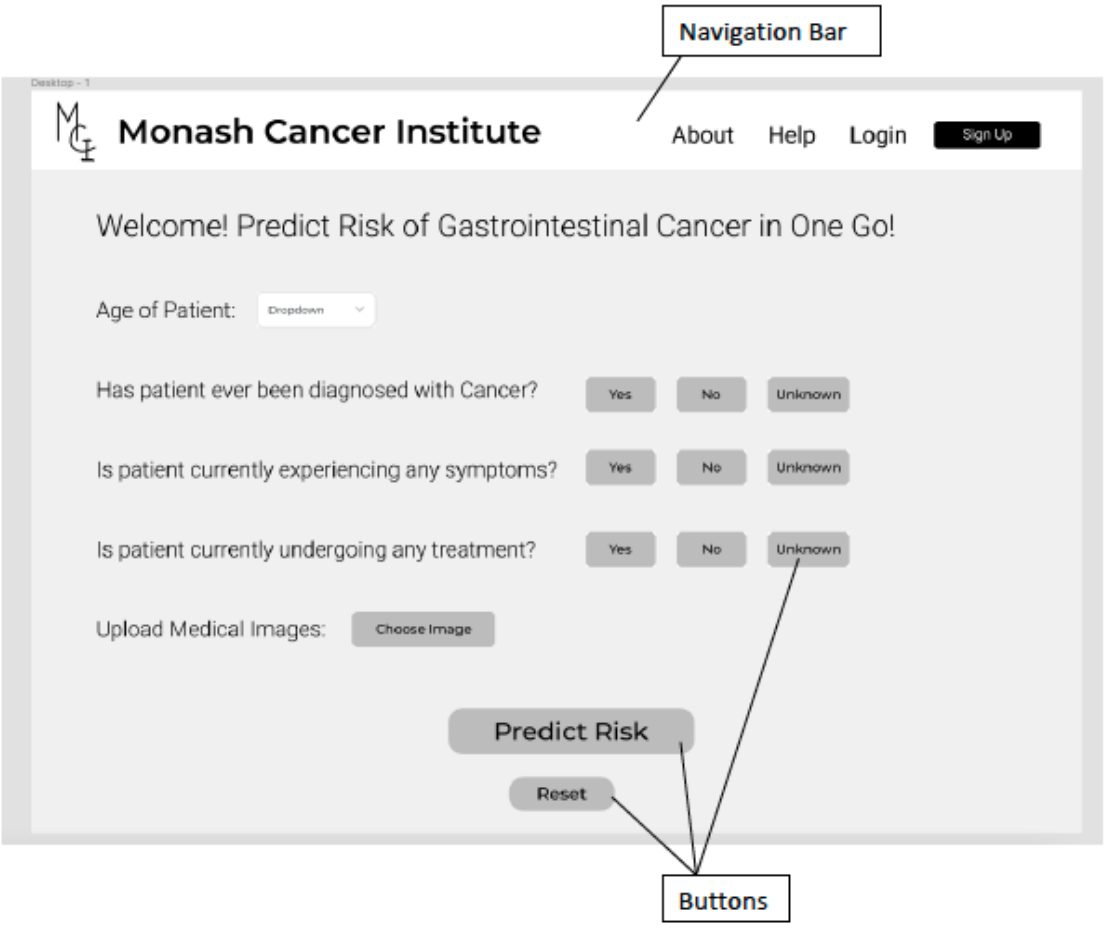
			outputs most of the time								
6	5	Losing source code	Source code is deleted and unable to be recovered	Technical risk	Source code accidentally deleted	Improper storage of project source code	Use GitHub for backup to minimise the risk	Team	5%	High	Potential
7	7	Website downtime	The platform that we decide to host our website in is unreliable and can't handle major traffics	Operational Risk	Could be caused by massive traffic or overall unreliability	Not choosing a good platform to host the website	Researching on pros and cons of several website hosting platforms and choose the platform with the least risk	Team	10%	Medium	Potential
8	10	Slow Stakeholder Actions that delays overall project completion	Poor communication with stakeholders and lack of verbal support from stakeholders		Attempting to have Communicative Measures with stakeholders but stakeholders remain unresponsive	Stakeholders are difficult to get hold of	Conduct regulatory meetings and emails to stakeholders. Have a stakeholder communication plan and update it accordingly. Make sure stakeholders are updated through every changes in the	Team	5%	Low	Unlikely

							project				
9	11	Scope Creep	Addition of unnecessary extra functionality not originally stated in the project scope	Scope Risk	Wanting to add new functionalities to enhance user experience but not addressing triple constraints of project	Excessive ideas given by team members	Clearly and succinctly state the requirements and scope in the project proposal. Update business case in case of changes	Project Manager / Team Members	5%	Low	Unlikely
10	12	Incomplete project design and deliverable definition	Not following preliminary designs created and risking to create a whole new environment that may have different deliverables	Technical Risk	Unclear preliminary designs that may not align with project scopes and definitions	Failure in understanding the given project and what is deemed an appropriate design and deliverable.	Create a preliminary design of the website, and build on it accordingly at each phase of the project.	Team	5%	Low	Unlikely
11	2	Software does not	Customers are	Software Risk	Customers may not be inclined to	When customers are not satisfied	Ensure that a substantial amount of testing is		15%	High	Potential

		fulfill customer requirements	collectively unable to obtain, view, save their results due to difficulty in accessing the user interface of the website		use the website for their purposes due to it being obsolete and not user-friendly	with the product	done on the website				
12	3	Lack of coding capabilities leading to failure of the whole project	Lack of self-awareness in coding capabilities may result in our team abandoning the whole project	Technical risk	Lack of desire to understand the resources required to build the predictive model and website.	Lack of self-awareness on coding skills that may be exaggerated in order to	Being honest with current coding capabilities, and attending extra workshops/consultations both from Monash and online resources to ensure basic foundational knowledge.	Personal	15%	High	Potential

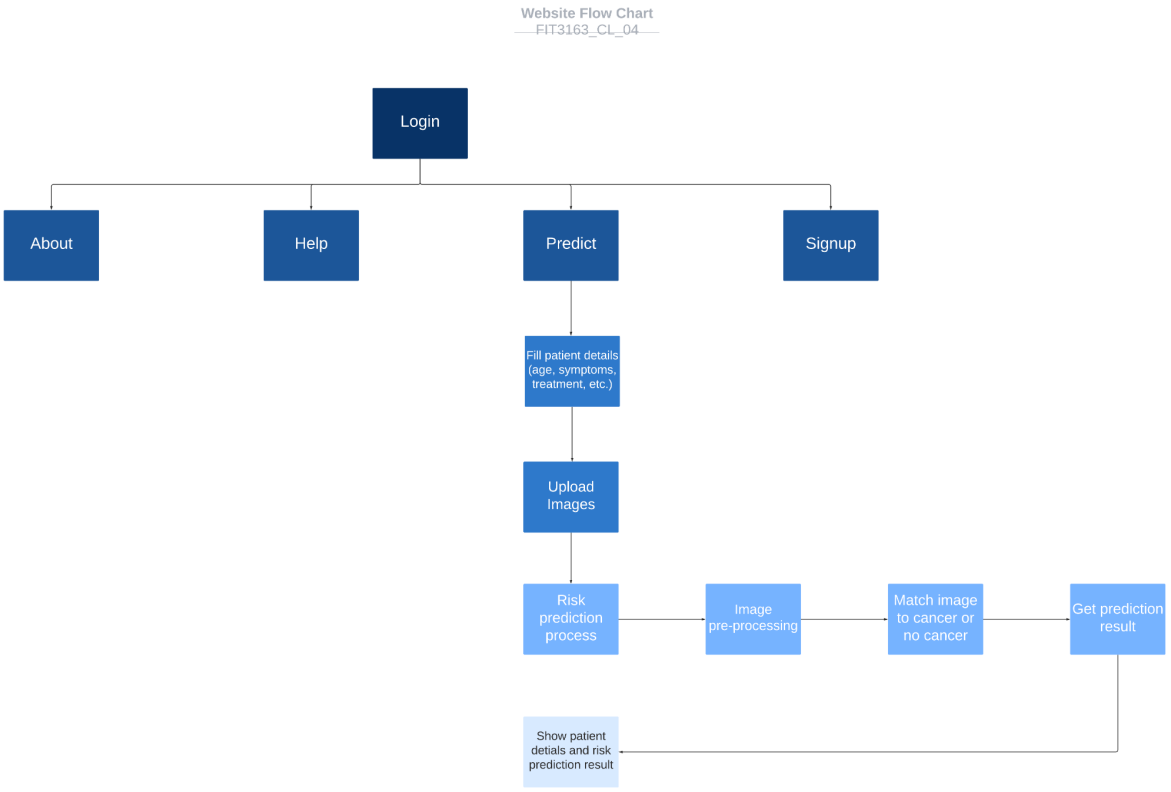
Appendix F

User Interface Design



Appendix G

Website Backend Flowchart



Annex

Team Member's Contributions	
Team Members	Contributions (Sections)
Elaine Liong	1. Introduction to Literature Review 2. Project Deliverables 3. Requirements Traceability Matrix 4. Process Model 5. Risk Management 6. Stakeholder and Communication Plan 7. Schedule 8. External Design - Backend Flowchart 9. Methodologies 10. Conclusion 11. References
Jack Ooi	1. Literature Review Content 2. Literature Review Conclusion 3. Gantt Chart 4. Work Breakdown Structure 5. User Interface Justification 6. Version Control System 7. Data Management 8. Requirements Traceability Matrix 9. Communication Plan 10. Project Responsibilities 11. References
Vionnie Tan	1. Introduction 2. Project Overview

	3. Project Deliverables 4. Functional and Non-functional requirements 5. Product User Acceptance Criteria 6. Project Organisation 7. Project Management Process 8. Resource Requirements 9. External Design - User Interface 10. Methodologies 11. Test Planning

10. References

Ansary, U, & Sarode, T. Skin Cancer Detection Using Image Processing. Retrieved May, 12, 2021, from https://d1wqtxts1xzle7.cloudfront.net/53555135/IRJET-V4I4702-with-cover-page.pdf?Expires=1620894864&Signature=DE~Uokb0e9tF5K6Dlv8FRj5kuA4u9nf0sz5q1vqpoYikD24WSZSI8SeWbFS7DPznq6vOoLolS2OG40dqqRhOgZVPV8LMtGQqL5YdRxbf5oMoFBurAhYpW22d6YEZuql7JjdQtWlJcvfggeA4jchsWpO9kCtFuzyb-gQDVW5~6bJZxRHzGW4JvSALlBWPfm4DmrDxKp5S3iYpCqSuUWbHVNP5Z6P2qfrPVTl9VunTZ19abjwqMwTbupdvYtN5u9biJEpjW0fX0KdktcEpYwYRuEB4sLw0p9Z0xC5LOx3a4P1wF6CBVtsxaLCDiBM4msv8uw4VjLO6OhEEiC5Q4njuYw__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

Australian Institute of Health and Welfare. (2020, Nov 13). *Cancer data in Australia.* Rankings. Retrieved March 29, 2021, from https://www.aihw.gov.au/reports/cancer/cancer-data-in-australia/contents/rankings

lom M, Z, Taha, T, M, Yakipcic, C & Westberg, S. (2018) The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches. Retrieved May 23, 2021 from https://www.researchgate.net/publication/323570864_The_History_Began_from_AlexNet_A_Comprehensive_Survey_on_Deep_Learning_Approaches

Cancer Council Australia. (n.d.). *Cancer Statistics in Australia. Facts and Figures.* Retrieved March 29, 2021, from https://www.cancer.org.au/cancer-information/what-is-cancer/facts-and-figures

Delen, D. (2009). Analysis of cancer data: a data mining approach. Retrieved May 12, 2021 https://onlinelibrary.wiley.com/doi/full/10.1111/j.1468-0394.2008.00480.x

Deng, Q., He, B., Liu, X., Yue, J., Ying, H., Pan, Y., Sun, H., Chen, J., Wang, F., Gao, T., Zhang, L., & Wang, S. (2015). Prognostic value of pre-operative inflammatory response biomarkers in gastric cancer patients and the construction of a predictive model. *Journal of Translational Medicine*, *13*(1), 66. https://doi.org/10.1186/s12967-015-0409-0

Demir, A & Yilmaz, F. (n.d). Early Detection of Skin Cancer Using Deep Learning Architectures: Resnet-101 and Inception-v3. Retrieved May 24, 2021. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8972045

Furey, T, S , Cristianini, N, Duffy, N, Bednarski, D,W, Schummer, M & Haussler, D. (2000). Support vector machine classification and validation of cancer tissues samples using microarray expression data. Retrieved May 24, 2021. https://watermark.silverchair.com/160906.pdf?token=AQECAHi208BE49Ooan9kkhW_Ercy7Dm3ZL_9Cf3qfKAc485ysgAAArswggK3BgkqhkiG9w0BBwagggKoMIICpAIBADCCAp0GCSqGSIb3DQEHATAeBglghkgBZQMEAS4wEQQMGkOu07HHTGFA51ofAgEQgIICbr5TBibsaoxzqIqgib_ZvoNNTo_ODdXl4vyeTMxFZzzaLfF_1tvQdVtJX8WsZyW6_BLbKJOokHB-p5gmO2RwQAcAHwtT2dJ5h-ic9yL2X0eUQZyvFIVsMWS1hkuHTJZPVFzmSmeqYm08R4Q1TyGwBcQY-bxWx7h85Gdv0NZtMD6AU2mK8FmveleTJ984vnMU2BYad4gl1zfTr0UM4ycM3JWSX1XJo412rHeFs2DWaMUEC_J7KOyMlMQBQJvnf6RGRW_vOuMhhwPwttl8siZA6rlF17rbeJIK_rGhdFk6wARp83Tq4_vfQ94Z3PFAYLz2th5WG4_E2E5LrKwjbHweqaYPAwTKoVRp9fFn3o3MsY9NzDkQlSUIpE8NWZXzoECozxinp7xJN7egiLimm5_59ESUiA6tGsu1_IHIlsiIY9acFUKYw9NMfvul6SwiPp3iltLVYFoyuBhQySTDZuXxCsOI_WDWti4frLpQcmP0LgB5cBGXba5UJAzQhy7uiV7E8o7cToAV18kwhyP5nzWNAN5rmd8N5zzMshX6lYPkOJPymtX1tFhyXtmUQORfCU219rn35zxtabzPbmyLQx37Xo5uWu-wA_FL5LDsh8N9G0tN44Ka8SVIlXJkUo2OgoNhxXRbNZJU0XLHxKWRncC5ZvvKEsVpZOn12BfdBgX_oi2ZJabV32NqNrzwae8DfLygQ1wroP2ttHMU8xYWTElNXB6FeYIAbj3ezOH1GPNcKrLfhAXpNiNE5s6bpwHt6q7vfY5y1FybFjgfSFdUxPaH_HqmLDKrz3D_jx48coy2t4PJyc6Lc2Q7QOvM1pYPEaI

Gupta, S., Kumar, D., & Sharma, A. (2011). Data mining classification techniques applied for breast cancer diagnosis and prognosis. *Indian Journal of Computer Science and Engineering, 2*(2), 188-195. https://www.researchgate.net/publication/268054343_Data_mining_classification_techniques_applied_for_breast_cancer_diagnosis_and_prognosis

Indolia, S, Goswami A, K, Mishra S, P & Asopa, P. (2018). Conceptual Understanding of Convolutional Neural Network- A Deep Learning Approach. Retrieved May 24, 2021 from https://www.sciencedirect.com/science/article/pii/S1877050918308019#:~:text=Convolutiona

l%20Neural%20Network%20(CNN)%20is,of%20traditional%20machine%20learning%20approaches.&text=This%20study%20provides%20the%20conceptual,common%20architectures%2C%20and%20learning%20algorithms.

Jain, S. Jagtap, V. & Pise, N. Computer Aided Melanoma Skin Cancer Detection Using Image Processing. Retrieved May, 12, 2021, from https://www.sciencedirect.com/science/article/pii/S1877050915007188

Kumar, S., Govardhan, A., & Srinivas, B. S. (2014). Data Mining Issues and Challenges in Healthcare Domain. *International Journal of Engineering Research, 3*(1), 857-861. https://www.ijert.org/research/data-mining-issues-and-challenges-in-healthcare-domain-IJERTV3IS10306.pdf

Munir, K., Elahi, H., Ayub, A., Frezza, F., & Rizzi, A. (2019). Cancer Diagnosis Using Deep Learning: A Bibliographic Review. *Cancers, 11*(9), 1235. https://res.mdpi.com/cancers/cancers-11-01235/article_deploy/cancers-11-01235.pdf

Serrador, P., & Pinto J. K. (2015). Does Agile Work? - A quantitative analysis of agile project success. *International Journal of Project Management, 33*(5), 1040-1051. https://www.sciencedirect.com/science/article/pii/S0263786315000071

Spanhol, F, A, Oliveira, L, S, Petitjean, C & Heutte, L. (2016). Breast Cancer histopathological image classification using Convolutional Neural Networks. Retrieved May 24, 2021, from https://ieeexplore.ieee.org/abstract/document/7727519

TCGA COAD MSI vs MSS Prediction (JPG). Retrieved May 19, 2021, from https://www.kaggle.com/joangibert/tcga_coad_msi_mss_jpg

Titoriya, A & Sachdeva, S (2019). Breast Cancer Histopathology Image Classification using Alexnet. Retrieved May 24, 2021, from https://ieeexplore.ieee.org/abstract/document/9036160

World Health Organization. (n.d). *Cancer*. Retrieved May 11, 2021, from https://www.who.int/health-topics/cancer#tab=tab_3