Accurate Cancer Prediction Using AI

Team Number: sdmay24-10
Client: Ashraf Gaffar
Advisers: Ashraf Gaffar, Ashfaq Khokhar
Team Members and Roles:

- Mark Hanson, Dev
- Eric Schmitt, Dev
- Christopher Tague, Dev
- Norfinn Norius, Client Communications
- Thriambak Giriprakash, Minutes and Admin
- Bishal Ghataney, Senior Engineer

Team Email: sdmay24-10@iastate.edu
Team Website: https://sdmay24-10.sd.ece.iastate.edu

Revised: 11/29/23 V2

Executive Summary

Development Standards & Practices Used

Agile software development patterns will be used. All code will go through Git/Google Colab, and will require approvals from another team member.

Summary of Requirements

- AI training model
- Cloud infrastructure for AI model
- Frontend to interact with model
- Backend to connect frontend to AI model

Applicable Courses from Iowa State University Curriculum

- Software testing SE 317
- SE 309 (API calls)
- COM S 319

New Skills/Knowledge acquired that was not taught in courses

- How to train AI with test data: In machine learning, training data and test data serve different purposes:
    - **Training Data:** This is the dataset used to train the AI model. The model learns patterns and relationships within this data.
    - **Test Data:** This dataset is used to evaluate the model's performance. It contains data that the model has not seen during training. The performance on this data helps assess how well the model generalizes to new, unseen examples.

When training an AI model, you don't use test data for training. Instead, you split your available data into three main subsets:

- **Training Set:** This subset, usually the largest portion of the data, is used to train the model.
- **Validation Set:** A smaller subset of the data used to fine-tune model hyperparameters and assess performance during training. It helps prevent overfitting by guiding decisions like when to stop training or adjusting hyperparameters.

- **Test Set:** Another separate subset, kept entirely unseen by the model during training and validation. This set is used at the very end to assess the final model's performance. It gives an unbiased estimate of how well the model might perform on new, unseen data.

The typical approach involves randomly splitting your dataset into these subsets. For instance, an 80-10-10 split means 80% of the data for training, 10% for validation, and 10% for testing.
To train an AI model:

- **Use the Training Set:** Train the model using the training data. Adjust the model's parameters to minimize the error between predicted and actual outcomes.
- **Use the Validation Set:** Fine-tune the model's hyperparameters (like learning rate, number of epochs, or architecture adjustments) based on its performance on the validation set. This helps prevent overfitting and ensures better generalization.
- **Assess on Test Set:** Finally, after fine-tuning and adjusting the model based on the validation set, use the test set to evaluate the model's performance. This provides an unbiased estimate of its performance on new, unseen data.

- How AI works to predict new data?

AI prediction involves using a trained model to make predictions or classifications on new, unseen data. Below is the overview of how AI typically works to predict new data:

- **Model Training:** During the training phase, the AI model learns patterns, features, and relationships within the provided training data. This involves adjusting its internal parameters to minimize the difference between predicted outcomes and actual outcomes.
- **Model Architecture:** The architecture of the model (like neural networks, decision trees, etc.) defines how data flows through it and how it learns from the input to generate predictions.
- **Feature Extraction:** The model identifies relevant features from the input data. These features can be various attributes, characteristics, or patterns that are informative for making predictions.
- **Prediction Process:**
    a. **Data Preprocessing:** New data, before prediction, needs to undergo the same preprocessing steps that the training data underwent. This ensures that the data format and features align with what the model has been trained on.
    b. **Prediction Phase:** Once the new, unseen data is preprocessed, it's fed into the trained model. The model then uses the learned patterns and relationships to make predictions or classifications based on this input.
- **Output:** The model generates predictions or classifications based on the input data. For example, it might predict whether a given medical image contains signs of cancer, forecast the price of a stock, classify an email as spam or not, etc.

- **Evaluation:** The predictions made by the model on the new data can be evaluated against known ground truth (if available). This evaluation helps assess the model's accuracy, precision, recall, or other performance metrics.
- **Iterative Improvement:** If the predictions are satisfactory, the model might be deployed for practical use. If not, this process becomes iterative—feedback from these predictions can be used to retrain the model, adjust parameters, or improve the data quality to enhance future predictions.

Table of Contents

List of figures/tables/symbols/definitions (This should be the similar to the project plan)

1 Team, Problem Statement, Requirements, and Engineering Standards

Problem Statement:

The project is trying to solve the problem of predicting the occurrence and recurrence of cancer. Despite advancements in medicine, cancer remains a significant challenge due to the limited ability to predict its occurrence and recurrence. This project aims to leverage artificial intelligence (AI) to provide more accurate predictions than human doctors alone. It will involve building and training a simple AI model for cancer prediction, with the goal of improving cancer treatment. The project will also provide training for students on AI, equipping them with skills to build and train a medical AI diagnosis tool using common AI tools and libraries like Tensorflow and Keras. The ultimate goal is to create an AI model capable of recognizing and predicting the risk of cancer occurrence and recurrence, which will be continually improved through training on multiple sets of data provided by leading research institutes and hospitals.

Requirements:

Learning Simple AI principles and tools

- Familiarize ourselves with Tensorflow and Keras' existing AI models
- Learning how the image classification or detection works
- Understanding the frameworks of convolutional neural networks and deep neural networks
- Exploring the concept of transfer learning and its application to enhance model accuracy

Construct a rudimentary AI model fit for our needs

- Find multiple models which are capable of processing medical data
- Select one to use the transfer learning on

Data Preprocessing Tools

- Use data preprocessing tools to clean, normalize, and transform raw data into a format suitable for model training

Train & Validate the Model

- Use transfer learning to train our model by providing data and expected outcomes. This will take time.

Security and Privacy

- Implement robust security measures to protect sensitive cancer-related data privacy regulations

Display the model

- Develop a UI interface which can be easily accessible for people to input their data and see results while enhancing user experience (UX)

Engineering Standards

Back end Development Language: Python

- Most AI libraries are python-based

Data Format: SQL

- SQL can be easily integrated into python code
- We have the most experience with SQL

Front End development Language: JavaScript/ReactJS

- Our team has the most experience with JS and ReactJS

General Software Standards

2 Clear and concise comments on code
3 Properly named variables
4 Files named properly
5 Files must be placed and go through our shared Git repository (Minimal local work)

1.1 Team Members

Eric Schmitt

Mark Hanson

Chris Tague

Norfinn Norius

Thriambak Giriprakash

Bishal Ghataney

1.2 Required Skill Sets for Your Project

AI development and training. AI infrastructure design. Backend development. Front-end development.

1.3 Skill Sets covered by the Team

Eric Schmitt - Full stack development

Mark Hanson - Full stack development

Chris Tague - Full stack development

Norfinn Norius - Full stack development

Thriambak Giriprakash - Full stack development

Bishal Ghataney - Full stack development, AI development

1.4 Project Management Style Adopted by the team

Agile/scrum

1.5 Initial Project Management Roles

Eric Schmitt - Development

Mark Hanson - Development

Chris Tague - Development

Norfinn Norius - Client Communications

Thriambak Giriprakash - Minutes and Administration

Bishal Ghataney - Senior Engineer

1.6 Problem Statement

The project is trying to solve the problem of predicting the occurrence and recurrence of cancer. Despite advancements in medicine, cancer remains a significant challenge due to the limited ability to predict its occurrence and recurrence. This project aims to leverage artificial intelligence (AI) to provide more accurate predictions than human doctors alone. It will involve building and training a simple AI model for cancer prediction, with the goal of improving cancer treatment. The project will also provide training for students on AI, equipping them with skills to build and train a medical AI diagnosis tool using common AI tools and libraries like Tensorflow and Keras. The ultimate goal is to create an AI model capable of recognizing and predicting the risk of cancer occurrence and recurrence, which will be continually improved through training on multiple sets of data provided by leading research institutes and hospitals.

1.7 Requirements & Constraints

Learning Simple AI principles and tools

- Familiarize ourselves with Tensorflow and Keras' existing AI models
- Learning how the image classification or detection works
- Understanding the frameworks of convolutional neural networks and deep neural networks
- Exploring the concept of transfer learning and its application to enhance model accuracy

Construct a rudimentary AI model fit for our needs

- Find multiple models which are capable of processing medical data
- Stitch them together in a logical sense in which there is a clear path for a set of data to follow from start to finish

Data Preprocessing Tools

- Use data preprocessing tools to clean, normalize, and transform raw data into a format suitable for model training

Train the Model

- Use transfer learning to train our model by providing data and expected outcomes. This will take time.

Security and Privacy

- Implement robust security measures to protect sensitive cancer-related data privacy regulations

Display the model

- Develop a UI interface which can be easily accessible for people to input their data and see results.

1.8 Engineering Standards

Backend Development Language: Python

- Most AI libraries are python-based

Data Format: SQL

- SQL can be easily integrated into python code
- We have the most experience with SQL

Front End development Language: JavaScript/ReactJS

- Our team has the most experience with JS and ReactJS

General Software Standards

- Clear and concise comments on code
- Properly named variables
- Files named properly
- Files must be placed and go through our shared Git repository (Minimal local work)

1.9 Intended Users and Uses

Who benefits from the results of your project? Who cares that it exists? How will they use it? Enumerating as many "use cases" as possible also helps you make sure that your requirements are complete (each use case may give rise to its own set of requirements).

The goal of the project is to advance research in the area of cancer prediction. This benefits researchers, hospitals, and laypeople. It gives researchers and hospitals a tool to utilize medical information and use it to predict cancer. It helps advanced researchers predict a layperson's risk of cancer using their basic medical data and gives them the ability to do extra screening and take precautions, and potentially save their lives.

A layperson could also input their own data (images, symptoms) to the algorithm and receive a response as to the likelihood of cancer. However, without proper medical data acquired through actual medical tests, any results produced by the AI would be similar to google searching your symptoms. There will be many false positives, and negatives if a layperson attempts to use this algorithm without proper medical data and counsel.

2 Project Plan

2.1 Task Decomposition

Tasks:

- 1. Get an AI model from Keras and Tensorflow

- 1.1 Search through the existing AI models and find the one that is the closest to the model we need

- 1.2 Edit the model to accept and use the data we will get from the client

- 2. Data Collection and Preprocessing

- 2.1 Clean, modify and extract the data to meet the project needs to make it ready for training

- 2.2 Divide the data into two sets, training and testing, using an 80-20 split

- 2.3 Address data privacy and HIPPA compliance when handling medical information

- 3 Train the AI model with data given by the client

- 3.1 Utilize GPU to speed up model training

- 3.2 Document the training process and any issues encountered

- 4 Test the AI model with data given by the client

- 4.1 Validate the accuracy of the model with the data given by the client

- 4.2 Maximize the accuracy and minimize the latency of the model using the transfer learning

- 5 User Interface and Visualization

- 5.1 Create a user-friendly interface for healthcare professionals to input patient data

- 5.2 Generate informative dashboards and reports to aid in decision-making

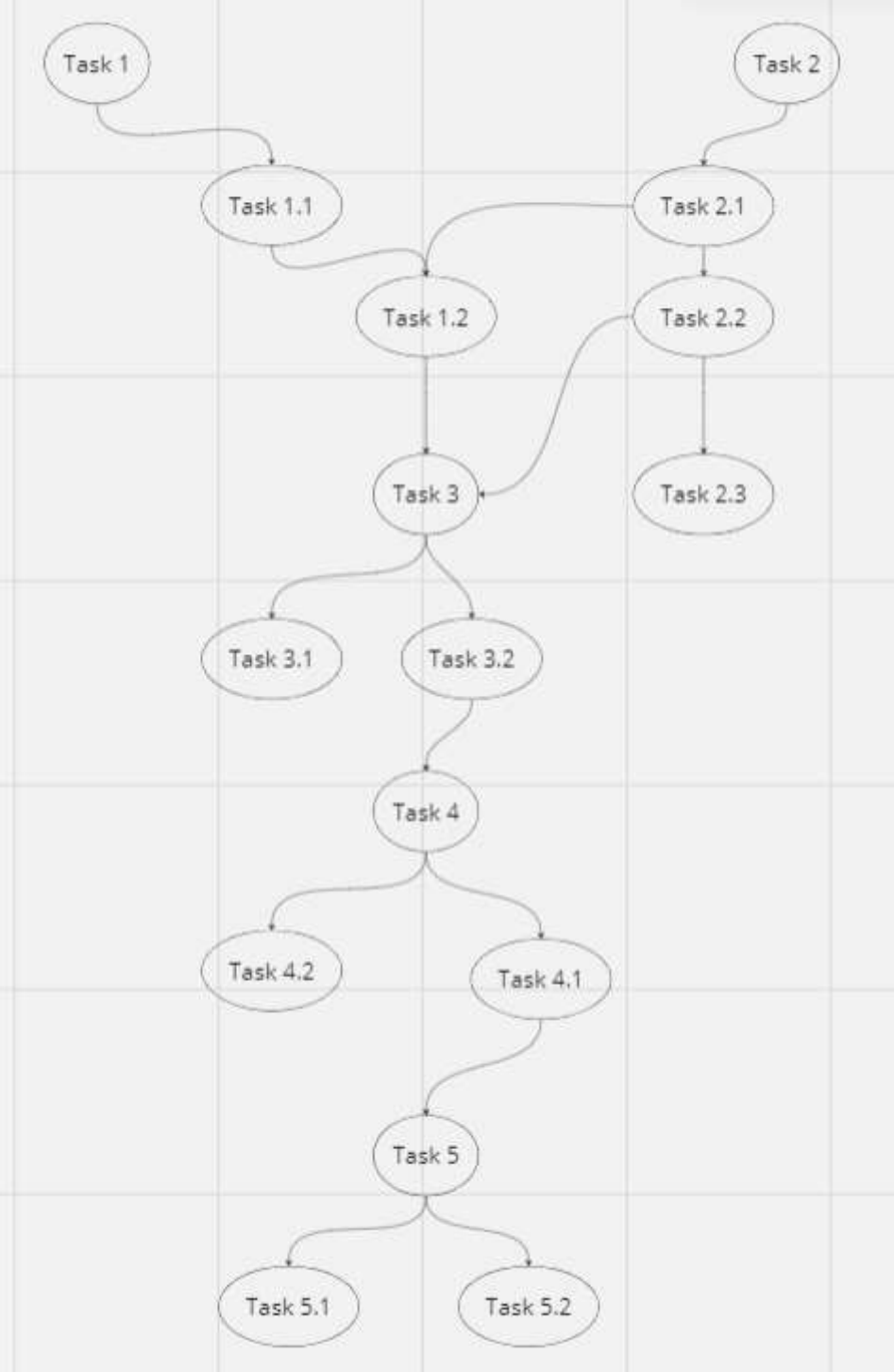


Image 1

2.2 Project Management/Tracking Procedures

We will be using the Agile project management style. Our goal is to develop an AI interface that accurately predicts the occurrence and recurrence of cancer. Our Agile management style will

allow us to ensure our AI is utilizing all the important factors for predicting cancer and is giving all the necessary feedback in a comprehensive way.

We will use the Gitlab issue board to track the progress of each part of the overall project. The Gitlab board will have 3 separate sections that we will move the respective story cards into as we continue through the development process. The sections will be “start” “in progress” and “done.”

2.3 Project Proposed Milestones, Metrics, and Evaluation Criteria

1: Construct an AI model which is capable of processing a data point from start to finish

- Understand how image classification and detection work.

2: Gather a proper dataset found from either our professor or from open sources, and make any changes necessary to prepare it to be used in our algorithm.

3: Train our algorithm with our data, the algorithm should classify with 90% accuracy with a pattern recognition of 100 ms. Test furthermore.

4: Build a proper Interface for the algorithm. With minimal lag and stress under high loads of processing. We will need to use cloud computing for this. AWS ELB might be of use here.

2.4 Project Timeline/Schedule

• A realistic, well-planned schedule is an essential component of every well-planned project

• Most scheduling errors occur as the result of either not properly identifying all of the necessary activities (tasks and/or subtasks) or not properly estimating the amount of effort required to correctly complete the activity

• A detailed schedule is needed as a part of the plan:

– Start with a Gantt chart showing the tasks (that you developed in 2.2) and associated subtasks versus the proposed project calendar. The Gantt chart shall be referenced and summarized in the text.

– Annotate the Gantt chart with when each project deliverable will be delivered

• Project schedule/Gantt chart can be adapted to Agile or Waterfall development model. For agile, a sprint schedule with specific technical milestones/requirements/targets will work.

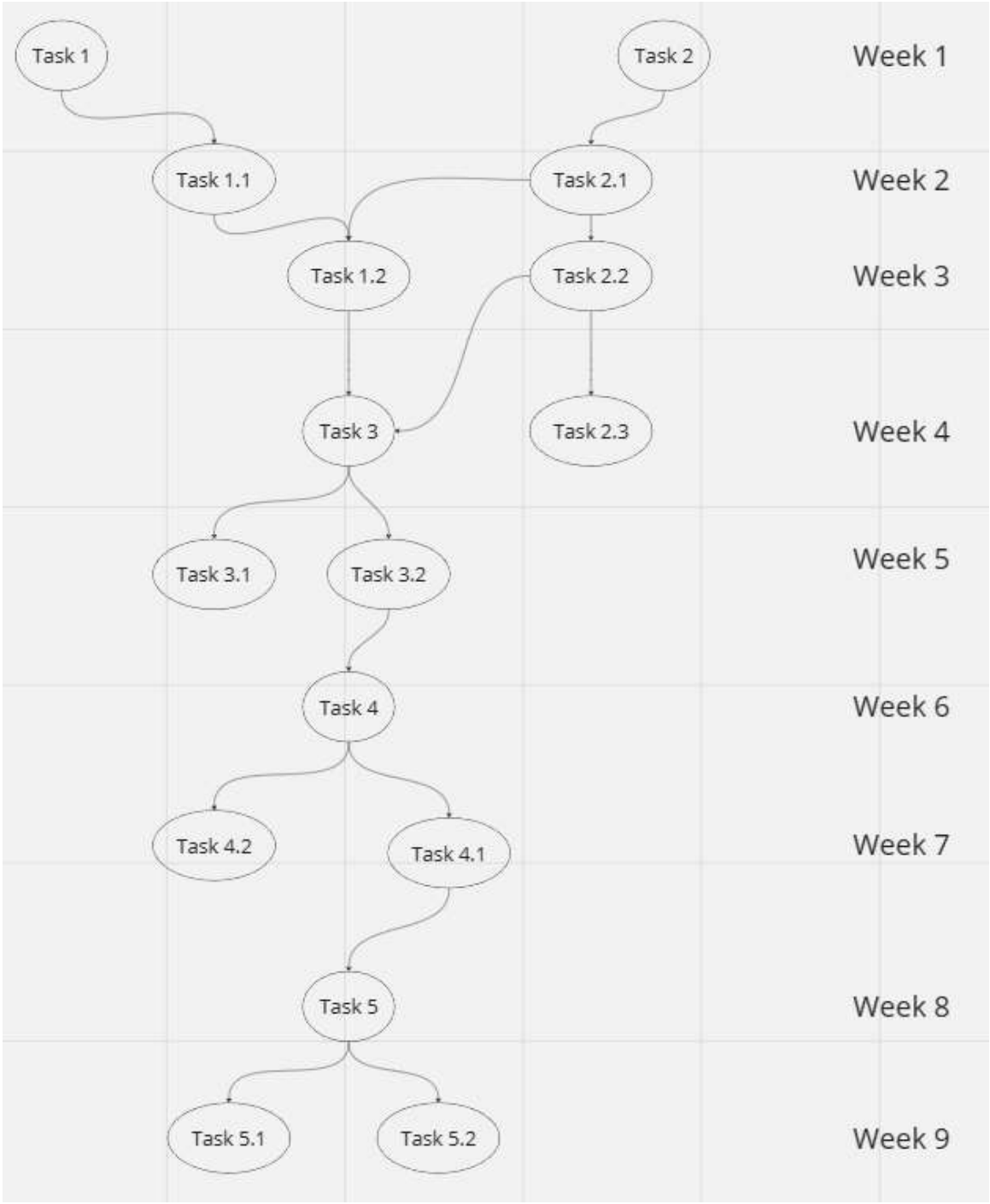


Image 2

*The above week deadlines might be subject to change due to unforeseen events

2.5 Risks And Risk Management/Mitigation

Learning how image classification or detection works

- **Risk:** Moderate risk due to complexity in computer vision. This moderate risk is due to the complexity involved in understanding image classification and detection techniques, which are essential for medical image analysis.
- **Risk Mitigation Plan:** To mitigate this risk, allocate more time for learning and provide support from experts in the field.

Understanding convolutional neural networks and deep neural networks

- **Risk:** Moderate risk due to the complexity of neural networks.
- **Risk Mitigation Plan:** Allocate more time for learning and provide access to relevant educational resources. Encourage team members to engage in hands-on practice with neural networks, possibly through coding exercises or small projects, to solidify their understanding.

Finding multiple models capable of processing medical data

- **Risk:** Moderate risk related to the availability and compatibility of models.
- **Risk Mitigation Plan:** Research and select models with a proven track record, consider custom model development if necessary. Conduct thorough research to identify existing models with a proven track record in processing medical data. Ensure compatibility by testing selected models with the project's data.

2.6 Personnel Effort Requirements

Names/Tasks	Thri	Chris	Eric	Bishal	Mark	Finn
Task 1	2.5hr	2.5hr	2.5hr	2.5hr	2.5hr	2.5hr
Task 2	2.5hr	2.5hr	2.5hr	2.5hr	2.5hr	2.5hr
Task 3	2.5hr	2.5hr	2.5hr	2.5hr	2.5hr	2.5hr
Task 4	2.5hr	2.5hr	2.5hr	2.5hr	2.5hr	2.5hr
Task 5	2.5hr	2.5hr	2.5hr	2.5hr	2.5hr	2.5hr

Task1: Get an AI model from Keras and Tensorflow

Task2: Data Collection and Preprocessing

Task3: Train the AI model with data given by the client

Task4: Test the AI model with data given by the client

Task5: User Interface and Visualization

We are expecting a fairly even distribution of effort across the board for all major tasks.

2.7 Other Resource Requirements

- Unlimited platform to train AI with a GPU
- Appropriate initial AI model to start with
- Keras and Tensorflow libraries
- Data set relevant to predicting cancer

- Server to host the website on

4 Design

4.1 Design Content

The project aims to leverage artificial intelligence (AI) to predict the occurrence and recurrence of cancer, with the ultimate goal of improving cancer treatment. It also seeks to provide AI training for students, equipping them with the skills necessary to build and train a medical AI diagnosis tool.

4.2 Design Complexity

The project's technical complexity can be justified by the following components/subsystems and their associated scientific, mathematical, or engineering principles:

1. Data Collection: This involves many csv files which have been provided to us from our client. Each csv file contains a linearization of an image of one cancerous cell along with a patient number and a survival rate in months.
2. Model Development: This involves the use of machine learning principles to build an AI model. It requires a deep understanding of algorithms, linear algebra, calculus, statistics, and probability.
3. Model Training: This involves the application of optimization techniques to train the AI model on the collected data. It requires knowledge of gradient descent algorithms, backpropagation, and other advanced mathematical concepts.
4. Model Evaluation: This involves the use of statistical measures to evaluate the performance of the trained model. It requires understanding of concepts like precision, recall, F1 score, ROC curves, etc. Should we not like accuracy of the simple AI model then, we will be using transfer learning to improve the accuracy and latency of our model.
5. AI Training for Students: This involves pedagogical skills to effectively teach students about AI and its applications in medical diagnosis. It requires knowledge of educational psychology and instructional design.

The problem scope also contains multiple challenging requirements that match or exceed current solutions or industry standards:

- The project aims to predict the occurrence and recurrence of cancer, which is a complex problem due to the myriad factors that can influence cancer development and progression.
- The project seeks to improve upon current solutions by leveraging AI for more accurate predictions than human doctors alone.
- The project also aims to equip students with skills to build and train a medical AI diagnosis tool, thereby contributing to the development of future professionals in this field.

4.3 Modern Engineering Tools

- Tensorflow- An open source library for machine learning and artificial intelligence. It will provide us with all the functionalities required for training and adjusting our AI model.

- Keras- An open source library for multiple pre-implemented neural networks. This library will provide us with a neural network model that we will use in transfer learning to adjust the model to better fit our needs.
- GitLab- Track the progress of our project and allow easy sharing of code. It will also allow us to revert back to older implementations of our project and start independent branches to work on specific tasks without interfering with others.
- Python- A programming language that lends itself greatly to the development and training of AI. All neural networks on Keras are implemented in python.
- SQL- A database language that we will use for storing and retrieving data for our AI.
- JavaScript- A programming language we will use for our front end development of the website for our healthcare professionals.

4.4 Design Context

Communities affected

– Patients at risk and those diagnosed with cancer.

– Families of patients, healthcare policy makers, and the general public.

Societal Needs Addressed

– The project addresses the need for early and accurate cancer detection, which often translates to better prognosis and survival rates. It also addresses the need for automation in diagnostics to help with the increasing number of cancer cases globally.

Public Health, Safety, and Welfare

– Positive impacts include, early and accurate cancer prediction could lead to improved patient survival rates, and may reduce the number of unnecessary procedures.

– Negative impacts include, potential false negatives or false positives. This could lead to missed treatments or unnecessary stress and medical interventions.

Global, Cultural, and Social

– Positive impacts: Universal access to such a tool could lead to standardized cancer care across different regions, which would benefit areas with less access to expert diagnostics. The system might also overcome bias in human-based diagnoses.

– Challenges: Cultural skepticism towards AI may make people hesitant to trust the diagnoses. Additionally, if we are too reliant on the technology it could lead to medical professionals to lose a bit of their skill.

Environmental

– Positive impact: If the system reduces the need for other means of testing due to improved accuracy, it could lead to decreased usage of medical resources and chemicals related to cancer screenings.

– Negative impact: The energy consumption of training and deploying the neural network could be significant.

Economic

– Positive impact: Faster and more accurate diagnosis could reduce the overall treatment cost by catching cancers early, leading to less aggressive treatments and shorter hospital stays. For healthcare providers the efficiency of an automated system could lead to reduced labor costs.

– Challenges: There might be potential job displacements if AI takes over roles traditionally held by medical professionals. However the tool will most likely be used as an aid rather than a replacement.

4.5 Prior Work/Solutions

A number of large studies have been done recently to assess the usefulness of AI in the realm of oncology. The resulting conclusion is that AI has strong potential in predicting and diagnosing cancer using pathology profiles and images studies (Zhang et al., 2023). The University of Pittsburgh in 2020 created a very accurate machine learning technique that diagnoses prostate cancer with a specificity of 98% and sensitivity of 98% (cite 1). Another AI technique that has been used recently was based on a Google DeepMind algorithm and was used to predict breast cancer more accurately than human specialists, also in 2020 (McKinney et al., 2020)

Oncology imaging studies using AI have an advantage in that training AI on images is relatively straightforward with huge results, such as the above mentioned case where breast cancer prediction was more accurate than a human specialist in that area. There are several disadvantages of using imaging for cancer research. One is that in some cases it is heavily biased, such as in detecting skin cancer the accuracy varies depending on the color of skin (Wen et al., 2021). Another study done showed that the AI could tell which institution had supplied the images and ended up lumping patients together by institution when training itself on the data which could lead to results based off of the institution rather than individual biology (Wood, 2021).

For AI training based on pathology data there is an issue of procuring good data. Training a model requires massive datasets to create accurate profiles, and this is tricky in the healthcare industry due to issues such as patient privacy, lack of data shared between institutions, and availability of data in general (Khan et al., 2023).

Due to some unforeseen circumstances with our advisor/client we are still in the process of getting information about the data we are using and our exact implementation goal. From what we know so far we will be training our AI model on sparse representation data pulled from images of cancerous and non-cancerous cells in the form of csv files. Each file is one image and column A gives a position x coordinate and column B is the corresponding value which together can be read as a vector.

The advantage to our approach is that we are specifically training just on the cells, so we can identify any kind of cancer since all cancer cells look the same. We also do not run into any bias such as encountered in skin cancer image studies.

The corresponding disadvantage is we cannot differentiate what kind of cancer it is.

Khan, B. *et al.* (2023) *Drawbacks of artificial intelligence and their potential solutions in the healthcare sector*, *Biomedical materials & devices (New York, N.Y.)*. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9908503/ (Accessed: 22 October 2023).

McKinney, S.M. *et al.* (2020) *International Evaluation of an AI system for breast cancer screening*, *Nature News*. Available at: https://www.nature.com/articles/s41586-019-1799-6 (Accessed: 22 October 2023).

Wen, D. (2021) *Characteristics of publicly available skin cancer image datasets: A ...*, *The Lancet Digital Health*. Available at: https://www.thelancet.com/journals/landig/article/PIIS2589-7500(21)00252-1/fulltext (Accessed: 22 October 2023).

Wood, M. (2021) *Artificial intelligence models to analyze cancer images can take shortcuts that introduce bias for minority patients*, *UChicago Medicine*. Available at: https://www.uchicagomedicine.org/forefront/research-and-discoveries-articles/artificial-intelligence-models-to-analyze-cancer-images-can-take-shortcuts-that-introduce-bias-for-minority-patients (Accessed: 22 October 2023).

Zhang, B., Shi, H. and Wang, H. (2023) *Machine learning and AI in cancer prognosis, prediction, and treatment selection: A critical approach*, *Journal of multidisciplinary healthcare*. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10312208/#:~:text=Machine%20learning%20(ML)%2C%20a,in%20predicting%20cancer%20than%20clinicians (Accessed: 22 October 2023).

4.6 Design Decisions

1. Choice of AI Model: The type of AI model to be used needs to be decided. This could be a decision between using a Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), or other types of models. The choice will depend on the nature of the data and the specific requirements of the problem.
2. Data Preprocessing: Decisions need to be made about how to preprocess the data. This could involve dealing with missing values, normalizing numerical data, encoding categorical data, etc.
3. Model Evaluation Metrics: The metrics used to evaluate the performance of the model need to be decided. This could include accuracy, precision, recall, F1 score, etc.
4. Training Methodology: Decisions need to be made about how to train the model. This includes choosing an optimization algorithm, deciding on the number of epochs, batch size, learning rate, etc.
5. Deployment Strategy: The trained model will be integrated into a web service. This web service will allow users to input their data and receive an assessment of their cancer occurrence and recurrence chances. However, it's important to note that this service is intended to be a supplementary tool and not a replacement for professional medical advice. Users are strongly advised to consult with medical professionals for a comprehensive evaluation of their health. This approach ensures that our AI tool aids in the process of medical diagnosis while emphasizing the irreplaceable value of professional medical consultation.

4.7 Proposed Design

We have begun transferring some of the projects from Keras into Google Collab projects to understand the general principles of implementing AI projects using python.

4.7.1 Design 0 (Initial Design)

Design Visual and Description

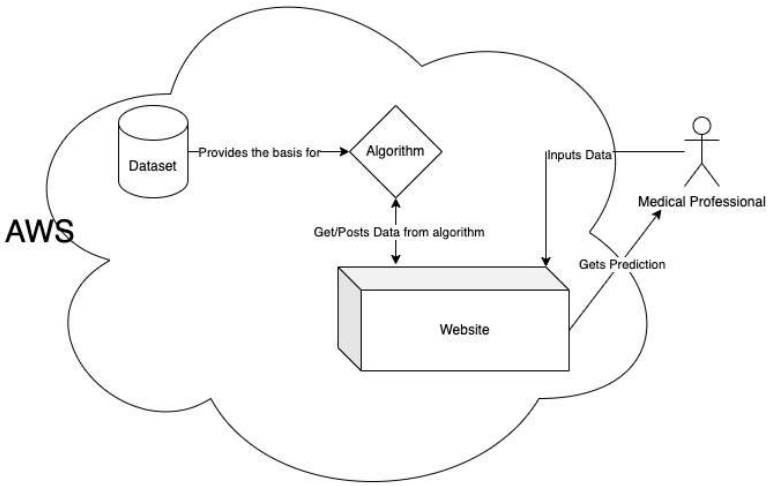
Include a visual depiction of your current design. Different visual types may be relevant to different types of projects. You may include: a block diagram of individual components or subsystems and their interconnections, a circuit diagram, a sketch of physical components and their operation, etc.

Describe your current design, referencing the visual. This design description should be in sufficient detail that another team of engineers can look through it and implement it.

Justify each component in the design with respect to requirements.

Cancer Prediction Software

This diagram details the data flow of our project. A medical professional inputs a spectrum .csv file into our webpage which then sends the csv file through our algorithm which will predict the time that the owner of that cell has left to live based on comparing it with the spectrums found in the dataset



(The components of this project will be hosted on AWS)

Functionality

Describe how your design is intended to operate in its user and/or real-world context. This description can be supplemented by a visual, such as a timeline, storyboard, or sketch.

How well does the current design satisfy functional and non-functional requirements?

On a practical level the intention is that a medical professional can input data in csv form on a website where our model is hosted and get back a percentage which is the chance that the patient has cancer.

4.7.2 Design 1 (Design Iteration)

We didn't have information on the data we would be training on until near the end of the semester. Once we received the data we realized the Keras model we originally chose, Structured data classification with FeatureSpace, would not work with this kind of data. The feedback received from our faculty mentor at this point indicated we should focus on getting familiar with

running general AI models on the platforms we will be using and not worry about choosing a specific Keras model just yet.

Design Visual and Description

Include a visual depiction of this design as well highlighting changes from Design 0. Describe these changes in detail. Justify them with respect to requirements.

NOTE: The following sections will be included in your final design document but do not need to be completed for the current assignment. They are included for your reference. If you have ideas for these sections, they can also be discussed with your TA and/or faculty adviser.

4.8 Technology Considerations

Highlight the strengths, weaknesses, and trade-offs made in technology available.

Discuss possible solutions and design alternatives

4.9 Design Analysis

- Did your proposed design from 4.7 work? Why or why not?
- What are your observations, thoughts, and ideas to modify or iterate further over the design?

5 Testing

Testing is an **extremely** important component of most projects, whether it involves a circuit, a process, power system, or software.

The testing plan should connect the requirements and the design to the adopting test strategy and instruments. In this overarching introduction, given an overview of the testing strategy. Emphasize any unique challenges to testing for your system/design.

5.1 Unit Testing

Once our AI model is at the point we can begin testing we could use either pytest or unit test for the unit tests. The AI model is one unit. Ideally we will be testing it to make sure it is behaving consistently and outputting the results we want. Unit testing for AI models is not the same as typical software testing and can be quite difficult. They can take a long time because an accurate test requires training the model which is costly even on a modest data set. Some recommendations from experts have been to consider instead of some unit tests using smoke tests. We will consult more with our faculty as we near the test stage to see what would be the best way for our particular model.

5.2 Interface Testing

Once our AI model is at the point we can begin testing we will use unit testing as stated in 5.1. We will also mock out our interfaces with unittest.mock in Python to test the interactions between the database and our machine learning model. In doing so we can test each interface separately to find and correct possible errors much faster.

5.3 Integration Testing

A few of our critical integration paths in our design are as follows.

- Front-end Application to AI Model: The integration of these is critical because it's the main interface the healthcare professionals will interact with. It must be tested rigorously to ensure seamless function.
- AI Model to Database: After the model makes a prediction, it may need to retrieve or store information in the database. The integration between the model and the SQL database is critical for maintaining patient records and making predictions based on historical data.
- Training Pipeline: From model development, training, to evaluation, this path is critical as it affects the overall accuracy and reliability of the AI model. Ensuring that data flows correctly and the model updates are critical for learning efficiency and accuracy.

For testing these paths we will use automated integration testing with frameworks such as Selenium for front-end testing or TensorFlow's testing module. We will also be using GitLab's CI/CD to automatically build and test our application when changes are made.

5.4 System Testing

System testing is a high-level testing phase that involves evaluating the complete and integrated software system to ensure compliance with specified requirements. For us system testing will validate the end-to-end functionality, performance and user interface. Our system testing strategy will include functional testing, performance testing, usability testing, and compatibility testing. The tools we can use for these include load testing tools such as JMeter or Locust for performance testing. Automated testing suits such as Selenium for functional and usability testing of web interfaces.

5.5 Regression Testing

Each time we make an addition we will repeat testing from our previous implementation afterwards to make sure we are still getting the same correct results. We do not have many different parts in our project so this will probably not be a big problem for us. The AI itself is self-contained so the only concern foreseen right now would be when connecting it to a website making sure it is correctly interacting with the model and creating the correct behavior. As we go through this process 'breaking' the model isn't really a concern, we just might need to troubleshoot interacting with it.

5.6 Acceptance Testing

For functional requirements we will set up some use case scenarios for different types of users interacting with the implemented product that also define the desired outcome of each case. We will manually run each of these cases and check the behavior and results we get and check them against the success criteria. When we create the use cases we will show them to our client and double check that they follow their desired functionality of the final product. We will also demonstrate to our client the functionality by walking through some cases in front of them so

they can see exactly how the product is working.

For non-functional design requirements we will collaborate with our client to establish the expectations for performance of the system and then create scenarios where we test the specific non-functional requirements. This could require several different types of testing such as load testing and reliability testing among others. We will work closely with our client to determine what type of testing is most important here.

5.7 Security Testing (if applicable)

We will be hosting our model on a website. The security of the website will be taken care of by the host we select. Once we get access to the data we will consult with our faculty advisor on if there are any additional security issues related to the data we need to worry about.

5.8 Results

What are the results of your testing? How do they ensure compliance with the requirements? Include figures and tables to explain your testing process better. A summary narrative concluding that your design is as intended is useful.

Since we are conducting a spectrum analysis to base the likelihood of an individual having cancer, we will need to test the analyzer by making sure that it can accurately identify the peaks of the spectrum by using differentiation of the slope to land on a result of zero. Though we should be wary of the fact that valleys will also give a slope of zero. We will need to make sure that the slope leading up the peak should be positive and the leaving slope should be negative. Once we are able to train the model to identify peaks correctly, we need to train the model to determine how similar an input csv file is to the aggregate that the model creates and predict how long a person has to live based on this aggregation.

6 Implementation

Describe any (preliminary) implementation plan for the next semester for your proposed design in 3.3. If your project has inseparable activities between design and implementation, you can list them either in the Design section or this section.

Building a model which can take a CSV as an input, with an excel as a reference.

Building a basic webpage which can take a CSV as input.

7 Professionalism

This discussion is with respect to the paper titled “Contextualizing Professionalism in Capstone Projects Using the IDEALS Professional Responsibility Assessment”, *International Journal of Engineering Education* Vol. 28, No. 2, pp. 416–424, 2012

7.1 Areas of Responsibility

Pick one of IEEE, ACM, or SE code of ethics. Add a column to Table 1 from the paper corresponding to the society-specific code of ethics selected above. State how it addresses each of the areas of seven professional responsibilities in the table. Briefly describe each entry added to

the table in your own words. How does the IEEE, ACM, or SE code of ethics differ from the NSPE version for each area?

Professional Responsibility	**NSPE**	**IEEE**	**ACM**
Hold paramount the safety, health, and welfare of the public	NSPE Code of Ethics requires engineers to prioritize public safety, health, and welfare in their professional work, ensuring projects are designed and executed to minimize risks to society.	IEEE Code of Ethics emphasizes the importance of safety, health, and the public welfare in technological innovations, requiring engineers to consider these factors in their work.	ACM Code of Ethics emphasizes the responsibility of computing professionals to ensure that their work does not cause harm to individuals or society. It includes considerations for privacy, security, and well-being in computing solutions.
Perform services only in areas of their competence	NSPE requires engineers to work within their expertise, seeking additional training or expertise if venturing into new areas.	IEEE stresses the importance of competence, requiring engineers to only undertake tasks that match their expertise or seek relevant education/training.	ACM highlights the responsibility of computing professionals to maintain and develop their professional skills, undertaking only those tasks for which they are qualified.
Issue public statements only in an objective and truthful manner	NSPE mandates that engineers maintain honesty and objectivity in all public statements related to their professional work.	IEEE requires members to be truthful and objective in their professional communications, providing accurate information without bias.	ACM requires computing professionals to be honest and truthful in their public communications, ensuring that information is accurate and not misleading.
Act for each employer or client as faithful agents or trustees	NSPE expects engineers to act in the best interest of their employers or clients, safeguarding their interests.	IEEE emphasizes the duty of engineers to act in the best interests of their employers or clients, maintaining confidentiality and avoiding conflicts of interest.	ACM stresses the importance of computing professionals acting in the best interests of their employers or clients, protecting confidential information and avoiding conflicts of interest.
Avoid deceptive acts	NSPE prohibits engineers from engaging in deceptive practices or acts that mislead	IEEE prohibits deceptive acts, requiring engineers to be transparent and truthful in their professional	ACM requires computing professionals to avoid deceptive practices, ensuring that their work and communications are

	others in their professional work.	dealings, avoiding misrepresentation.	transparent and do not deceive others.
Conduct themselves honorably, responsibly, ethically, and lawfully so as to enhance the honor, reputation, and usefulness of the profession	NSPE expects engineers to uphold the honor, reputation, and integrity of the engineering profession, acting ethically and lawfully.	IEEE stresses the importance of honorable conduct, ethical behavior, and compliance with laws to enhance the reputation of the engineering profession.	ACM emphasizes the responsibility of computing professionals to behave ethically, lawfully, and responsibly, contributing positively to the reputation of the computing profession.
Continue their professional development throughout their careers and provide opportunities for the professional development of those engineers under their supervision	NSPE requires engineers to continually develop their skills and support the professional development of others under their supervision.	IEEE emphasizes the importance of continuous professional development for engineers and providing opportunities for the development of those under their supervision.	ACM stresses the responsibility of computing professionals to continually enhance their professional skills and knowledge and support the professional development of colleagues under their supervision.

Each society's code of ethics shares fundamental principles such as prioritizing public welfare, maintaining competence, honesty, acting in the best interest of clients, and ethical conduct. However, they differ slightly in their emphasis and specific details. For instance:

- **Scope and Application:** Each code is tailored to the specific profession it represents, highlighting unique aspects relevant to engineering or computing.
- **Language and Emphasis:** While the core values align, the wording and emphasis on certain principles might differ, reflecting the nuances of each profession.
- **Specific Guidelines:** Each society provides unique guidelines or examples relevant to their field, addressing specific ethical challenges and considerations.

These differences arise from the specific contexts, practices, and challenges inherent in engineering and computing professions, but all aim to uphold the integrity and ethical conduct of their respective professions.

7.2 Project Specific Professional Responsibility Areas

For each of the professional responsibility area in Table 1, discuss whether it applies in your project's professional context. Why yes or why not? How well is your team performing (High, Medium, Low, N/A) in each of the seven areas of professional responsibility, again in the context of your project. Justify.

- **Hold paramount the safety, health, and welfare of the public:**
    - **Applicability:** Absolutely applicable. Predicting cancer occurrence and recurrence directly impacts public health and well-being.
    - **Team Performance:** High. Our project's goal to improve cancer prediction aligns with prioritizing public welfare.
- **Perform services only in areas of their competence:**
    - **Applicability:** Highly applicable. Working on cancer prediction demands specialized knowledge in AI and medical data analysis.
    - **Team Performance:** High/Medium. Our team has some expertise in AI and collaborates with medical professionals, it aligns well. Regardless of our expertise in AI, seeking expert advice is highly necessary.
- **Issue public statements only in an objective and truthful manner:**
    - **Applicability:** Relevant. While our team might not directly issue public statements, the accuracy and transparency of findings are crucial.
    - **Team Performance:** High. Ensuring the accuracy and objectivity of our predictive model and its outcomes demonstrates adherence to this responsibility.
- **Act for each employer or client as faithful agents or trustees:**
    - **Applicability:** Applicable if our team collaborates with specific organizations, institutions, or stakeholders.
    - **Team Performance:** Applicable, ensuring confidentiality and prioritizing the client's interests would be important.
- **Avoid deceptive acts:**
    - **Applicability:** Important for maintaining integrity in research, data reporting, and model presentation.
    - **Team Performance:** High. Transparent methodologies and truthful reporting contribute to fulfilling this responsibility.
- **Conduct themselves honorably, responsibly, ethically, and lawfully:**
    - **Applicability:** Essential in all aspects of the project, including data handling, model development, and professional conduct.
    - **Team Performance:** High. Upholding ethical conduct and compliance with laws demonstrates adherence to this responsibility.
- **Continue their professional development throughout their careers and provide opportunities for the professional development of those engineers under their supervision:**
    - **Applicability:** Relevant for ongoing skill enhancement and knowledge sharing within the team.
    - **Team Performance:** Medium/High. Depending on the team's efforts to stay updated with advancements in AI, medicine, and ethical practices. Providing learning opportunities would enhance this further.

Overall, it seems that our team is performing well in most areas, particularly concerning ethical and responsible conduct, transparency in reporting, and prioritizing public welfare. However, assessing collaboration, ongoing learning initiatives, and potentially seeking expertise outside the team in specific areas might further enhance our performance in some aspects.

7.3 Most Applicable Professional Responsibility Area

In this project, accurate cancer prediction using AI, the most applicable professional responsibility area would likely be "Hold paramount the safety, health, and welfare of the public."

This responsibility directly aligns with the core objective of your project, which is to leverage AI to predict cancer occurrence and recurrence. By focusing on improving cancer prediction, your team implicitly prioritizes the well-being and health outcomes of individuals. The accuracy and reliability of your predictive model could significantly impact patient outcomes, treatment strategies, and overall healthcare.

This responsibility extends beyond the technical aspects of AI and data analysis. It encompasses the ethical obligation to ensure that the predictions made by the AI model contribute positively to public health and patient care. Emphasizing the safety and welfare of individuals through accurate cancer predictions demonstrates a commitment to this professional responsibility area and its significance in your project's context.

8 Closing Material

8.1 Discussion

Our project is experiments oriented. Now that we have the data, we will begin running experiments next semester. Our hypothesis is as follows; If we have a large set of data of various cancer patients and how long they survived, can we create an amalgamation of this data to test outside data of other suspected cancer patients and predict the likelihood of them having cancer.

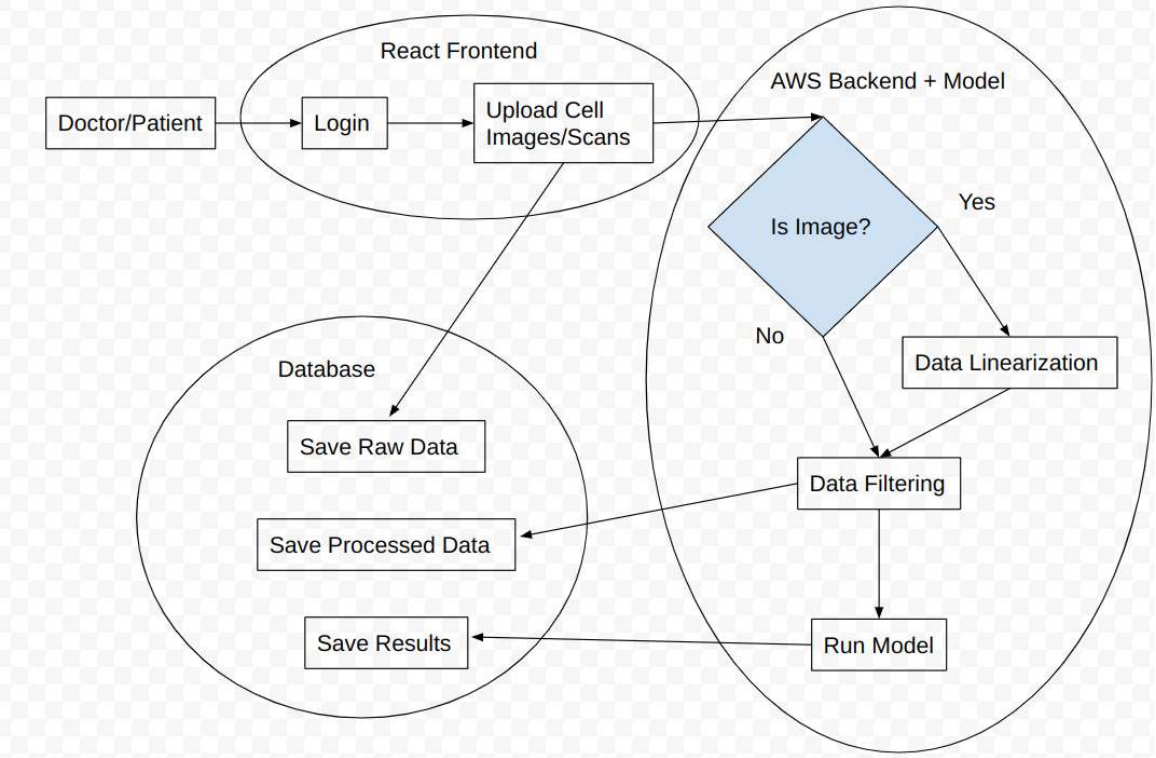
8.2 Conclusion

Our goal is to train an AI model that is trained to recognize and predict the risk of cancer occurrence and reoccurrence. We have completed the goal for this semester of designing our implementation of this and on the practical side getting some hands-on experience running general AI models on Colab and AWS in preparation for starting actual implementation next semester. We also signed an NDS and formally received the data we will be training the model on. We were a bit constrained in developing our design since due to unforeseen circumstances we lost contact with our faculty mentor/ client early in the semester and were unable to get answers to some of our questions on the how the deliverable should work, the type of data we would be training it on, and details on the end use of the project.

8.3 References

No sources to be cited, all of the information in this document was formed from our previous 4 years of University education and Dr. Gaffar’s lectures on this project.

8.4 Appendices



8.4.1 Team Contract

Team Members:

1) Eric Schmitt
2) Norfinn Norius
3) Mark Hanson
4) Thriambak Giriprakash
5) Chris Tague
6) Bishal Ghataney

Team Procedures

1. Day, time, and location (face-to-face or virtual) for regular team meetings:
   Wed 11am, SICTR whatever room is available
2. Preferred method of communication updates, reminders, issues, and scheduling (e.g., e-mail, phone, app, face-to-face): Discord and Text, and important documents and stories on Gitlab
3. Decision-making policy (e.g., consensus, majority vote): Majority vote
4. Procedures for record keeping (i.e., who will keep meeting minutes, how will minutes be shared/archived): Thriambak will keep meeting minutes

Participation Expectations

1. Expected individual attendance, punctuality, and participation at all team meetings:
   a. Communicate, try to be there
2. Expected level of responsibility for fulfilling team assignments, timelines, and deadlines:
   a. we will meet deadlines, and communicate with the teammates and the client if we think they cannot be met
3. Expected level of communication with other team members:

    a. High expectations
4. Expected level of commitment to team decisions and tasks:
    a. High expectations (don't bite off more than you can chew, if you do communicate)

Leadership

1. Leadership roles for each team member (e.g., team organization, client interaction, individual component design, testing, etc.):
    a. Client Communication: Norfinn
    b. Minutes and Administration: Thriambak
    c. Senior Engineer: Bishal
    d. Devs: Chris, Mark, Eric will be taking the lead on development whenever Norfinn or Thriambak are doing something related to their roles.
2. Strategies for supporting and guiding the work of all team members:
    a. Meet regularly, communicate frequently, ask for help if you need it
3. Strategies for recognizing the contributions of all team members:
    a. Cookie and “Good job”

Collaboration and Inclusion

1. Describe the skills, expertise, and unique perspectives each team member brings to the team.
    a. Bishal: experience with AI, Python, Java, C/C#, SQL
    b. Thriambak: Python, Java, C/C++, SQL
    c. Eric: Java, C, C#, C++, SQL
    d. Mark: Java, C, C++, SQL
    e. Chris: Python, Java, C++, Rust, Full stack development
    f. Norfinn: Java, C, C++, SQL, Javascript, webGL
2. Strategies for encouraging and supporting contributions and ideas from all team members:
    a. Give time for everyone to give an opinion and ask questions.
3. Procedures for identifying and resolving collaboration or inclusion issues (e.g., how will a team member inform the team that the team environment is obstructing their opportunity or ability to contribute?) Don't encourage these behaviors and nip them in the bud.

Goal-Setting, Planning, and Execution

1. Team goals for this semester:
    a. Develop structure for project
    b. Begin development of system
2. Strategies for planning and assigning individual and team work:
    a. Team will outline a task that needs to be completed, and will make story cards which go in the direction of completing said goal. and members can pick up and

complete stories at will. Make sure to put names on selected stories. Todo information is also available on discord.

3. Strategies for keeping on task:
    a. Follow tasks on story cards, if there is a situation where an individual is working on something fruitless for an extended period of time, they should bring it up with the team. Preferably we would like to preempt this by having our Senior Engineer and experienced individuals weigh in on story cards and tasks that might end up being a rabbit hole.

Consequences for Not Adhering to Team Contract

1. How will you handle infractions of any of the obligations of this team contract?
    a. Help each other to the best of our abilities.
2. What will your team do if the infractions continue?
    a. If it continues we will elevate it to the TA or faculty advisor.

****

a) I participated in formulating the standards, roles, and procedures as stated in this contract.

b) I understand that I am obligated to abide by these terms and conditions.

c) I understand that if I do not abide by these terms and conditions, I will suffer the

consequences as stated in this contract.

1) Eric Schmitt	DATE: 09/01/23
2) Norfinn Norius	DATE: 09/01/23
3) Mark Hanson	DATE: 09/01/23
4) Chris Tague	DATE 09/01/23
5) Thriambak G.	DATE 09/01/23
6) Bishal Ghataney	DATE 09/01/23