



HIGH LEVEL DESIGN DOCUMENT

From Diagnosis to Remission: Understanding Cancer Detection and Staging

UE20CS390A – Capstone Project Phase – 1

Submitted by:

Name: Akshat Bhandari	SRN : PES2UG20CS030
Name: Ayush Singh	SRN : PES2UG20CS080
Name : Ayan Aggarwal	SRN : PES2UG20CS079
Name: Ankur Kumar Dubey	SRN : PES2UG20CS054

Under the guidance of

Dr. Prema R

Designation
PES University

January - May 2023

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
FACULTY OF ENGINEERING
PES UNIVERSITY**

(Established under Karnataka Act No. 16 of 2013)

Electronic City, Hosur Road, Bengaluru – 560 100, Karnataka, India

TABLE OF CONTENTS

1. Introduction	4
2. Current System	4
3. Design Considerations	4
3.1 Design Goals	4
3.2 Architecture Choices	4
3.3 Constraints, Assumptions and Dependencies	4
4. High Level System Design	5
5. Design Description	5
5.1 Master Class Diagram	6
5.2 Reusability Considerations	6
6. ER Diagram / Swimlane Diagram / State Diagram	6
7. User Interface Diagrams	6
8. Report Layouts	6
9. External Interfaces	7
10. Packaging and Deployment Diagram	7
11. Help	7
12. Design Details	7
12.1 Novelty	7
12.2 Innovativeness	7
12.3 Interoperability	7
12.4 Performance	7
12.5 Security	7
12.6 Reliability	7
12.7 Maintainability	7
12.8 Portability	7
12.9 Legacy to Modernization	7
12.10 Reusability	7
12.11 Application Compatibility	7

12.12 Resource Utilization	7
Appendix A: Definitions, Acronyms and Abbreviations	8
Appendix B: References	8
Appendix C: Record of Change History	8
Appendix D: Traceability Matrix	8

Note:

Section – 1 & Section 2	Common for Product Based and Research Projects
Section 3 to Section 11	High-Level Design for Product Based Projects.
Section 12	High-Level Design for Research Projects.
Appendix	Provide details appropriately

1. Introduction

The lung cancer detection website is a web-based application designed to assist medical professionals in the early detection and diagnosis of lung cancer. The website leverages machine learning algorithms, image processing techniques, and specialized knowledge to analyze medical images and provide a diagnosis based on the stage of cancer present.

The website's main functionality will be divided into three parts. The first part will be responsible for uploading medical images, which will be processed and analyzed by the machine learning model. The second part will display the results of the analysis in an easily understandable format, highlighting the areas of the image that are potentially cancerous and indicating the stage of cancer present. The third part of the website will include a reports section, providing additional analysis and information for medical professionals.

The reports section will include analysis covering treatment plans available, specialized hospitals for such treatment plans, and the patient's life expectancy based on the stage of cancer present. This information will be available only to radiologists and doctors, providing valuable insights into the best possible treatment options for the patient.

The website's success will be measured by its ability to accurately detect the stage of cancer present in medical images, assist medical professionals in providing the best possible treatment options for patients, and reduce the time required for manual interpretation of medical images.

2. Current System

The existing system for identifying tumors in CT scans involves manual identification by doctors and radiologists, which can be error-prone. The process of manually identifying and characterizing lung nodules is challenging and time-consuming, as it requires a high level of expertise and experience. Moreover, different radiologists may interpret the same scan differently, leading to inconsistency and variability in the identification and characterization of lung nodules.

Manual identification of tumors also carries the risk of false-positive or false-negative results. False-positive results occur when a radiologist identifies a nodule as malignant when it is not, leading to unnecessary further testing and potential harm to the patient.

False-negative results occur when a radiologist fails to identify a malignant nodule, leading to delayed diagnosis and treatment, which can result in a worse patient outcome.

Furthermore, manual identification of tumors is subjective, as it relies on the radiologist's interpretation of the scan. The process is also labor-intensive and time-consuming, which can lead to long wait times for patients to receive their diagnosis and treatment plan. In addition, manual identification of tumors is not scalable, as it is not feasible to hire enough radiologists to keep up with the growing demand for lung cancer screening.

Overall, the existing system of manual identification of tumors by doctors and radiologists is error-prone, subjective, labor-intensive, time-consuming, and not scalable. Therefore, there is a need for more efficient and accurate methods of identifying lung nodules in thoracic CT scans.

3. Design Considerations

3.1. Design Goals

Design Goals, Guidelines, and Principles:

The design of a new system for identifying lung nodules in thoracic CT scans should have the following goals, guidelines, and principles:

Accuracy: The new system should be more accurate than the existing system of manual identification by doctors and radiologists. It should minimize false-positive and false-negative results and provide a more reliable and consistent way of identifying and characterizing lung nodules.

Efficiency: The new system should be more efficient than the existing system, with faster processing times and shorter wait times for patients to receive their diagnosis and treatment plan.

Usability: The new system should be user-friendly and easy to use for both medical professionals and patients. It should have a simple and intuitive interface that does not require extensive training to use.

Scalability: The new system should be scalable and able to handle a large volume of thoracic CT scans. It should be able to keep up with the growing demand for lung cancer screening and diagnosis.

Accessibility: The new system should be accessible and available to all medical professionals and patients who need it. It should be web-accessible and have no geographical or technological barriers to access.

Why the newly proposed system is better than the existing system :

The newly proposed system for identifying lung nodules in thoracic CT scans is better than the existing system of manual identification by doctors and radiologists because it is more accurate, efficient, and scalable. It uses advanced machine learning algorithms and computer vision techniques to analyze thoracic CT scans and identify lung nodules with a high level of accuracy and reliability.

The new system is also more efficient than the existing system, with faster processing times and shorter wait times for patients to receive their diagnosis and treatment plan. It is designed to handle a large volume of thoracic CT scans, making it more scalable and able to keep up with the growing demand for lung cancer screening and diagnosis.

The "look and feel" of the new system:

The new system for identifying lung nodules in thoracic CT scans should have a simple and intuitive interface that is easy to use for both medical professionals and patients. It should have a clean and modern design with clear and concise information displayed in an organized and easy-to-read format.

The quality of services characteristics:

The new system for identifying lung nodules in lung CT scans should have high-quality services characteristics, including availability, security, privacy, and speed. It should be available to all medical professionals who need it, regardless of geographical or technological barriers. It should have strong security measures in place to protect patient data and maintain privacy. The system should also be fast and efficient, with short processing times and minimal wait times for patients to receive their diagnosis and treatment plan.

3.2. Architecture Choices

When developing a system for identifying lung nodules in thoracic CT scans using image segmentation techniques, SVM, and CNN, we considered various architecture choices. Some of the alternative choices that we evaluated include:

Traditional Machine Learning Techniques: One alternative to using machine learning models such as SVM and CNN is to use traditional machine learning algorithms such as random forests or support vector machines. While these algorithms can achieve high accuracy in some cases, they typically require extensive feature engineering and do not perform as well as deep learning models when dealing with complex, high-dimensional data such as medical images. Therefore, we decided to use these learning models for our system.

Other Deep Learning Architectures: Another alternative to using SVM and CNN is to use other deep learning architectures such as recurrent neural networks (RNNs) with attention mechanisms. While these architectures can also achieve high accuracy, they may require more complex training procedures and may not be as well-suited for image segmentation tasks as water-shed segmentation ,Otsu's segmentation. Therefore, we decided to use water-shed and Mask otsu for image segmentation in our system.

Hybrid Approaches: A third alternative is to use hybrid approaches that combine deep learning models with traditional machine learning algorithms. For example, one could use a CNN to extract features from medical images, and then use an SVM to perform classification. While this approach can potentially achieve high accuracy, it may also be more complex to implement and may require more resources. Therefore, we decided to use a fully deep learning-based approach for our system.

Given these considerations, we believe that our choice of using image segmentation techniques, SVM, and CNN in our system is the most appropriate. By using deep learning models, we can achieve high accuracy without requiring extensive feature engineering, and by using water-shed , otsu's method for image segmentation, we can accurately identify lung nodules in thoracic CT scans. Additionally, by using SVM and CNN for classification, we can make the final decision on whether a given lung nodule is malignant or benign.

However, there are also some potential downsides to our chosen approach. For example, deep learning models can be computationally expensive and may require significant resources for training and inference. Additionally, the accuracy of the system may be limited by the quality and quantity of training data, and the interpretability of the models may be limited. Nonetheless, we believe that the benefits of our chosen approach outweigh the potential drawbacks..

3.3. Constraints, Assumptions and Dependencies

Assumptions

- **Availability of DICOM-formatted CT scan images:** The website assumes that CT scan images of the lungs are available in DICOM format, and can be uploaded to the website by users.
- **User privacy and data security:** The website assumes that user privacy and data security are maintained by implementing appropriate measures such as data encryption, secure storage, and access controls.
- **Reliable internet connection:** Users are assumed to have access to a reliable internet connection for uploading the CT scan images and accessing the website's services.
- **Sufficient computing power:** The website assumes that it has sufficient computing power to process and analyze the uploaded CT scan images using deep learning algorithms.

Constraints:

- **Quality of imaging:** The accuracy and quality of the images used for detection can be a significant constraint in the diagnosis of lung cancer. Low-quality images can result in false positives and false negatives, leading to incorrect diagnosis and treatment.
- **Technical limitations of imaging techniques:** Imaging techniques may not be able to detect small tumors or may not provide sufficient detail to accurately determine the stage of cancer.
- **Variability of tumor growth:** Lung tumors can grow at different rates, making it difficult to accurately determine the stage of cancer at a given time.
- **Legal repercussions:** Working with medical data may have legal repercussions, such as adhering to data privacy laws or receiving regulatory approval. Legal repercussions or reputational harm may follow failure to adhere to these requirements.

Dependencies:

- Programming language: We will be using Python as our programming language.
- Machine learning framework: We plan to use TensorFlow as our machine learning framework to build and train our model.
- Image processing libraries: For image processing, we will be using OpenCV and Pillow libraries to preprocess the medical images.
- Data visualization libraries: We will use Matplotlib and Seaborn libraries to visualize our data and model results.
- Web application framework: If we decide to build a web application, we will use either Flask or Django as our web application framework.
- Database: We will use a database to store patient information and medical images. We plan to use either MySQL, PostgreSQL, or MongoDB as our database.
- Cloud computing services: If we decide to deploy our application in the cloud, we may use services such as AWS or GCP.

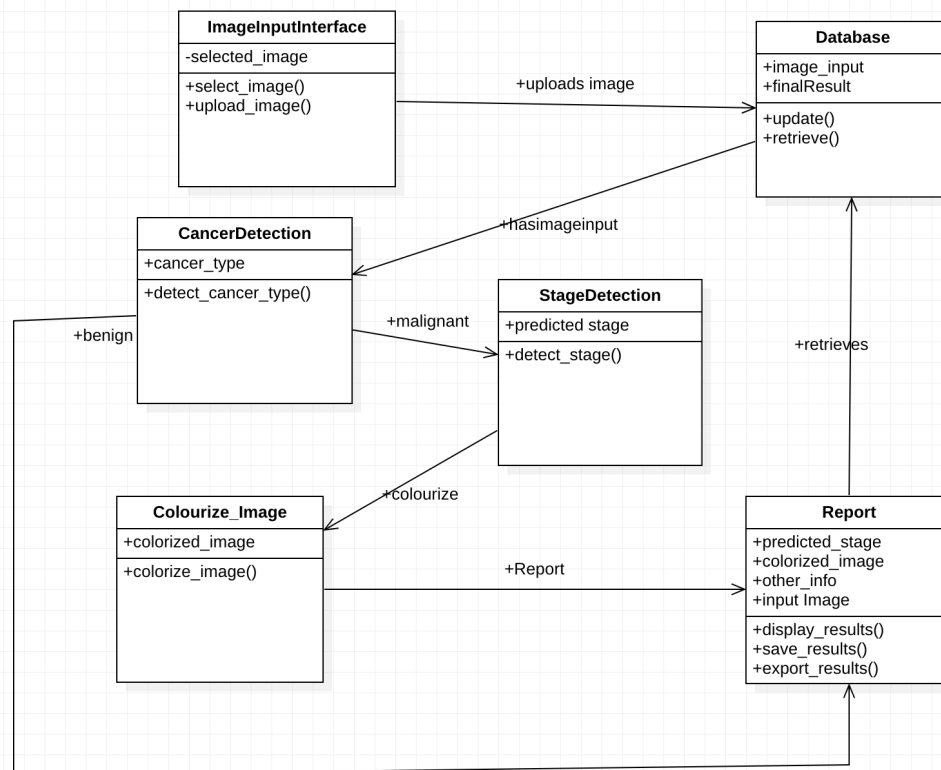
4. High Level System Design

1. User Interface: The system will have a user interface that allows the user to upload or select a lung cancer image. The interface will support various image formats. Once the user selects the image, the system will display a preview of the image.
2. Image Processing: The system will use image processing algorithms to detect and segment the lung nodule from the image. This will involve analyzing the size, shape, texture, and other features of the nodule to determine whether it is cancerous or not.
3. Cancer Detection: The system will use machine learning algorithms to analyze the features of the nodule and predict whether it is benign or malignant. This will involve training the machine learning model on a dataset of lung cancer images to learn the patterns and characteristics of cancerous nodules.
4. Cancer Stage Detection: If the nodule is predicted to be cancerous, the system will use the size of the nodule to predict the stage of the cancer. This will involve comparing the size of the nodule to established size criteria for each stage of lung cancer.
5. Results Interface: The system will display a summary of the analysis results, including the predicted stage of lung cancer, the colorized image, and any other relevant information.

6. Database: The system will store patient data and analysis results in a database for later reference. This will include patient personal details, medical history, and imaging data, as well as the results of the cancer detection and stage detection algorithms.
7. Doctor Interface: The system will provide a separate interface for doctors to access patient data and analysis results. This will allow doctors to review patient records and make informed decisions about diagnosis and treatment

5. Design Description

5.1. Master Class Diagram



5.2. Reusability Considerations

- Project Components that are and can be generated with available reusable components.

One way to maximize reusability is to identify existing components that can be integrated into the project. For instance, there might be libraries or frameworks that can be used to process CT scan images, extract features, or

train machine learning models. By leveraging existing code, the project team can save time and reduce the risk of errors.

Some possible reusable components that could be used in this project include:

- Image processing libraries: Libraries such as OpenCV or Pillow can be used for tasks such as image filtering, segmentation, or feature extraction. These libraries are widely used and have extensive documentation and community support, which can make it easier to integrate them into the project.
 - Machine learning frameworks: Frameworks such as TensorFlow, PyTorch, or Scikit-learn can be used to develop and train machine learning models. These frameworks have many pre-built models and algorithms that can be used as a starting point for developing a tumor detection model.
- Components that can be built in the project for reuse in the project.]

Another way to promote reusability is to design project components in a modular and reusable way. This can involve creating functions or classes that can be used in multiple parts of the project, or creating separate modules that can be imported and used across the project.

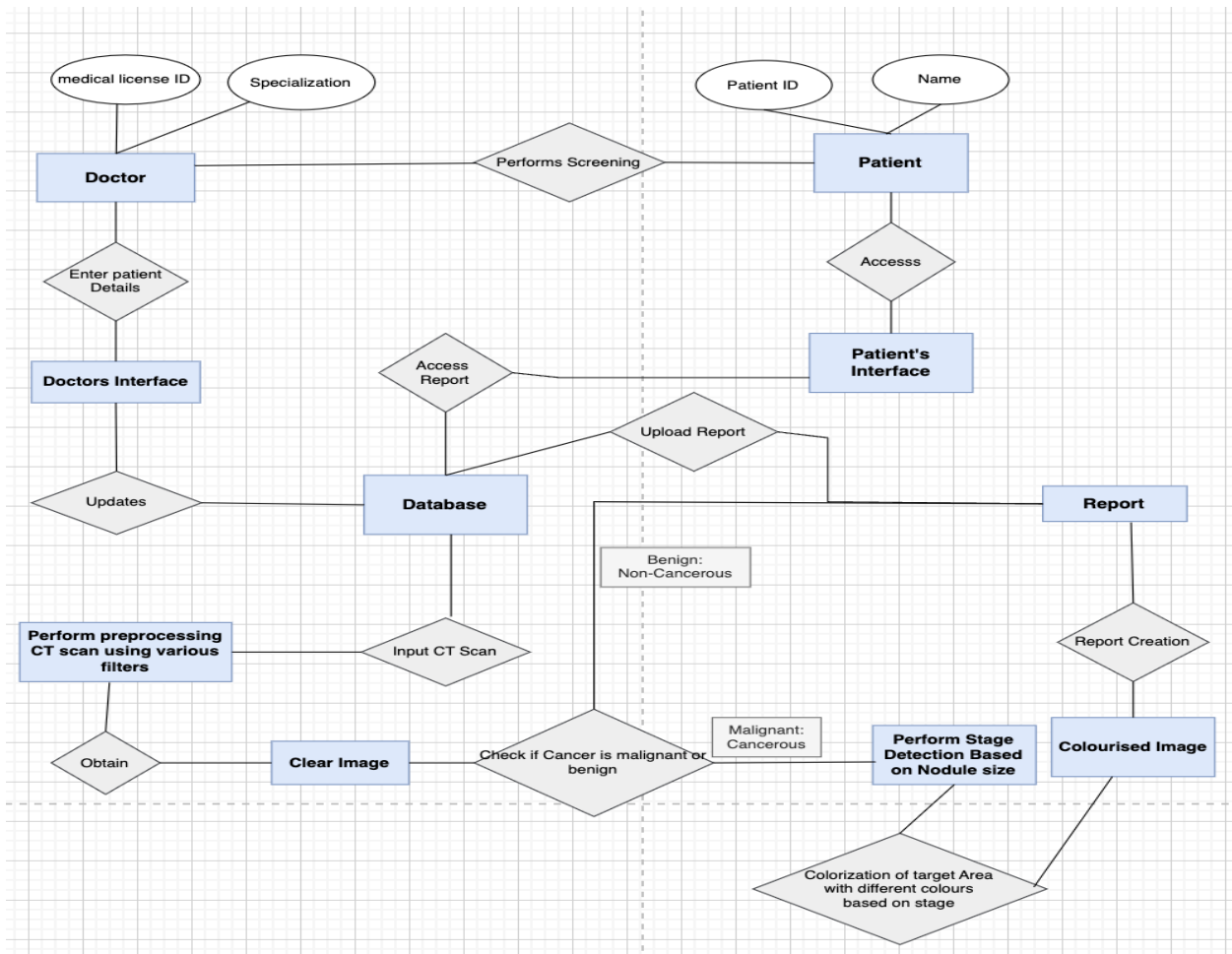
Some possible components that could be built for reuse in this project include:

- Data preprocessing functions: Functions that handle common data preprocessing tasks, such as resizing images, normalizing pixel values, or splitting data into training and validation sets, can be built and reused across the project.
- Feature extraction modules: Modules that implement different feature extraction techniques, such as texture analysis, edge detection, or color histograms, can be developed and used as part of the tumor detection pipeline.
- Visualization functions: Functions that generate visualizations of the CT scan images, such as heatmaps or overlays, can be built and used for debugging or presenting results.

Overall, by leveraging existing reusable components and designing new components for reuse, the project team can increase the efficiency of development, reduce errors, and ensure that the project is maintainable and adaptable over time.

6. ER Diagram / Swimlane Diagram / State Diagram (include as appropriate)

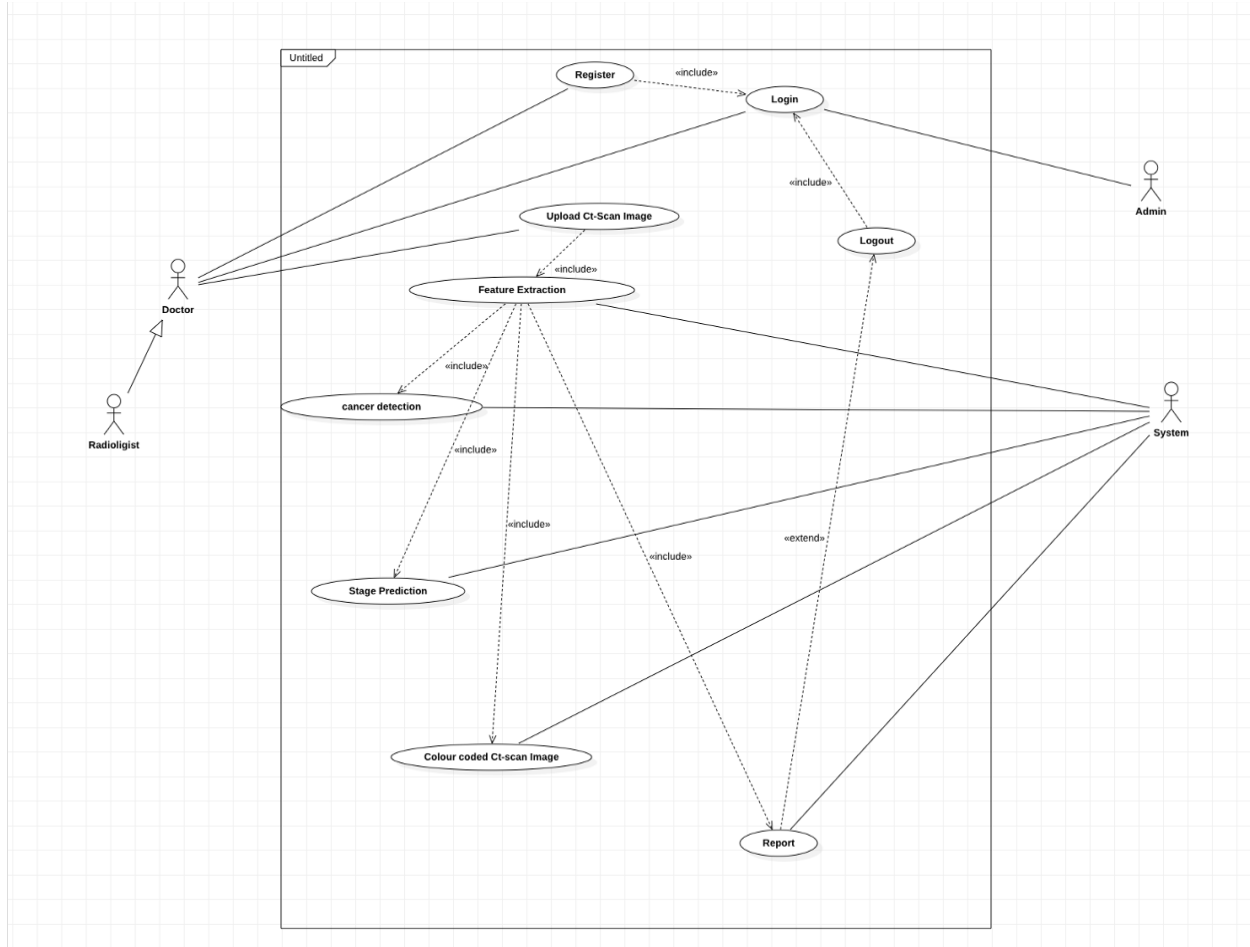
ER Diagram:



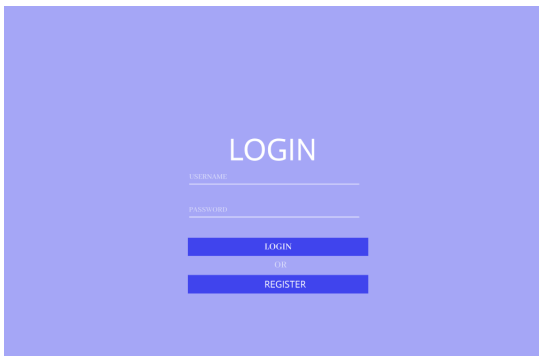
#	Entity	Name	Type
ENTITIES			
1.	Doctor	Doctor	Strong
2.	Patient	Patient	Strong
3.	Patient's Interface	Patient's Interface	Strong

4.	Doctor's Interface	Doctor's Interface	Strong
5.	Database	Database	Strong
6.	Performs Preprocessing of CT scan images	Performs Preprocessing of CT scan images	Strong
7.	Clear Image	Clear Image	Strong
8.	Colourised Image	Colourised Image	Strong
9.	Perform stage detection Based on Nodule size	Perform stage detection Based on Nodule size	Strong
10.	Report		Strong
#	Attribute	Name	Type (size)
DATA ELEMENTS			
1.	Medical License id	Medical_License_id	Primary Key, int
2.	Specialization	Specialization	string
3.	Patient id	Patient_id	Primary Key, int
4.	Name	Name	string

7. User Interface Diagrams



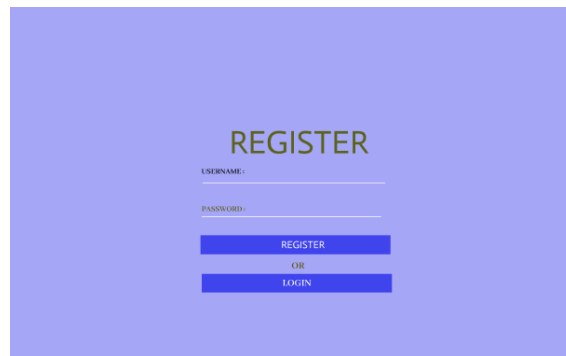
8. External Interfaces



LOGIN

USERNAME:

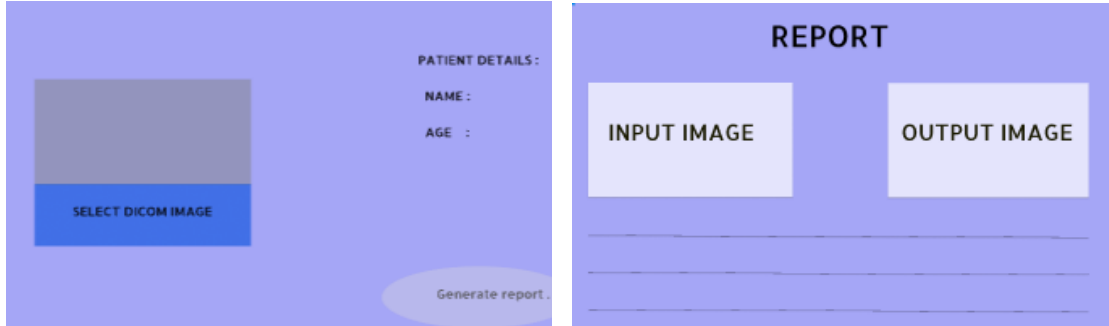
PASSWORD:



REGISTER

USERNAME:

PASSWORD:



The mockup consists of two side-by-side panels. The left panel has a light blue background. On the left side, there is a grey rectangular box representing an image, with a blue button labeled 'SELECT DICOM IMAGE' below it. On the right side, under the heading 'PATIENT DETAILS:', there are two labels: 'NAME:' and 'AGE:'. At the bottom right of this panel is a light blue oval button labeled 'Generate report.'. The right panel has a light blue background and is titled 'REPORT' at the top. It contains two white rectangular boxes, one labeled 'INPUT IMAGE' and one labeled 'OUTPUT IMAGE'. Below these boxes are three horizontal lines representing text input or output.

- If the user already has an account, he or she must Login.
- If the user does not already have an account, he or she must establish one by going to the Register page.
- After creating an account, the user must log in using the credentials he provided.
- The user will be directed to the input screen, where he must enter the CT scan image.
- A report will be generated after you click the Generate report button.
- The report will include input and output images, as well as information regarding the type of cancer diagnosed and available treatment plans.

9. Help

Online Help: Provide online help that is accessible through the system's user interface. This help can provide guidance and instructions on how to use the system's features and functionality, and can include frequently asked questions (FAQs) or troubleshooting guides.

Context-Sensitive Help: Provide context-sensitive help that is available when a user is interacting with a specific feature or function in the system. This help can provide targeted guidance and instructions that are tailored to the user's current context.

User Manual: Provide a user manual that provides comprehensive instructions on how to use the system, including detailed explanations of each feature and function. This manual can be distributed in electronic or printed form, and can be organized in a logical and intuitive manner to make it easy for users to find the information they need.

Technical Manual: Provide a technical manual that provides detailed information on the system's architecture, algorithms, and other technical details. This manual can be targeted at developers or other technical staff who need to understand the system's underlying workings.

Training Materials: Provide training materials such as videos, webinars, or in-person training sessions to help users learn how to use the system effectively. These materials can provide hands-on instruction and guidance on how to use the system's features and functionality, and can be tailored to different user groups or levels of experience.

10. Design Details

Novelty and Innovativeness

- Our project aims to introduce a novel approach to detecting lung cancer from CT scan images by developing a machine learning model that not only identifies the presence of cancer but also determines the stage of the cancer automatically. By colorizing the tumor region based on its stage, we aim to create a system that is easy to use, reliable, and can save time for medical professionals, thereby improving overall patient outcomes.
- The novelty of our project lies in its unique and innovative approach to lung cancer detection, which goes beyond traditional methods and provides valuable information that can aid medical professionals in planning appropriate treatment options. By incorporating the colorization of the tumor region based on its stage, we aim to provide an additional visual aid for medical professionals and improve the accuracy of the diagnosis.

Maintainability:

The proposed solution is designed to be easily maintainable, with regular updates and improvements made to ensure optimal performance.

Portability:

Like interoperability application should be portable across different platforms and devices.

Legacy to Modernization:

The website should be designed with modern web development standards in mind, allowing it to be easily updated and modernized as new technologies emerge.

Reusability and Application Compatibility:

The website will be designed with reusability in mind, making it easy to integrate into different applications and workflows

Interoperability:

The website should be designed and developed to be compatible with different web browsers and devices, ensuring interoperability across different platforms. We recommend using responsive web design techniques to optimize the website's layout and functionality on different screen sizes and resolutions.

Performance:

The website should be optimized for performance, ensuring that it loads quickly and efficiently. This can be achieved by minimizing the size of images and other media files and optimizing code.

Security:

The proposed solution adheres to standard healthcare security protocols to ensure patient data is protected and confidential.

Reliability:

To make our project more reliable, we can take several steps such as thorough testing, effective bug tracking and fixing, high-quality code, continuous integration, and error monitoring. By implementing these measures, we can ensure that our project performs consistently and accurately under different circumstances and user conditions, helping to build trust and confidence among users, reduce the likelihood of errors, and improve the overall user experience.

Appendix A: Definitions, Acronyms and Abbreviations**Appendix B: References**

- [1] - **Lung Cancer Prediction using Machine Learning: A Comprehensive Approach**
- Syed Saba Raoof , M A. Jabbar, Syed Aley Fathima – 2020.
- [2] - **Lung cancer prediction and Stage classification in CT Scans Using Convolution Neural Networks -A Deep learning Model** - V.Deepa, P.Mohamed Fathimal – 2022.
- [3] - **Multi-Stage Lung Cancer Detection and Prediction Using Multi-class SVM Classifier** - Janee Alam , Sabrina Alam , Alamgir Hossain -2022.
- [4] - **Isoline Based Image Colorization** -Adam Popowicz, Bogdan Smolk – 2014.
- [5] - **Lung Cancer Detection and Classification using CT Scan Image Processing** - Nusrat Nawreen , Umma Hany ,Tahmina Islam Department of Electrical and Electronic Engineering-2021.
- [6] - **Lung nodules: A comprehensive review on current approach and management**
- Konstantinos Loverdos, Andreas Fotiadis, Chrysoula Kontogianni, Marianthi Iliopoulou, and Mina Gaga.
- [7]-**Lung Cancer Detection and Prediction of Cancer Stages Using Image Processing**
– S.A.D.L.V. Senarathna ,S.P.Y.A.A. Piyumal ,R. Hirshan,W.G.C.W. Kumara-2021.
- [8]-**Image Acquisition and Pre-processing for Detection of Lung Cancer using Neural Network** - B C Kavitha; K B Naveen -2022.
- [9] -**Proposed methodology for Early Detection of Lung Cancer with low-dose CT Scan using Machine Learning** - Gagan Thakral, Sapna Gambhir , Nagender Aneja – 2022.
- [10] -**A Comparative Study of Image Segmentation Technique applied for Lung Cancer Detection** - Mohd Firdaus Abdullah; Muhammad Safwan Mansor; Siti Noraini Sulaiman; Muhammad Khusairi Osman-2019.

Appendix C: Record of Change History

#	Date	Document Version No.	Change Description	Reason for Change
1.				
2.				
3.				

Appendix D: Traceability Matrix

Project Requirement Specification Reference Section No. and Name.	DESIGN / HLD Reference Section No. and Name.