

**FACULTY OF ENGINEERING**

**DEPARTMENT OF COMPUTER ENGINEERING**

**MULTIVARIATE RISK PREDICTORS FOR LUNG CANCER DIAGNOSIS.**

**By**

**KISITU WILFRED**

**Reg. No: BU/UG/2021/0018**

**Tel: +256 778184388, Email:** [**wktdcourses@gmail.com**](mailto:wktdcourses@gmail.com)

And

**KAKOMO RAYAN AHMED**

**Reg. No: BU/UP/2021/1285**

**Tel: +256 780248288, Email: kakomoahmed@gmail.com**

**Final Project Report in partial fulfillment of the requirements for the award of Bachelor’s Degree of Science in Computer Engineering at Busitema University.**

**April, 2025**

# DECLARATION

I KAKOMO RAYAN AHMED and KISUTU WILFRED hereby declare that this final project report is our original work except where explicit citation has been made and it has not been presented to any Institution of higher learning for any academic award.

Signature: Date:

…………………………. ………………….

KAKOMO RAYAN AHMED

Signature: Date:

……………………… ………………….

KISITU WILFRED

# APPROVAL

This is to certify that the final project report titled “Multivariate Risk Predictors for Lung Cancer Diagnosis.” has been done under my supervision and is now ready for examination.

Signature Date

Dr. OWOMUGISHA GODLIVER

Final Project supervisor

Department of Computer engineering.

# ACKNOWLEDGEMENT

We take this opportunity to extend our sincere gratitude first to the Almighty God who has been the source of love, provision and protection in the entire period of writing this report.

We thank our supervisors Dr. Owomugisha Godliver, and Dr. Steven for the guidance, support and encouragement rendered throughout the writing of this report. Their dedicated and unwavering support and love to help us learn and develop our skills has led to the successful writing of this report. We also thank Dr. Alunyu Andrew for the help in direction of this project idea and guidance at the very start of the idea development.

We thank our beloved parents for the emotional, financial and mental support offered to us throughout this entire period. We thank our fellow classmates, friends and all the staff of Busitema University for providing a conducive environment making it able for us to finish this assignment.

# LIST OF ACRONYMS

GLCM - Gray Level Co-event Matrices

MV-KBC - multi-view information based collective

CT - Computed Tomography

KNN - K-Nearest Neighbors

PCA - Principal Component Analysis

ML - Machine Learning

LCP-CNN - Lung Cancer Prediction Convolutional Neural Network

NLST - National Lung Screening Trial

NCI - National Cancer Institute

CDC - Centers for Disease Control and Prevention

WHO - World Health Organization

ANNs - Artificial Neural Networks

API - Application Programming Interface

# LIST OF TABLES

[Table 1: Showing the comparison between the existing systems basing on strengths, limitations and gaps in and of the system. 9](#_Toc196140738)

[Table 2 Description of attributes 15](#_Toc196140739)

[Table 3 Category of risk factors 16](#_Toc196140740)

[Table 4 Label encoding mapping of categorical with ordinal relationship 19](#_Toc196140741)

[Table 5 Hyper parameters of KNN Classifier 24](#_Toc196140742)

[Table 6 Hyper parameters of Decision tree classifier 25](#_Toc196140743)

[Table 7 Hyper parameters of Random forest classifier 26](#_Toc196140744)

[Table 8 Hyper parameters of AdaBoost Classifier 28](#_Toc196140745)

[Table 9 Hyper parameters of ANN Classifier 30](#_Toc196140746)

[Table 10 Comparison of model Performance 33](#_Toc196140747)

[Table 11 Model Summary of Image Classification model 37](#_Toc196140748)

[Table 12 Tools and libraries for API development 40](#_Toc196140749)

[Table 13 Authentication API documentation 41](#_Toc196140750)

[Table 14 User API documentation 42](#_Toc196140751)

[Table 15 Diagnose API Documentation 42](#_Toc196140752)

[Table 16 Ct Diagnose API Documentation 43](#_Toc196140753)

[Table 17 Status codes 43](#_Toc196140754)

[Table 18 Packages for frontend development using flutter 50](#_Toc196140755)

[Table 19 GitHub repository for project source codes 50](file:///C:\Users\RAYAN\Downloads\EARLY%20LUNG%20CANCER%20DIAGNOSIS%20UPDATED%20II%20%20REPORT.docx#_Toc196140756)

[Table 20 Resources to entire project 66](#_Toc196140757)

# LIST OF FIGURES.

Figure 1 Distribution of Attributes 17

Figure 2 Visualization of missing values 18

Figure 3 Distribution of targets 20

Figure 4 Mutual information scores 21

Figure 5 Correlation Matrix 22

Figure 6 Performance report of KNN Classifier 24

Figure 7 Performance of the Decision tree classifier 25

Figure 8 Performance report of Random Forest classifier 27

Figure 9 Performance report of AdaBoost Classifier 28

Figure 10 Confusion matrix of the four machine learning models 29

Figure 11 ANN model Summary 30

Figure 12 Performance report of ANN Classifier 31

Figure 13 ANN Training and validation accuracy 31

Figure 14 CT, PET scans of Lung Cancer patients 34

Figure 15 Class distribution of Imagery dataset 35

Figure 16 Total number of model parameters 37

Figure 17 Training parameters of CT classification image model 37

Figure 18 Training and validation accuracy of the Cancer Variant Classification model 38

Figure 19 Entity Relation Diagram of the Backend service 41

Figure 20 Postman showing the API endpoints developed 44

Figure 21: logging into the remote server 44

Figure 22: successfully logged into server. 45

Figure 23: system information of the remote server. 45

Figure 24: kernel information of the remote server. 46

Figure 25: RAM information of the remote server. 46

Figure 26: disk information of the remote server. 46

Figure 27 Wireframes for the frontend application 48

Figure 28 Mutual information of each risk factor in descending order 55

# TABLE OF CONTENTS.

Contents

[DECLARATION i](#_Toc196259699)

[APPROVAL ii](#_Toc196259700)

[ACKNOWLEDGEMENT iii](#_Toc196259701)

[LIST OF ACRONYMS iv](#_Toc196259702)

[LIST OF TABLES v](#_Toc196259703)

[LIST OF FIGURES. v](#_Toc196259704)

[TABLE OF CONTENTS. vii](#_Toc196259705)

[CHAPTER ONE: INTRODUCTION. 1](#_Toc196259706)

[1.1 Background. 1](#_Toc196259707)

[1.1.1 Why machine learning? 1](#_Toc196259708)

[1.2 Problem statement. 2](#_Toc196259709)

[1.3 Objectives of study 3](#_Toc196259710)

[1.3.1 Main objective 3](#_Toc196259711)

[1.3.2 Specific objectives. 3](#_Toc196259712)

[1.4 Justification 3](#_Toc196259713)

[1.5 Significance 4](#_Toc196259714)

[1.6 Scope of study 4](#_Toc196259715)

[1.6.1 Content scope. 4](#_Toc196259716)

[1.6.2 Geographical scope. 4](#_Toc196259717)

[1.6.3 Time scope. 5](#_Toc196259718)

[1.6.4 Unit of measurement 5](#_Toc196259719)

[CHAPTER TWO: LITERATURE REVIEW 6](#_Toc196259720)

[2.1 INTRODUCTION 6](#_Toc196259721)

[2.2 RELATED WORKS. 6](#_Toc196259722)

[2.3 EXISTING WORKS. 8](#_Toc196259723)

[2.4 COMPARISON OF THE EXISTING SYSTEMS 9](#_Toc196259724)

[2.5 DEVELOPED SYSTEM 11](#_Toc196259725)

[CHAPTER THREE: METHODOLOGY 13](#_Toc196259726)

[3.0 INTRODUCATION 13](#_Toc196259727)

[3.1 SPECIFIC OBJECTIVE ONE: TO IDENTIFY AN EXISTING DATASET THAT CAPTURES THE LIFESTYLE, DEMOGRAPHICS, MEDICAL HISTORY AND CLINICAL SYMPTOMS OF A PATIENT. 13](#_Toc196259728)

[3.1.1 Exploration of Dataset Repositories 13](#_Toc196259729)

[3.2 SPECIFIC OBJECTIVE TWO: TO EXTRACT FEATURES FOR EARLY LUNG CANCER DIAGNOSIS 14](#_Toc196259730)

[3.2.1 Data Processing and Preparation 14](#_Toc196259731)

[3.2.2 Mutual Information Analysis 17](#_Toc196259732)

[3.2.3 Correlation Matrix Analysis 17](#_Toc196259733)

[3.3 SPECIFIC OBJECTIVE THREE: MODEL TRAINING EVALUATION OF MACHINE LEARNING MODELS 18](#_Toc196259734)

[3.3.1 Development of Various Machine Learning Models 18](#_Toc196259735)

[3.3.2 Hyper parameter Tuning 23](#_Toc196259736)

[3.3.3 Model Evaluation 23](#_Toc196259737)

[3.3.4 Model Validation Using Image-Based Analysis 24](#_Toc196259738)

[3.2 SPECIFIC OBJECTIVE FOUR: TO DEVELOP A MOBILE APPLICATION INTERFACE AND INTEGRATE IT WITH MACHINE LEARNING MODEL 30](#_Toc196259739)

[3.4.2 Mobile User Interface Development 35](#_Toc196259740)

[3.4.4 Unit and Integration Testing 39](#_Toc196259741)

[CHAPTER FOUR: TESTING, RESULTS AND DISCUSSIONS 41](#_Toc196259742)

[1.2 DISCUSSION OF RESULTS. 41](#_Toc196259743)

[4.2.1 Most Suitable Data Selection 41](#_Toc196259744)

[Dataset Description and structure. 42](#_Toc196259745)

[CHAPTER FIVE: RECOMMENDATIONS AND CONCLUSION 59](#_Toc196259746)

[5.1 Challenges 59](#_Toc196259747)

[5.2 Feedback 60](#_Toc196259748)

[5.3 Recommendations 61](#_Toc196259749)

[5.3 Conclusion 63](#_Toc196259750)

[**References** 64](#_Toc196259751)

[APPENDIX 66](#_Toc196259752)

[Lung Cancer Diagnosis Using Multi Risk Factor Analysis 67](#_Toc196259753)

[Lung-PET-CT-Dx | A Large-Scale CT and PET/CT Dataset Second Training 67](#_Toc196259754)

[Mobile application mockups 67](#_Toc196259755)

[Implementation of Restful APIs 67](#_Toc196259756)

[Implementation of mobile application 67](#_Toc196259757)

# CHAPTER ONE: INTRODUCTION.

## 1.1 Background.

Globally, lung cancer has been one of the most common malignancies in the last few decades, with the highest incidences and is the leading cause of death. In 2018, there was approximately 2.1 million new lung cancer diagnosis accounting for 12% of the global cancer burden [1]. In 1950, cigarette smokers were found more likely to be diagnosed with lung cancer [2]. However research has found other risk factors that contribute to lung cancer such as toxic environment, exposure to radon, immune system diseases, and indoor air pollution from cooking and heating residues, and genetics [3]. Notably, the 5-year survival rate for patients with lung cancer is low, at 18%. However if early diagnosis can be achieved, the survival rate can be increased to approximately 55%. It has been reported that patients with early-stage cancer have a 5-year survival rate of up to 40% if they receive appropriate treatment [4]. Unfortunately, over 70% of patients are diagnosed when their tumor has progressed to an advanced stage, and most of these cases are not suitable for surgery. This is related to the fact that existing diagnosis methods are not sensitive and accurate enough. The current gold standard for diagnosis of lung cancer is CT-guided transthoracic aspiration biopsy. However it is expensive and has high risks of pulmonary embolism and significant trauma. Other diagnostic methods, such as; blood tumor biomarkers and bronchoscopy, for lung cancer screening still have limitations [5]. Advances in machine learning have provided tools to assist in detection of lung cancer based on CT-scans, X-ray, and metabolomics. These machine learning models use computers to analyze, model and train a large amount of medical data to reveal relationship between various medical indicators [6].

### 1.1.1 Why machine learning?

In cancer, machine learning has been used to explore survival and prognosis prediction models for pancreatic, bladder, advanced nasopharyngeal and breast cancer [7]. Models such as; XGBoost models have been applied to identify lung cancer, colon cancer subtypes [8], prediction of lung cancer metastases from thyroid cancer, and risk models for identifying lung cancer, with all performing remarkably. In context of early diagnosis of lung cancer, several models have been developed but they still have limitations; Guan et al [9] proposed an XGBoost model for early lung cancer prediction based on metabolic indices. The model achieved a high sensitivity of 74% with 75.29% accuracy, identifying metabolic biomarkers ornithine and palmitoylcarnitine as potential biomarkers to screen for lung cancer. However, the proposed model lacked external validation from public datasets, few demographic indicators were used, and further laboratory studies in the biological mechanism of the identified biomarkers were required to better validate the model [9]. Wiratama, Pangga Kurnia Patra [11] proposed a novel Random Forest based Risk Factor analysis model for lung cancer Prediction using combination of lifestyle, environmental data and health conditions of a patient. The model achieved 99% accuracy and 100% recall through k-fold cross validation. However, the proposed model completely excluded demographic variables, medical history, and clinical symptoms consistent with lung cancer which are crucial in clinical diagnosis of the disease [10]. To address these limitations, we propose a novel machine learning framework that integrates demographics, lifestyle, medical history and symptoms consistent with lung cancer in early detection of the disease. This approach aims to integrate multiple risk factors to develop a machine learning model with minimal false positives and enhanced predictive sensitivity, and accuracy especially for high risk patients.

## 1.2 Problem statement.

Notably, the 5 year survival rate of lung cancer patients is still low at 8% due to over 70% of the patients diagnosed at advanced stages (III, IV) where surgery is not suitable [5]. However, if diagnosed early, the survival rate can be increased to approximately 55%. Advances in machine learning have provided promising tools for clinical diagnosis by analyzing large medical data. However these still exhibit limitation. For example; Guan et al [9] proposed an XGBoost model for prediction of lung cancer based on metabolic indices with a high a sensitivity. However the biomarkers used required further laboratory studies in their biological mechanism to better validate the model. Wiratama [11] proposed a Random Forest based risk factor analysis model using a combination of lifestyle, environmental conditions, demographics and medical conditions. This achieved 99% accuracy. However they eliminated demographics, lacked medical history and clinical symptoms of patients in their study. Failure to address these limitations of current diagnostic methods and machine learning models has significant consequences such as; late stage diagnosis, increased false positives, and low sensitivity models. This study aims to fill in these gaps by developing a machine learning framework that integrates lifestyle, demographics, medical history, and clinical symptoms of a patient for more sensitive and accurate early detection of lung cancer.

## 1.3 Objectives of study

### 1.3.1 Main objective

To develop a comprehensive machine learning framework that integrates lifestyle, demographics, medical history, and clinical symptoms of patients into a single multiple risk factor analysis model for early diagnosis of lung cancer.

### 1.3.2 Specific objectives.

1. To identify an existing dataset that captures the lifestyle, demographics, medical history, and clinical symptoms of lung cancer patients
2. To extract features for early lung cancer diagnosis
3. To develop a machine learning model and evaluate its performance based on; accuracy, sensitivity, and false positive through k-fold cross validation
4. To develop a mobile application interface and integrate it with machine learning model

## 1.4 Justification

Over 70% of patients are diagnosed at advanced stages of lung cancer [5], where the 5-year survival rate is low at approximately 18%. This study stems from the urgent need to improve the survival rates of lung cancer patients by addressing the limitations in current diagnostics methods such as CT-guided transthoracic aspiration biopsy, blood tumor biomarkers and bronchoscopy, which are not accurate and sensitive enough for early diagnosis. Furthermore, the existing machine learning models such as; XGBoost model for early lung cancer prediction based on metabolic indices proposed by Guan at el [9], and Random Forest based Risk Factor analysis model for lung cancer Prediction using combination of lifestyle, environmental data and health conditions proposed by Wiratama, Pangga Kurnia Patra [11], exclude critical factors such as; demographics, medical history, and clinical symptoms which are essential for accurate early diagnosis of lung cancer. This study aims to fill in these gaps by developing a comprehensive machine learning model which integrates lifestyle, demographics, medical history, and clinical symptoms into a single multiple risk factor analysis model for early diagnosis of lung cancer. The results of this study will be applicable to clinical diagnosis of lung cancer in health settings. The machine learning model will provide a tool that can potentially reduce late-stage lung cancer diagnosis, and provide a less invasive diagnostic procedure for high risk individuals such as; active smokers. The research will also inform future studies related to developing advanced predictive tools in lung cancer diagnosis.

## 1.5 Significance

This study aims to assess the significance of clinical symptoms, and medical history of patients as integral risk factors in the diagnosis of lung cancer using machine learning techniques. The existing diagnostic models such as XGBoost model for lung cancer prediction based on metabolic indices, and Random Forest based Risk factor analysis model show remarkable performance but exclude; symptoms and medical history of patients which are essential in clinical diagnosis especially when combined with the lifestyle, environmental conditions, and demographics of individuals. By combining a broad set of risk factors that capture the lifestyle, demographics, medical history, and clinical symptoms of a patient, this study enhances the depth and scope of predictive modelling in early lung cancer diagnosis. From theoretical standpoint, this study expands on the existing machine learning frameworks by integrating clinical symptoms, and medical history into predictive models which provides insights into how diverse variables interact in the early detection of lung cancer. This study employs a methodologically rigorous approach by utilizing; feature selection, model regularization, hyper parameter tuning, and k-fold cross validation techniques to ensure the results are generalized. This study aims to contribute to the adoption of machine learning tools in clinical and health settings as well as inspire further studies in early diagnosis of lung cancer thereby reducing the societal burden of the disease.

## 1.6 Scope of study

### 1.6.1 Content scope.

This study aims to develop a comprehensive machine learning framework that integrates lifestyle, demographics, medical history, and clinical symptoms of patients into a single multiple risk factor analysis model for early diagnosis of lung cancer. The study will utilize an existing dataset that captures these risk factors, providing an analysis of how they interact to influence progression and risk assessment of lung cancer.

### 1.6.2 Geographical scope.

This study will focus on lung cancer patients and high risk individuals in Uganda. This will include population exposed to risk factors such as; smoking, indoor air pollution from cooking and burning residues, exposure to random, and genetic disposition.

### 1.6.3 Time scope.

The research will be conducted over a period of four (4) months during which an existing dataset will be identified and analyzed, and a machine learning model which integrates the suggested risk factors will be developed and evaluated.

## 1.6.4 Unit of measurement

The unit of measurement will be individual lung cancer patients and those exposed to high risk factors of the disease. The identified existing dataset will capture the lifestyle, demographics, medical history, and clinical symptoms of individuals that will be utilized to develop and evaluate a machine learning model with an appreciable sensitivity and accuracy for early lung cancer diagnosis.

# CHAPTER TWO: LITERATURE REVIEW

## 2.1 INTRODUCTION

This section includes critical review of previous work and systems related to the proposed system as well as an analysis of the existing knowledge related to the study and the technologies to be used in the proposed system. The review includes research work from journals and books cited with the objective of revealing contributions, weakness and gaps within the subject.

## 2.2 RELATED WORKS.

With recent advancements in Machine Learning, it is now possible to predict and diagnose various medical data with Machine Learning and Random Forest techniques [11, p. 10]. Specifically, cancer prevention and management can be hugely supported with the use of Machine Learning [12].

YutongXie, states that a multi-view information based collective (MV-KBC) deep model was used to isolate malignant tumor from normal lung nodules utilizing chest CT information. They used 9 KBC [13] sub-models to train the model. The model was tested on LIDC-IDRI data set and compared with the five modern classification approaches.

LilikAnifah et.al proposed the detection of lung cancer utilizing Artificial Neural Network Back-propagation based Gray Level Co-event Matrices (GLCM) features. The lung information is utilized from the Cancer imaging archive Database, comprised of 50 CT-pictures. The steps of this process are: image pre-processing, segmentation, feature extraction, and recognition of tumor growth using Neural Network Back-propagation method which has 3 layers. The result showed that framework [14] can differentiate between ordinary lung and lung malignancy with accuracy of over 80%.

Pudjihartono et al. [15]discuss various techniques of feature selection for disease risk prediction. They describe multiple feature selection techniques and compare them in terms of evaluation metrics, computational complexity, and the potential to detect redundancies between features. They conclude that each method has its strengths and weaknesses, and the best practice is to use several methods or to combine them and use a hybrid approach.

J. Chen compares different machine learning models for lung cancer prediction, including K-Nearest Neighbors (KNN), Random Forest, Logistic Regression, K-Means clustering, and Principal Component Analysis (PCA). Previous related works have used neural networks, deep learning, and other ML techniques for lung cancer diagnosis and risk estimation. However, this paper focuses specifically on selecting the optimal ML approach for lung cancer prediction using a dataset with features like smoking, age, gender etc. After experimentally comparing the performance of the different algorithms using evaluation metrics like accuracy, AUC, F1-score, the Random Forest model emerges as the best performer. The paper attributes this to Random Forest's ensemble nature making it suitable for classification tasks, as well as its ability to reduce over fitting. The analysis provides insights into the applicability of different ML algorithm types for the lung cancer prediction problem.

M. Heuvelmans et al [9] validates a deep learning model called the Lung Cancer Prediction Convolutional Neural Network (LCP-CNN) for identifying benign lung nodules detected on CT scans. The LCP-CNN was previously trained on data from the National Lung Screening Trial (NLST) in the United States. In this study, the authors validated the LCP-CNN on an independent multi-center dataset of 2,106 lung nodules from Europe, including 205 malignant nodules. They found that the LCP CNN could correctly rule out malignancy in 22.1% of nodules while maintaining a sensitivity of 99.0%, allowing 18.5% of patients to avoid unnecessary follow-up scans. This builds on previous studies that showed the LCP-CNN can outperform other risk prediction models. The results demonstrate the potential of using deep learning to guide clinical decision-making for incidental lung nodules detected on CT scans.

Various studies have shown that Random Forest produce good results: [4] shows that it produced one of the best accuracies for Alzheimer’s disease prediction, and, as shown in [16], it is possible to use the Random Forest technique to diagnose breast cancer with 100% accuracy.

Prof. AnuradhaDeshpande and DhaneshLokhande [16] focused on lung cancer prediction using image processing strategies followed by watershed segmentation and SVM. In this combination procedure, the critical characteristics of various pictures are consolidated together to acquire the required data in a Fused Image format. CT picture examines the denser tissues and MRI filters the delicate tissues, so by joining pertinent data of the two pictures, proper data of melded picture is obtained. This procedure additionally enhances the quality of the melded picture.

“A Lightweight Deep Learning Model for Automatic Diagnosis of Lung Cancer”. The project focuses on developing a lightweight deep learning model for the automatic diagnosis of lung cancer. Extensive research and review have shown that the model is effective and efficient at accurately categorizing lung cancer cases based on medical imaging data. The proposed approach appears to be a promising method for speeding up and improving the accuracy of lung cancer diagnosis. This could result in early detection and better outcomes for patients. Further testing and approval in clinical settings are required to fully assess the therapeutic utility and versatility of the suggested paradigm.

## 2.3 EXISTING WORKS.

A research conducted by E. Dritsas [17] uses importance scores to research the relevant features and uses multiple classifiers to predict lung cancer in patients. The study shows that age, allergy, alcohol, and wheezing show high correlation with lung cancer.

D. Endalie [18] analyses lung cancer risk factors using tree-based ranking algorithm and proposes an ML model to predict cancer severity in medical records in Ethiopia. They claim that blood coughing, air pollution, and obesity are the most severe lung cancer risk factors. This research provides the most important risk factors, but only for a very restricted geographic region.

LUNA16. The Lung Nodule Analysis (LUNA16) challenge was a widely recognized competition in the field of lung cancer detection from CT scans. While it's not a system per se, it provided a benchmark dataset and a platform for researchers to develop and evaluate lung cancer prediction algorithms. Many research papers and code implementations have been built upon this challenge.

DeepLung. DeepLung is an open-source project developed by researchers at the University of Central Florida, aimed at detecting lung nodules from CT scans using deep learning techniques. It provides a TensorFlow-based implementation of various deep learning models for lung nodule detection and classification.

CheXNet. Although primarily focused on chest X-ray analysis for various pathologies, including pneumonia, CheXNet, developed by researchers at Stanford University, demonstrated the potential of deep learning in medical image analysis. While not specifically for lung cancer prediction, the techniques and methodologies used in CheXNet could be adapted for similar tasks.

IBM Watson for Oncology. IBM Watson for Oncology is a commercial system that utilizes artificial intelligence, including deep learning algorithms, to assist oncologists in making treatment decisions for cancer patients. While it's not solely focused on lung cancer prediction, it showcases the integration of AI into clinical decision-making processes.

Google AI's Research on Lung Cancer Prediction. Google's DeepMind Health team has conducted research on using deep learning algorithms to predict lung cancer from CT scans. While their work is primarily research-oriented, it highlights the potential of deep learning in improving early detection and diagnosis of lung cancer.

## 2.4 COMPARISON OF THE EXISTING SYSTEMS

Table 1: Showing the comparison between the existing systems basing on strengths, limitations and gaps in and of the system.

|  |  |  |  |
| --- | --- | --- | --- |
| **System** | **Strengths** | **Limitations** | **gaps** |
| **LUNA 16** | Focuses specifically on nodule detection, enhancing sensitivity and specificity in identifying lung nodules from CT scans  Provides a platform for researchers to benchmark their algorithms against a standardized dataset, promoting further advancements in nodule detection technologies | High false positive rates can lead to unnecessary follow-ups and anxiety for patients  Challenges in detecting nodules in peripheral and apical regions of the lungs | Limited capabilities in assessing the malignancy of detected nodules; it primarily focuses on detection rather than classification |
| **DeepLung** | Utilizes multiple CNN architectures to improve the accuracy of nodule detection and classification, thereby reducing false negatives  Designed to provide detailed segmentation of lung nodules, which aids in distinguishing benign from malignant growths | Computationally intensive, requiring significant resources for training and inference, which may limit accessibility in low-resource settings  Performance can be affected by the quality and diversity of training datasets | More research is needed to optimize the model for real-time applications in clinical settings |
| **CheXNet** | Achieves performance on par with radiologists for detecting various pathologies in chest X-rays, making it versatile  Offers the potential for early detection of multiple diseases, not limited to lung cancer | Primarily designed for 2D X-ray images, which may not provide the detail needed for accurate lung nodule assessment compared to CT scans  Limited focus on lung cancer specifically, as it targets broader respiratory conditions | Further development is needed to enhance its application specifically for lung cancer diagnosis, including differentiating between benign and malignant nodules |
| **IBM Watson for Oncology** | Integrates a vast amount of medical literature and patient data to provide personalized treatment recommendations, enhancing clinical decision-making  Utilizes a holistic approach, considering various patient factors, which can improve treatment outcomes​ | More focused on treatment guidance than direct diagnosis or nodule detection  Performance can vary based on the data quality and the specific cancer types it has been trained on​ | Limited availability and adaptation in diverse clinical settings, particularly in low-resource environments​ |
| **Google AI's Research on Lung Cancer Prediction** | Uses advanced deep learning algorithms to analyse 3D CT scans, outperforming traditional models in some cases​  Demonstrates strong performance in predicting malignancy over time, which is crucial for early detection | The complexity of the model requires substantial computational power, which may not be feasible in all clinical settings  May have a limited dataset in terms of diversity, affecting generalizability | More research is needed to refine its capabilities for broader application in diverse populations and clinical contexts |

## 2.5 DEVELOPED SYSTEM

This machine learning framework integrates lifestyle, demographics, medical history and clinical symptoms into a single multiple risk factor analysis model for early detection of lung cancer. It utilizes thirteen (21) risk factors / attributes of a patient, these include; smoking history, yellowing of fingers, anxiety, peer pressure, chronic disease, fatigue, allergy, wheezing, alcohol consumption, coughing, shortness of breath, swallowing difficulty, chest pain and others. The system employs;

* Data preprocessing to handle missing data points, categorical values, and normalize data
* Feature extraction to identify key risk factors through correlation matrix analysis
* Training and testing of various machine learning models such as; SVMs, XGBoost models, decision trees, random forest model, and an artificial neural network
* Accessing the performance of the models based on False Positive rate, F1 score, recall, precision, and accuracy through k-fold cross validation
* Visualization of performance of models and selection of the best performing model on the dataset

The diagram below illustrates the machine learning framework for early diagnosis of lung cancer.

Patient Data (coughing, wheezing shortness of breath, chest pain, yellow fingers, age, gender, etc.)

Data pre-processing

(Handling categorical, and missing values, normalization, splitting)

Feature extraction

(Correlation matrix analysis)

Machine learning Model Training

SVM

ANN

XgBoost

Decision tress

Random Forest

Visualization of performance of the models and selection of best performing model

Model evaluation and comparison (based on false positive rate, F1 score, precision, and accuracy using k-fold cross validation)

# CHAPTER THREE: METHODOLOGY

## 3.0 INTRODUCATION

This chapter explains the methods, procedures and strategies that were used to meet the objectives of this project.

## 3.1 SPECIFIC OBJECTIVE ONE: TO IDENTIFY AN EXISTING DATASET THAT CAPTURES THE LIFESTYLE, DEMOGRAPHICS, MEDICAL HISTORY AND CLINICAL SYMPTOMS OF A PATIENT.

To achieve the objective of identifying a suitable dataset for lung cancer diagnosis using risk factors, the following steps were undertaken:

### ****3.1.1 Exploration of Dataset Repositories****

A thorough search was conducted across multiple reputable data repositories, including the

* **Public Health** and Prevention (CDC), and World Health **and Epidemiological Databases.** The National Cancer Institute (NCI), National Lung Screening Trial (NLST), Centers for Disease Control Organization (WHO) databases were explored for structured epidemiological datasets containing lung cancer-related attributes.
* **Open Source Data Repositories.** Kaggle, UCI Machine Learning Repository, and other publicly available machine learning datasets were reviewed to identify datasets containing patient risk factors, symptoms, and demographic details.

During this search, datasets were assessed based on their availability, comprehensiveness, and diversity of patient data collected in regards to the objectives of the study

## 3.2 SPECIFIC OBJECTIVE TWO: TO EXTRACT FEATURES FOR EARLY LUNG CANCER DIAGNOSIS

This objective was focused on preprocessing of the dataset for future analysis, feature extraction of key risk factors for lung cancer diagnosis, and for model training, evaluation and testing. This mainly involved data preprocessing, mutual information analysis, and correlation matrix analysis.

### ****Data Processing and Preparation****

Before extracting relevant features, the dataset underwent preprocessing to ensure high-quality data and reliable analysis. The following steps were implemented

**a) Handling Missing Values**

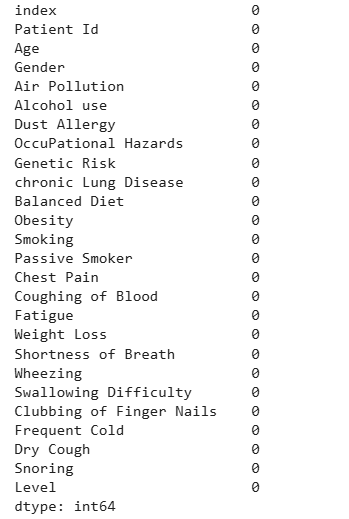
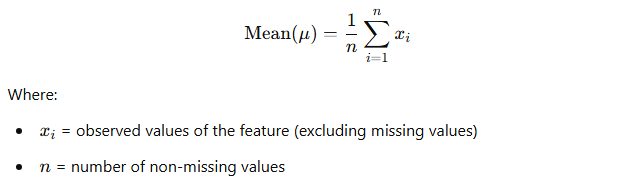
The dataset was first examined for missing value by summation of all Nan (Not a Number) values in each column. The following results were obtained.

Figure 2 Visualization of missing values

From the above figure, all columns contained no missing values. Therefore, there was no relevancy for treatment of missing values in the dataset. However, in case of any, we would improvise mean;

Where; x = observed values of the feature, n = number of non-missing values

##### **Handling of categorical variables**

We applied the **label encoding** technique to convert categorical variables into numerical form. This approach was appropriate due to the **ordinal nature** of the data, where the attributes reflected increasing degrees of severity. The categories—**very low**, **low**, **moderate**, **high**, and **very high**—were encoded in a way that preserved their **natural order**, enabling the model to interpret the progressive severity of each feature accurately. The label encoding would map attributes in a way illustrated below

|  |  |
| --- | --- |
| **Category** | **Encoded Value** |
| extremely low | 1 |
| very low | 2 |
| low | 3 |
| moderate | 4 |
| high | 5 |
| very high | 6 |
| extremely high | 7 |

Table 4 Label encoding mapping of categorical with ordinal relationship

##### **Normalization and Scaling**

To ensure uniformity across features, numerical attributes were scaled using the **Min-Max Scalar**. This method transforms values to a fixed range, typically [0, 1] making the data suitable for machine learning models that rely on distance-based computations. The Min-Max Scaling formula is:

Where X is original value, Xmin and Xmax are the minimum and maximum values of features respectively

##### **c) Handling Imbalanced Data**

An **imbalance check** was conducted to examine the distribution of low, moderate, and high risk cases of lung cancer patients. This step is crucial to prevent the model from becoming biased toward the **majority class**, which can lead to poor performance in predicting the **minority class**. The distribution of the target classes is illustrated below

Figure 3 Distribution of targets

The pie chart above shows that the target classes are **equally distributed**, indicating a balanced dataset. In cases where **class imbalance** is detected, we would apply techniques like the **Synthetic Minority Oversampling Technique (SMOTE)**, which addresses the issue by generating **synthetic samples** for the minority class rather than simply duplicating existing ones. Formula;

Where

##### **d) Dataset Splitting**

To enable effective model evaluation, the dataset was divided into **75% for training** and **25% for testing**. This split ensured the model had enough data to learn from while preserving a dedicated portion for assessing its performance on unseen data

### ****3.2.2 Mutual Information Analysis****

To identify the most informative features for lung cancer diagnosis**, Mutual Information (MI) Analysis** was applied. The following steps were performed:

1. **Computation of MI Scores;** mutual information scores were calculated for each independent variable against the target variable (lung cancer diagnosis).
2. **Feature Ranking;** features were ranked based on their MI scores to determine their relevance.

### ****3.2.3 Correlation Matrix Analysis****

To evaluate the relationships between variables, a **correlation matrix** was computed and visualized using a **heat map**, allowing for quick identification of both strong and weak correlations. Based on this analysis, features that showed a **high correlation with the target variable** and **highly correlated independent features** were identified

## 3.3 SPECIFIC OBJECTIVE THREE: MODEL TRAINING EVALUATION OF MACHINE LEARNING MODELS

This objective details the approaches we used to train and evaluate various machine learning models for early lung cancer diagnosis. These included the following;

### ****3.3**.**1 Development of Various Machine Learning Models****

To determine the most effective model for predicting lung cancer based on risk factors, multiple machine learning models were developed and trained both traditional machine learning models and a deep learning model which handles complex and high dimensional datasets

**Traditional Machine Learning Models Trained**

1. **KNN Classifier;** this is a non-parametric, and instance-based machine learning algorithm used for classification tasks. It works by comparing a new data point to the 'k' closest data points in the training set (based on a distance metric like Euclidean distance) and assigns the class that is most common among those neighbors. **The following provides a summary of the parameters we used to train and test the model**

|  |  |
| --- | --- |
| Aspect | Parameter |
| Model | K-Nearest Neighbors (KNN) Classifier |
| Tuning Method | RandomizedSearchCV |
| Scoring Metric | Accuracy |
| Cross-Validation (CV) | 10-fold |
| Hyper parameter Tuned | n\_neighbors |
| Search Range | 2 to 49 |
| Best Parameter Found | n\_neighbors = X *(replace with actual)* |
| Best Cross-Validation Score | Y *(replace with actual)* |
| Final Model Used | Best estimator from RandomizedSearchCV |

Table 5 Hyper parameters of KNN Classifier

1. **Decision Tree Classifier; is a supervised machine learning algorithm, also used for classification tasks. It works by splitting the dataset into subsets based on the value of input features, forming a tree-like structure where each internal node represents a decision on a feature, each branch represents an outcome, and each leaf node represents a class label. The following provides a summary of the parameters we used to train and test this model**

|  |  |
| --- | --- |
| Aspect | Details |
| Model | Decision Tree Classifier |
| Tuning Method | GridSearchCV |
| Scoring Metric | Accuracy |
| Cross-Validation (CV) | 10-fold |
| Hyper parameter Tuned | ccp\_alpha (Cost-Complexity Pruning Alpha) |
| Search Range | Values from dt.cost\_complexity\_pruning\_path() |
| Best Parameter Found | ccp\_alpha = X *(replace with actual value)* |
| Best Cross-Validation Score | Y *(replace with actual value)* |
| Final Model Used | DecisionTreeClassifier with ccp\_alpha = X |

Table 6 Hyper parameters of Decision tree classifier

1. **Random Forest Classifier; is an ensemble learning algorithm that builds multiple decision trees and combines their outputs to improve classification accuracy and robustness. Each tree is trained on a random subset of the data and features, introducing diversity and reducing over fitting compared to a single decision tree. During prediction, each tree "votes" for a class, and the final output is determined by majority voting. The following provides a summary of the parameters we used to train and test this model**

|  |  |
| --- | --- |
| Aspect | Details |
| Model | Random Forest Classifier |
| Tuning Method | RandomizedSearchCV |
| Scoring Metric | Accuracy |
| Cross-Validation (CV) | 10-fold |
| Hyper parameters Tuned | criterion, min\_samples\_split, min\_samples\_leaf, max\_features, n\_estimators |
| Parameter Search Space | criterion: ['gini', 'entropy']min\_samples\_split: 2 to 40 min\_samples\_leaf: 2 to 40, max\_features: ['sqrt', 'log2', None], n\_estimators: 1000 |
| Best Parameters Found | *(replace with nrf.best\_params\_ output)* |
| Best Cross-Validation Score | *(replace with nrf.best\_score\_ output)* |
| Final Model Used | Best estimator from RandomizedSearchCV |

Table 7 Hyper parameters of Random forest classifier

1. **Ada Boost Classifier; is an ensemble learning technique that combines multiple weak learners (typically shallow decision trees) to create a strong classifier. It works by training models in sequential steps, where each new model focuses more on the previously misclassified instances by assigning them higher weights.**

**With each iteration, AdaBoost adjusts the weights of training samples and combines the predictions of all models through a weighted majority vote. The following provides a summary of the parameters we used to train and test this model**

|  |  |
| --- | --- |
| Aspect | Details |
| Model | AdaBoost Classifier |
| Boosting Algorithm | SAMME |
| Tuning Method | RandomizedSearchCV |
| Scoring Metric | Accuracy |
| Cross-Validation (CV) | 10-fold |
| Hyper parameters Tuned | n\_estimators, learning rate |
| Parameter Search Space | n\_estimators: 300, learning rate: 0.01 to 2.0 (step 0.01) |
| Best Parameters Found | *(replace with nada.best\_params\_ output)* |
| Best Cross-Validation Score | *(replace with nada.best\_score\_ output)* |
| Final Model Used | Best estimator from RandomizedSearchCV |

Table 8 Hyper parameters of AdaBoost Classifier

**Deep Learning Model**

**We developed an Artificial Neural Network (ANN) which can capture complex patterns and high-dimensional relationships among risk factors. The model was trained using the processed dataset, and initial performance metrics were recorded to establish a baseline. The ANN architecture was kept flexible to allow tuning of hidden layers and activation functions for optimal performance. The following is a summary of the model architecture;**

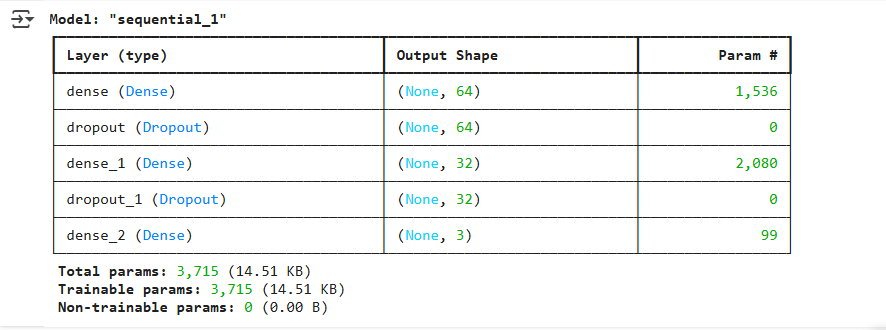
**The following table illustrates the model parameters used to compile and train the ANN**

Figure 6 ANN model Summary

|  |  |
| --- | --- |
| Parameter | Value |
| Loss Function | sparse\_categorical\_crossentropy |
| Optimizer | Adam |
| Learning Rate | 0.001 |
| Metrics | Accuracy |
| Epochs | 50 |
| Batch Size | 32 |
| Validation Split | 0.2 (20% of training data) |

Table 9 Hyper parameters of ANN Classifier

**The following illustrates the training progress of the model’s training and validation accuracy.**

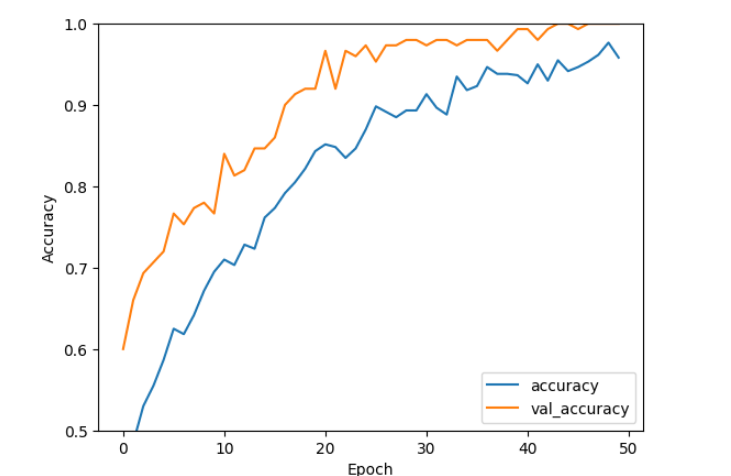
****

Figure 7 ANN Training and validation accuracy

### ****3.3**.**2 Hyper parameter Tuning****

For the traditional machine learning models, we obtained the optimal initial parameters through grid and random search methods as illustrated in each models hyper parameters selected. For each model we focused on obtaining the optimal; learning rate, number of estimators for ensemble models, and the type of kernel for Support Vector Classifiers. For the ANN model, we performed auto tuning using the keras auto tune library to determine the optimal learning rate, activation function and epochs

### ****3.3.3 Model Evaluation****

From the presented performance reports, we evaluated each model using 10 **K-Fold cross-validation** to ensure robust performance across different subsets of the dataset. The following key metrics where focused on

1. **Accuracy;** measured the proportion of correctly diagnosed cases.
2. **False Positive Rate (FPR);** assessed the proportion of high risk individuals incorrectly diagnosed with lung cancer.
3. **Recall (Sensitivity); d**etermined the proportion of correctly identified lung cancer patients among actual cases.
4. **F1 Score;** provided a balanced measure of precision and recall.

### ****3.3.4 Model Validation Using Image-Based Analysis****

To validate the effectiveness of the risk-factor-based model, a separate CT scan image-based classification model was developed for comparison. This involved training an independent deep learning model for medical image analysis to assess the consistency and reliability of the risk-factor predictions. The overarching goal was to establish a unified framework for early lung cancer diagnosis, enabling patients to first assess their risk level and, if identified as high-risk, proceed to detect the likely variant of lung cancer using the same system. The following steps were undertaken

1. **Dataset Selection**

We utilized the publicly available “[Lung PET CT Dx](https://www.kaggle.com/datasets/rangan2510/lung-pet-ct-dx)” dataset from Kaggle, which contains **13,494 images**, including CT, PET, and fused scans. The images are categorized into three major variants of lung cancer: **Adenocarcinoma, Small Cell Carcinoma,** and **Squamous Cell Carcinoma**. This dataset is specifically curated for the classification of lung cancer types based on imaging, making it highly suitable for our study objectives.

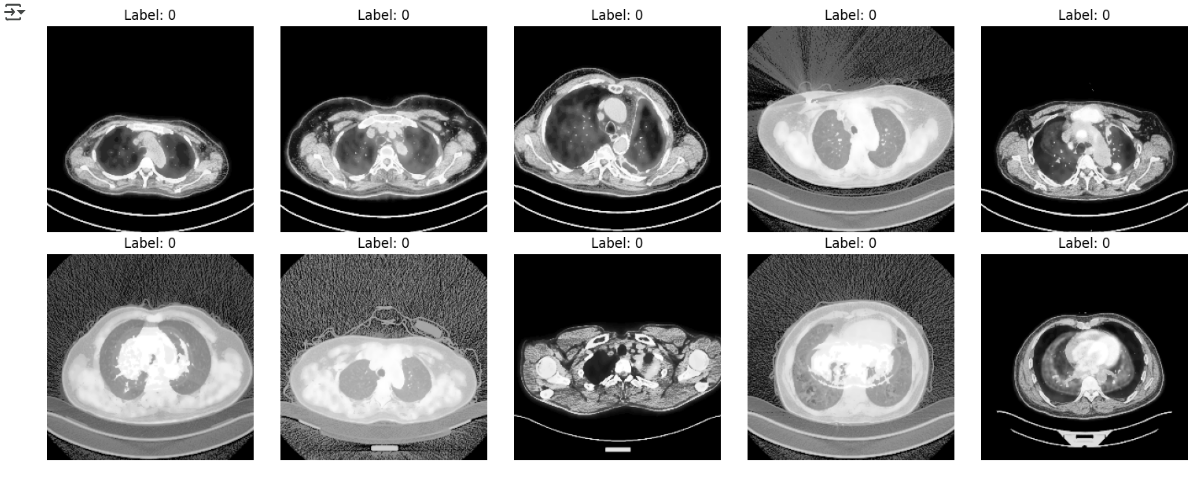


Figure 8 CT, PET scans of Lung Cancer patients

1. **Data Preprocessing**

Various data preprocessing techniques were utilized to prepare the images for model training and evaluation. These included the following

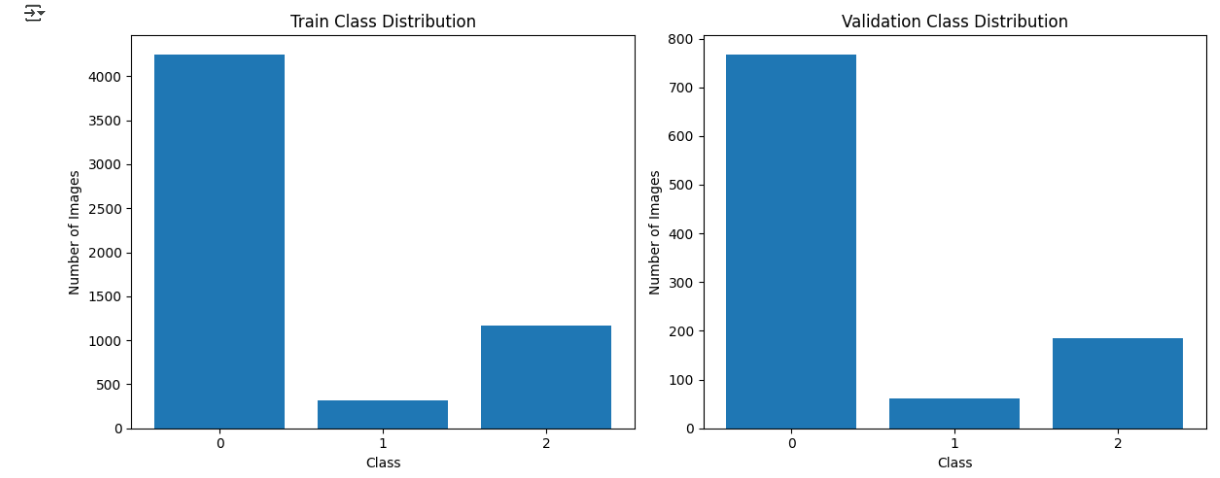
1. **Image resizing**; the original CT scans had a high resolution of **1024 × 1024**, which could slow down loading and training. To optimize performance, the images were resized to **224 × 224**, significantly accelerating the training process while retaining essential diagnostic details
2. **Histogram equalization**; this helped to enhance the contrast of the CT scans. This improved the visibility of the images by spreading out the most intensity values
3. **Label binarization**; the targets were initially categorical. We applied label binarization to transform them into a binary format suitable for model training; 0 – for adenocarcinoma, 1- for small cell carcinoma, and 2 – for squamous cell carcinoma
4. **Checking for class balance**; we assessed the distribution of targets. Imbalanced dataset creates model biases in favor of majority class and poor performance of minority classes

Figure 9 Class distribution of Imagery dataset

From the above visualization the dataset was imbalanced which would lead to poor model performance on the minority class. This imbalance was solved during the data argumentation stage by increasing the number of samples of the underrepresented classes that is; small cell carcinoma, and squamous cell carcinoma

1. Data augmentation; we applied data argumentation especially focusing on generating more samples of the underrepresented classes. Argumentation techniques applied include; horizontal rotation, flipping images, zooming in, and contrast adjustment
2. Normalization; the image pixels were normalized by division by 255, the maximum pixel value. This is helped to speed up model training and model stability
3. **Model Architecture Selection**

The architecture leveraged utilizes the power of **transfer learning** with fine-tuning, combining deep features learned from ImageNet with custom classification layers tailored for a **3-class lung cancer variant classification**. The following provides a summary of the model structure

|  |  |
| --- | --- |
| Component | Details |
| Base Model | ResNet50 (pretrained on ImageNet, include\_top=False) |
| Input Shape | (224, 224, 3) |
| Top Layers Added |  |
| Global Average Pooling | GlobalAveragePooling2D() |
| Dense Layer 1 | Dense(1024, activation='relu') |
| Dropout Layer 1 | Dropout(0.5) |
| Dense Layer 2 | Dense(512, activation='relu') |
| Dropout Layer 2 | Dropout(0.3) |
| Output Layer | Dense(3, activation='softmax') – for 3-class classification |
| Final Model | Model(inputs=base\_model.input, outputs=predictions) |
| Trainable Layers | Last 30 layers of ResNet50 set to trainable=True (fine-tuning enabled) |
| Frozen Layers | All other layers of ResNet50 frozen (trainable=False by default) |

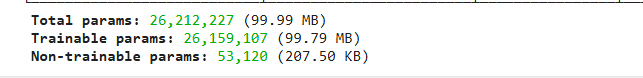
****Table 11 Model Summary of Image Classification model

Figure 10 Total number of model parameters

**In this study, we used the following training parameters**

|  |  |
| --- | --- |
| Parameter | Value / Description |
| Optimizer | Adam with learning rate = **0.001** |
| Loss Function | categorical\_crossentropy – suitable for multi-class classification |
| Metric | accuracy |
| Epochs | 50 (maximum) |
| Callbacks | EarlyStopping (monitors val\_loss, patience = **5**, restores best weights) |
| Training Data | train\_generator |
| Validation Data | validation\_generator |

Figure 11 Training parameters of CT classification image model

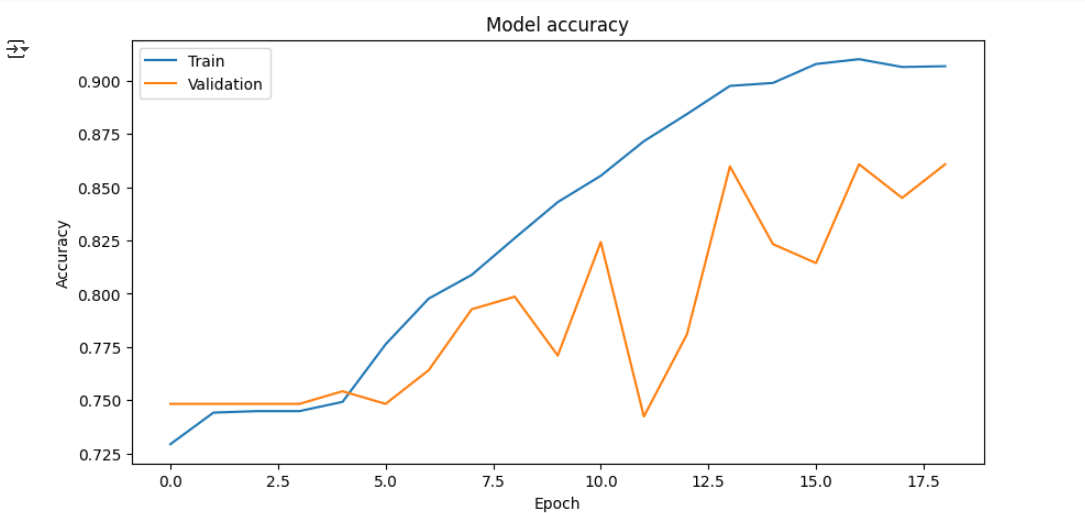
**The following illustrates the training and validation accuracy of the model**

Figure 12 Training and validation accuracy of the Cancer Variant Classification model

**Model Consistency testing**

**This was to be achieved using Cohen’s Kappa (κ). It is a statistical measure used to evaluate the degree of agreement between two models, beyond what would be expected by random chance. It considers both the observed agreement and the expected agreement. A higher κ value indicates stronger consistency between the models. The score is interpreted as follows: values less than 0.00 indicate poor agreement; 0.01–0.20, slight agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, substantial agreement; and 0.81–1.00, almost perfect agreement.**

**Where ​ is the observed agreement (i.e., the proportion of instances where both models agree, calculated as the sum of true positives and true negatives divided by the total number of cases), and ​ is the expected agreement by chance, based on the distribution of each model’s predictions.**

## ****3.4 SPECIFIC OBJECTIVE FOUR: TO DEVELOP A MOBILE APPLICATION INTERFACE AND INTEGRATE IT WITH MACHINE LEARNING MODEL****

To enable user-friendly interaction with the lung cancer risk assessment and image analysis models we developed mobile application in parallel with a RESTful APIs to ensure smooth and efficient model integration. The following steps were carried involved.

#### **3.4.1 API Development for Model Interaction**

This objective encompassed the development of four (4) key APIs to enable comprehensive user interaction: user authentication, cancer risk assessment powered by the risk-factor-based model, CT scan classification using the image analysis model, and user history management to support end-to-end engagement and tracking. This involved the following activities;

**Installation of Tools and Libraries**

|  |  |
| --- | --- |
| Tool / Library | Purpose |
| NestJS | Framework for building scalable and maintainable server-side applications |
| Node.js | JavaScript runtime environment required to run NestJS apps |
| Typescript | Primary language for NestJS, offering type safety and advanced features |
| Express (default) | Underlying HTTP server framework (NestJS wraps around it by default) |
| @nestjs/cli | Command-line tool to scaffold, build, and manage NestJS projects |
| @nestjs/core | Core NestJS package required for application structure and lifecycle |
| @nestjs/common | Contains decorators, services, and utility classes used across the app |
| @nestjs/platform-express | Enables Express platform integration with NestJS |
| @nestjs/swagger | For auto-generating interactive API documentation using Swagger |
| @nestjs/config | Allows environment-based configuration management |
| class-validator | Used for validating incoming request data in DTOs |
| class-transformer | Helps transform plain objects into class instances and vice versa |
| axios (optional) | For making HTTP requests to external services if needed |
| TypeORM / | (Choose one) ORM or ODM for database interaction |
| Jest | Testing framework integrated by default in NestJS for unit and e2e testing |

Table 12 Tools and libraries for API development

**Database Structure and Entity Relation Diagram**

**The database model mainly comprises three entities;**

* **User; for managing user authentication and credentials**
* **Diagnose; for managing risk assessment API calls and historical assessments**
* **CT Diagnose; for managing CT scan classification API calls and historical CT scan diagnoses**

**The following illustrates the Entity Relation Diagram which details; attributes of each entity with their respective data-types, plus associations or relationships with other entities**

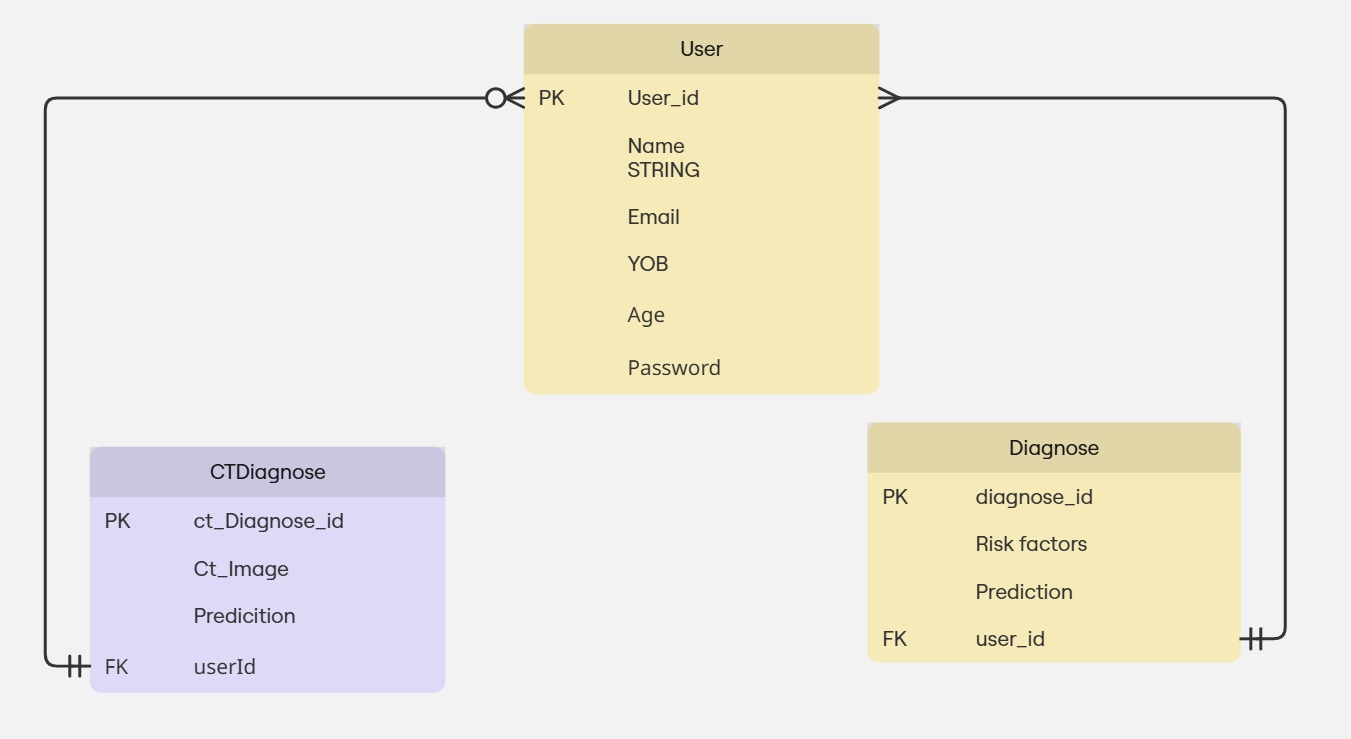
****

Figure 13 Entity Relation Diagram of the Backend service

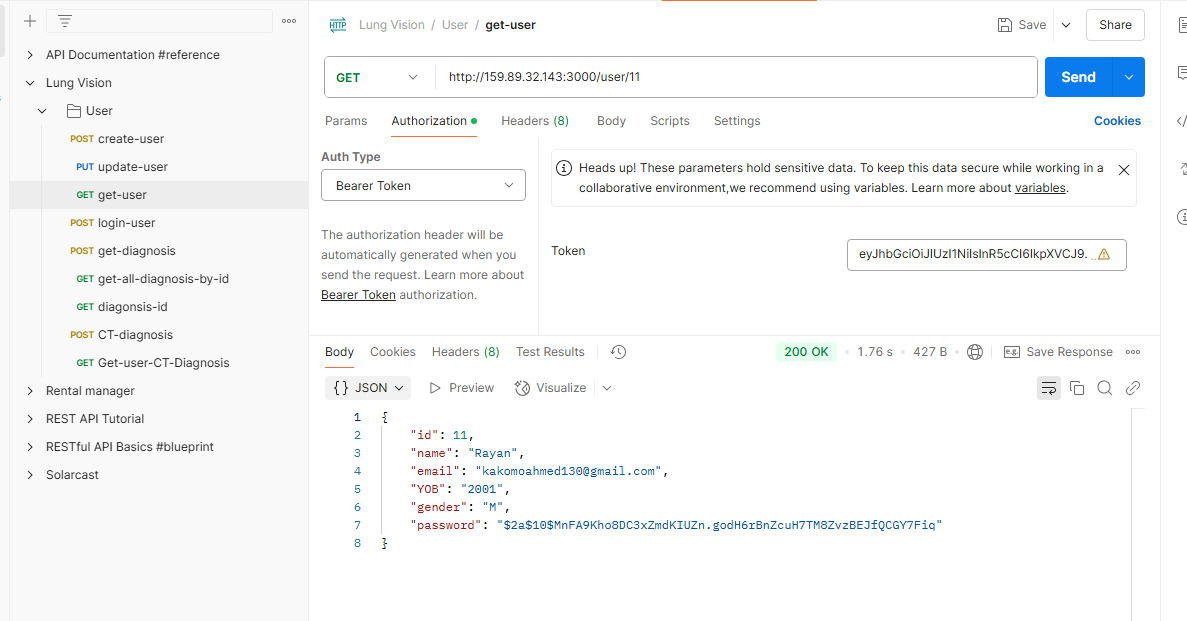
All the Backend service codes are hosted on the project repository. [Visit here to preview the source codes of the backend service.](https://github.com/wilfredKisitu/Lung_vision_backend_application)

Figure 14 Postman showing the API endpoints developed

**Deployment of Backend service**

Logging into the remote server using the IP address.

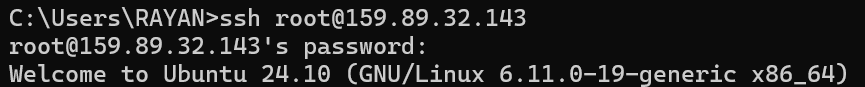


Figure 15: logging into the remote server

A snap shot of the successfully logged into server.

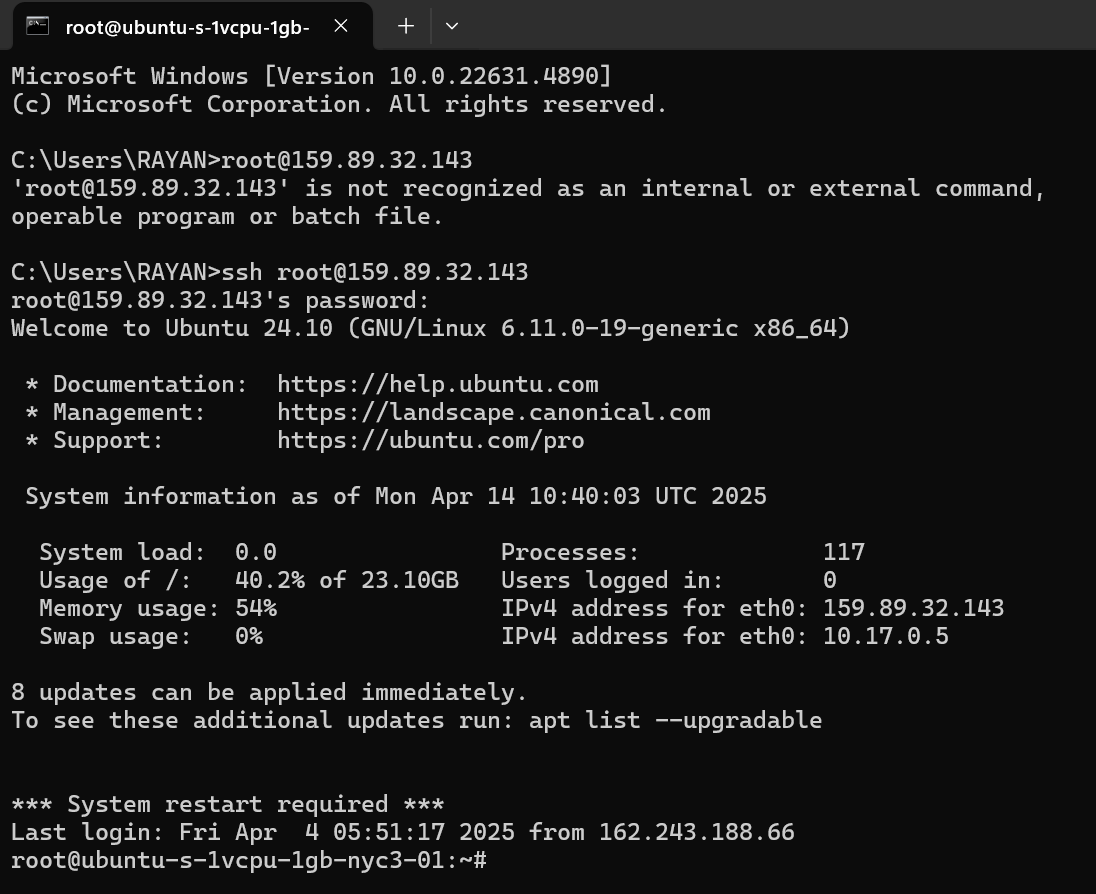


Figure 16: successfully logged into server.

A snap shot of the operating system information of the accessed remote server.

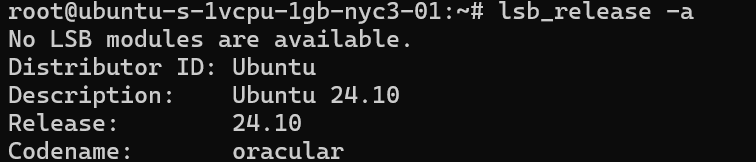


Figure 17: system information of the remote server.

A snap shot of the kernel information of the accessed remote server.



Figure 18: kernel information of the remote server.

A snap shot of the RAM information of the accessed remote server.

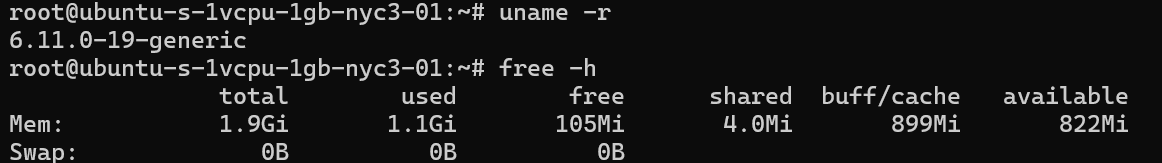


Figure 19: RAM information of the remote server.

A snap shot of the disk information of the accessed remote server.

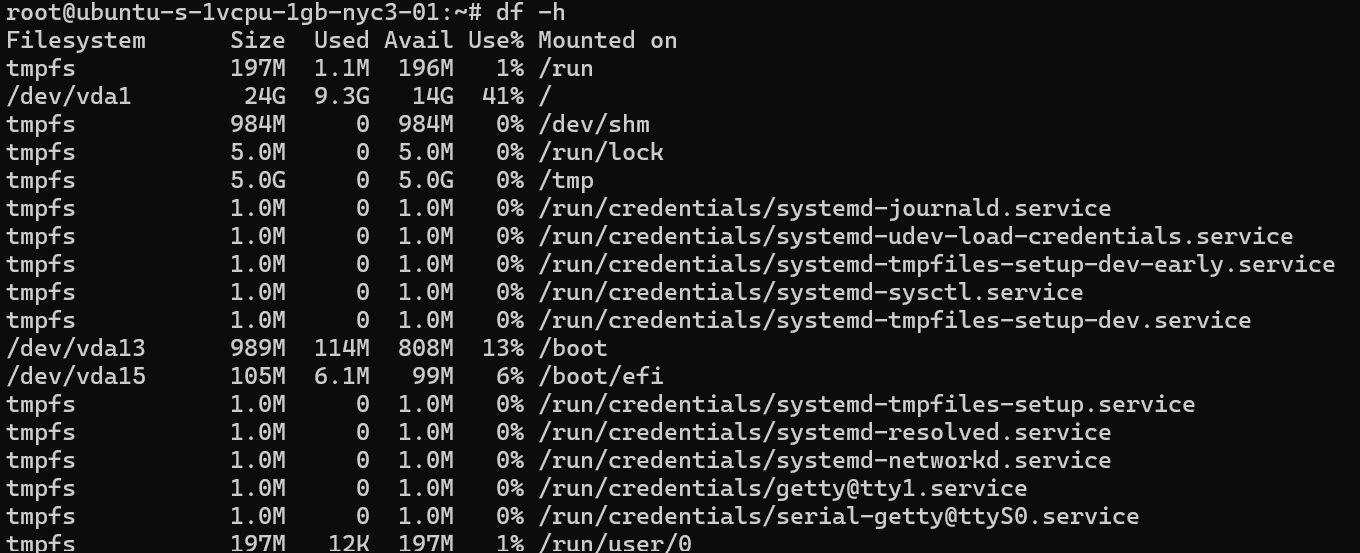


Figure 20: disk information of the remote server.

### ****3.4.2 Mobile User Interface Development****

The mobile interface was developed using **Flutter** for its cross-platform capabilities, allowing a single codebase to run seamlessly on both Android and iOS devices. Flutter’s rich widget library enabled the creation of a responsive and intuitive user interface, while its fast development cycle ensured efficient prototyping and testing. Integration with the backend API was achieved using HTTP packages, allowing real-time communication with the diagnosis models and secure access to user data and prediction results. This objective also involved multiple activities which include the following;

**Wireframe prototypes**

**Using Figma, a rapid prototyping tool, we developed wire frames and prototypes of UI and UX of the application with focus on smooth and user friendly interaction of the end user**

**Installation of Tools and Libraries.**

|  |  |
| --- | --- |
| Package / Tool | Purpose |
| Flutter SDK | Core framework for building cross-platform mobile applications |
| Dart | Programming language used with Flutter |
| Android Studio / VS Code | IDEs commonly used for Flutter development |
| flutter\_secure\_storage | Securely stores sensitive data like tokens and passwords |
| http | Makes HTTP requests to RESTful APIs |
| provider | State management solution for managing app-wide state |
| shared\_preferences | Stores simple key-value data locally (e.g., user settings, login status) |
| image\_picker | Allows users to pick images from the gallery or camera |
| firebase\_core | Initializes Firebase services (if using Firebase backend) |
| firebase\_auth | Handles user authentication via Firebase |
| cloud\_firestore | Cloud-based NoSQL database from Firebase |
| fluttertoast | Displays short messages in toast notifications |
| path\_provider | Provides access to commonly used device file paths |
| google\_fonts | Easily use custom Google Fonts in your app |
| flutter\_form\_builder | Simplifies form creation and validation |
| intl | Supports localization and date formatting |
| flutter\_launcher\_icons | Customizes app icons for different platforms |
| flutter\_native\_splash | Generates native splash screens |

Table 18 Packages for frontend development using flutter

**Debugging and Optimization**

Extensive debugging was carried out to resolve UI inconsistencies and eliminate runtime errors, followed by performance optimization to enhance the app’s speed, responsiveness, and overall user experience

* + 1. **API Integration with Mobile Interface**

Integrated the mobile application with the API endpoints using HTTP for seamless communication and Riverpod for state management. We implemented features that enables to;

* **Send user data:** Risk factors were sent to the API for model prediction.
* **Retrieve predictions:** The application received and displayed the diagnostic results.
* **Authentication Management:** Integrated login and registration features using the API’s authentication endpoint.
* **Error Handling:** Incorporated user-friendly error messages for invalid inputs and server issues.

### ****3.4.4 Unit and Integration Testing****

**Unit Testing**

Unit testing was performed on individual components to ensure reliability and accuracy. This included evaluating API responses across various input scenarios, verifying the validation logic within both the mobile interface and backend API, and testing UI elements to confirm they responded correctly and intuitively to user interactions

**Integration Testing.**

We tested the entire workflow, including; submitting data from the mobile app and receiving accurate prediction, ensuring authentication layers worked correctly for secure API access, verifying that error handling mechanisms triggered appropriately.

# CHAPTER FOUR: DISCUSSIONS OF RESULTS AND TESTING

## ****4.1**** DISCUSSION OF RESULTS.

### 4.1.1 TO IDENTIFY AN EXISTING DATASET THAT CAPTURES THE LIFESTYLE, DEMOGRAPHICS, MEDICAL HISTORY, AND CLINICAL SYMPTOMS OF LUNG CANCER PATIENTS

### Most Suitable Data Selection

The “[Lung Cancer Prediction](https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link)” dataset from Kaggle was selected for analysis. It comprises 1,000 patient records, each containing 21 attributes related to lung cancer risk factors. These features include age, gender, and various health and lifestyle indicators such as air pollution exposure, alcohol consumption, dust allergy, occupational hazards, genetic predisposition, chronic lung disease, diet quality, obesity, smoking habits, passive smoking, and symptoms like chest pain, coughing up blood, fatigue, weight loss, shortness of breath, wheezing, difficulty swallowing, clubbing of fingernails, and snoring. Analyzing this dataset allowed us to uncover patterns and gain valuable insights into the risk factors and early indicators associated with lung cancer. The dataset was mainly tailored for;

* Predicting the likelihood of a patient developing lung cancer
* Identifying risk factors for lung cancer
* Determining the most effective treatment for a patient with lung cancer

This dataset therefore aligned with the predefined dataset characteristics for the study that is; it comprised the four (4) categories of risk factors relevant for study namely; **lifestyle factors** such as smoking history and alcohol consumption, **demographics** such as age, and gender of diagnosed patients, **medical history** including chronic diseases, allergies, and **clinical symptoms** such as wheezing, coughing, shortness of breath, and chest pain

### ****Dataset Description and structure.****

Upon selection, this dataset was examined to understand its structure and contents. The key steps involved:

1. **Data Exploration.** A summary of dataset attributes, including feature distribution, and nature of the attribute (either numeric or categorical). The following table provides a description of each attribute in the dataset

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Description** | **Attribute type** |
| Age | The age of the patient. | Numeric |
| Gender | The gender of the patient. | Categorical |
| Air Pollution | The level of air pollution exposure of the patient. | Categorical |
| Alcohol use | The level of alcohol use of the patient. | Categorical |
| Dust Allergy | The level of dust allergy of the patient. | Categorical |
| Occupational Hazards | The level of occupational hazards of the patient. | Categorical |
| Genetic Risk | The level of genetic risk of the patient. | Categorical |
| chronic Lung Disease | The level of chronic lung disease of the patient. | Categorical |
| Balanced Diet | The level of balanced diet of the patient. | Categorical |
| Obesity | The level of obesity of the patient. | Categorical |
| Smoking | The level of smoking of the patient. | Categorical |
| Passive Smoker | The level of passive smoker of the patient. | Categorical |
| Chest Pain | The level of chest pain of the patient. | Categorical |
| Coughing of Blood | The level of coughing of blood of the patient. | Categorical |
| Fatigue | The level of fatigue of the patient. | Categorical |
| Weight Loss | The level of weight loss of the patient. | Categorical |
| Shortness of Breath | The level of shortness of breath of the patient. | Categorical |
| Wheezing | The level of wheezing of the patient. | Categorical |
| Swallowing Difficulty | The level of swallowing difficulty of the patient. | Categorical |
| Clubbing of Finger Nails | The level of clubbing of finger nails of the patient. | Categorical |

Table 2 Description of attributes

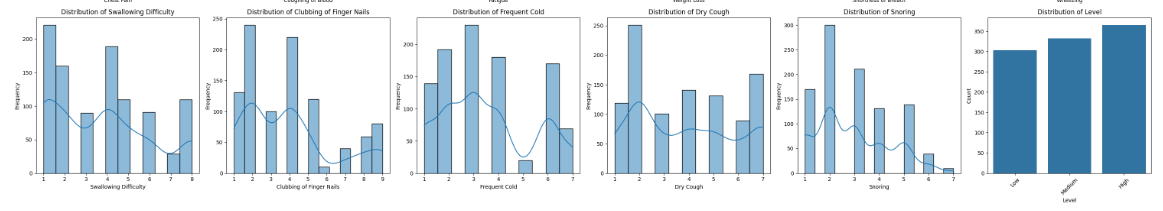
1. **Feature Categorization.** The attributes were categorized into four groups: lifestyle, demographics, medical history, and clinical symptoms. The categorization is illustrated below

|  |  |
| --- | --- |
| **Category** | **Features** |
| Demographics | Age, Gender |
| Lifestyle | Air pollution exposure, Alcohol consumption, Occupational hazards, Diet quality, Obesity, Smoking habits, Passive smoking |
| Medical History | Dust allergy, Genetic predisposition, Chronic lung disease |
| Clinical Symptoms | Chest pain, Coughing up blood, Fatigue, Weight loss, Shortness of breath, Wheezing, Difficulty swallowing, Clubbing of fingernails, Snoring |

Table 3 Category of risk factors

By following these steps, the study successfully identified and prepared a dataset that met the requirements for lung cancer diagnosis using risk factors.

**Attribute Distribution**

The following illustrates the distribution and scales used for the collection of each attribute of the dataset used the study

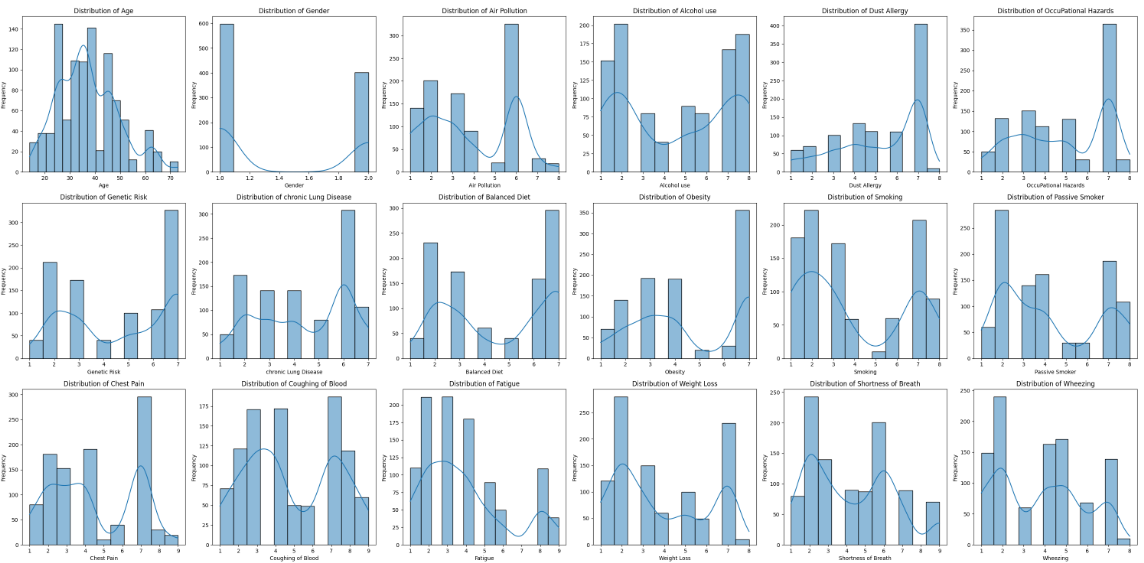


Figure 1 Distribution of Attributes

### 4.1.2 SPECIFIC OBJECTIVE TWO: TO EXTRACT FEATURES FOR EARLY LUNG CANCER DIAGNOSIS

**Selection of Most Informative Features;**

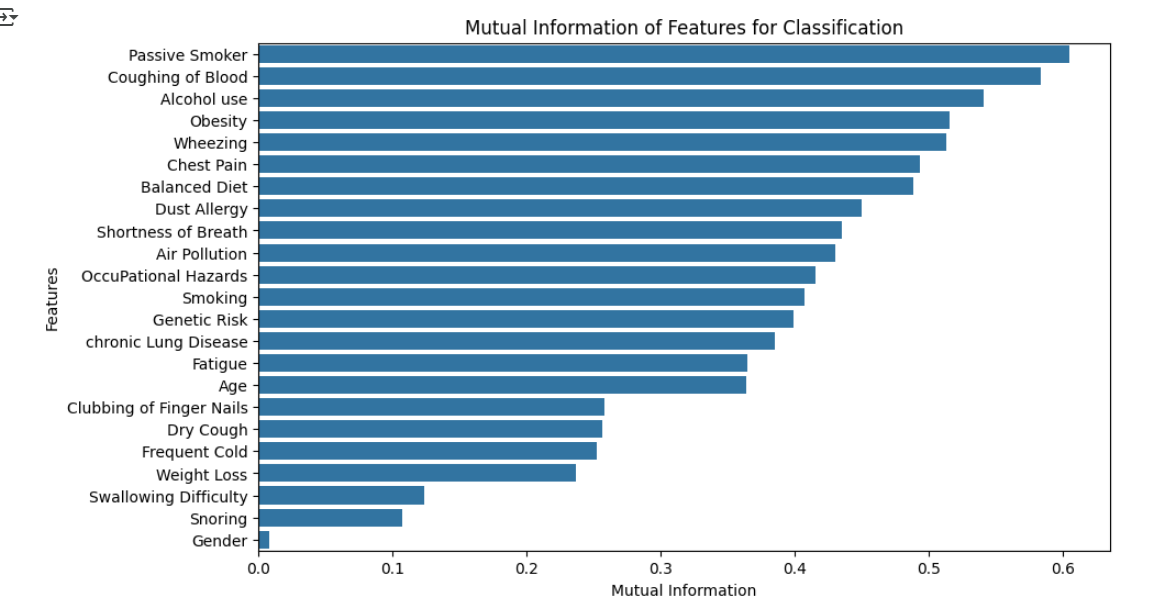
Features with the highest MI scores (taking a threshold of 10%) were prioritized as key risk factors, while those with low scores were considered for removal to improve model efficiency. From the illustration below, most attributes had significant information with respect to the target except the “Gender” who’s MI score didn’t meet the threshold. The following illustration shows the MI scores in descending order.

Figure 4 Mutual information scores

The correlation matrix analysis showed that **smoking history, age, chronic disease history, coughing, shortness of breath, and chest pain** had strong positive correlations with lung cancer diagnosis. These features were the most statistically significant and clinically relevant, confirming their importance for accurate prediction. Smoking history showed the highest correlation, supporting its role as a key risk factor.

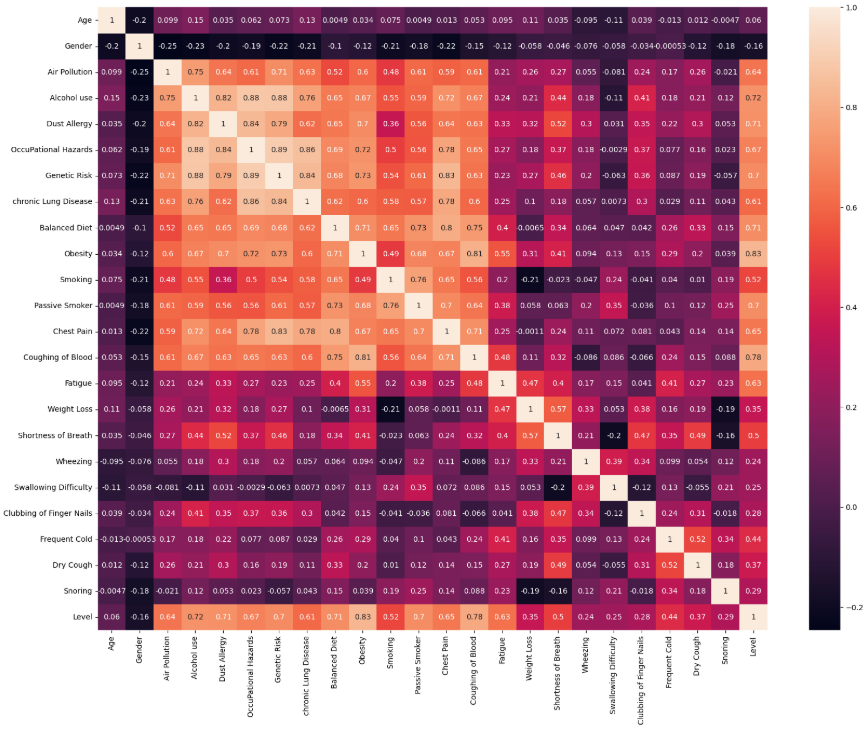


Figure 5 Correlation Matrix

### 4.3 OBJECTIVE THREE: TO DEVELOP A MACHINE LEARNING MODEL AND EVALUATE ITS PERFORMANCE BASED ON; ACCURACY, SENSITIVITY, AND FALSE POSITIVE THROUGH K-FOLD CROSS VALIDATION

**Model Performance Reports**

In the study, five (5) models were trained on the structured dataset:

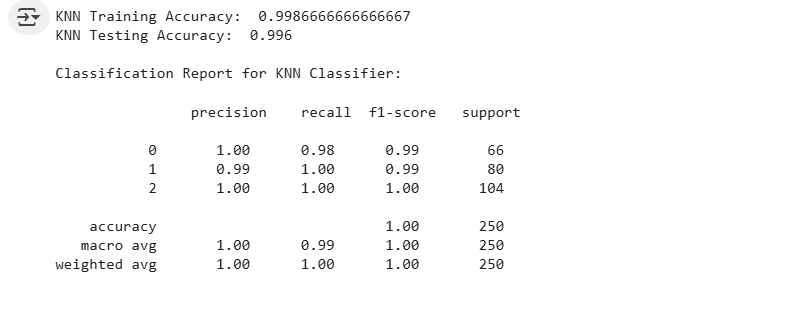
1. ****K-Nearest Neighbors (KNN) **achieved a 99.6% testing accuracy based on k-fold cross validation on the dataset, which demonstrated** remarkable performance in pattern recognition, although slightly less robust in handling complex, nonlinear relationships. **The following illustrates other evaluation metrics tested**

Figure 23 Performance report of KNN Classifier

1. Decision Tree; **the model achieved a 98.4% testing accuracy based on k-fold cross validation on the dataset. The following illustrates other evaluation metrics tested**

****

Figure 24 Performance of the Decision tree classifier

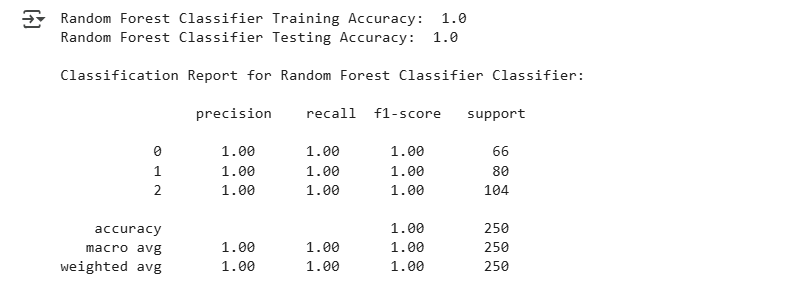
1. ****Random Forest classifiers: **The model achieved a 100% testing accuracy based on k-fold cross validation on the dataset** respectively, reflecting the strength of ensemble learning in medical datasets**. The following illustrates other evaluation metrics tested**

Figure 25 Performance report of Random Forest classifier

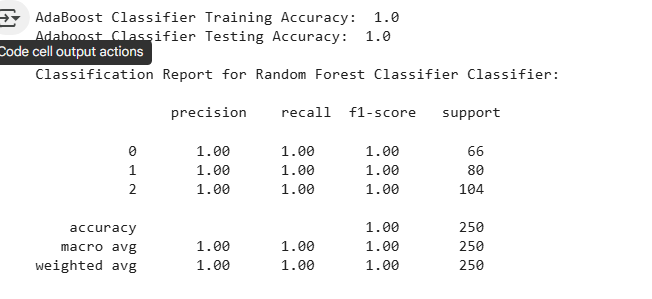
1. ****AdaBoost Classifier: **The model achieved a 100% testing accuracy based on k-fold cross validation on the dataset** benefiting from iterative refinement of weak learners. **The following illustrates other evaluation metrics tested**

Figure 26 Performance report of AdaBoost Classifier

**To better visualize and evaluate the performance of the above traditional machine learning models, the confusion matrix for each model was computed and visualized on a heat map. It shows the True positives, False Positives, True negatives, and false negatives of each model. Below is an illustration of the confusion matrices**

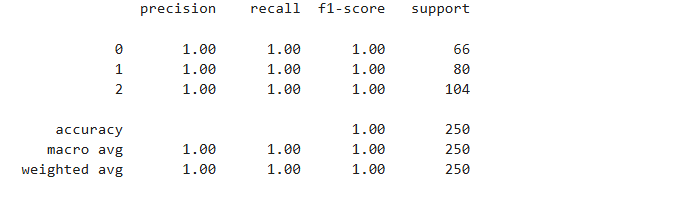
1. ****The Artificial Neural Network (ANN): **The ANN model also achieved a 100% testing accuracy based on k-fold cross validation on the dataset** indicating strong capability to model intricate dependencies between risk variables. **The following illustrates other evaluation metrics tested**

Figure 27 Confusion matrix of the four machine learning models

Figure 28 Performance report of ANN Classifier

**The following table provides a summary of all the models’ performance based on the above metrics**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Targets** | **Precision** | **Recall** | **F1-score** | **Accuracy** |
| KNN Classifier | 0 | 1.00 | 0.98 | 0.99 |  |
| 1 | 0.9 | 1.00 | 0.99 | 99.6% |
| 2 | 1.00 | 1.00 | 1.00 |  |
| Decision Tree Classifier | 0 | 1.00 | 0.94 | 0.97 |  |
| 1 | 0.95 | 1.00 | 0.98 | 98.4% |
| 2 | 1.00 | 1.00 | 1.00 |  |
| Random Forest Classifier | 0 | 1.00 | 1.00 | 1.00 |  |
| 1 | 1.00 | 1.00 | 1.00 | 100% |
| 2 | 1.00 | 1.00 | 1.00 |  |
| AdaBoost Classifier | 0 | 1.00 | 1.00 | 1.00 |  |
| 1 | 1.00 | 1.00 | 1.00 | 100% |
| 2 | 1.00 | 1.00 | 1.00 |  |
| ANN Classifier | 0 | 1.00 | 1.00 | 1.00 |  |
| 1 | 1.00 | 1.00 | 1.00 | 100% |
| 2 | 1.00 | 1.00 | 1.00 |  |

Table 10 Comparison of model Performance

These high accuracies indicate the dataset’s high productiveness and potential simplicity in classifying risk once appropriate features are selected.

**Model Selection**

**From the above study, both traditional and deep learning models presented remarkable performance on the dataset with Random Forest, AdaBoost and Artificial Neural Networks presenting 100% accuracy on test dataset based on 10 k-fold cross validation. Given the fact that Artificial Neural Networks have the ability to capture complex data patterns even at high dimensionality, we selected the Artificial Neural Network as the best performing and a better choice for our mobile application integration**

**Image Analysis**

To test for consistency between risk factor based models and imaging model, a separate CT-scan classifier was developed, aiming to classify CT-scans into one of three types: adenocarcinoma, squamous cell carcinoma, or small cell carcinoma**. The model managed to achieve a 90.2% training accuracy, an 86.08% validation accuracy. Due to the large size of the dataset with respect to the available computation resources, we trained the baseline model on half of total dataset size 6747 CT images achieving 79% accuracy. The model with its previous learnt weights was trained on the remaining sample of the dataset achieving a high validation accuracy of 86.08%.** While lower than risk factor models, this accuracy is substantial given the complexity of medical image interpretation. The performance was attributed to several factors:

* Visual similarity between different cancer types.
* Variability in CT image quality and acquisition settings.
* The need for more balanced datasets across cancer types.

This component introduced an important dimension to the system by integrating imaging diagnostics, which are often central to cancer confirmation and staging in real-world practice. This therefore provides a framework, where patients can assess their risks and at the same time identify the type of lung cancer in case they are malignant

**Model Consistency testing**

The aim of this objective was to test the consistence of the risk assessment model with the CT image classification model. This therefore required access to a diverse dataset that captures the risk factors of a patients as well as their respective CT scans respectively. However, depending on our research, we didn’t manage to identify such a dataset and recommend any further research in this study to build from there.

### 4.1.4 OBJECTIVE FOUR: TO DEVELOP A MOBILE APPLICATION INTERFACE AND INTEGRATE IT WITH MACHINE LEARNING MODEL

**API DEVELOPMENT**

We successfully developed four (4) APIs to facilitate seamless end-user interaction. These APIs support key functionalities, including user authentication, risk factor–based diagnosis, CT image–based diagnosis, and historical medical record tracking.

**Documentation for the APIs of the Backend Service**

1. Authentication

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Service | Method | Route (Server based URL) | Request Components | | | Response payload | Possible status codes |
| Header | Query | Body |
| Login | POST | /auth/login | None | None | Email, password | JWT Token | 201, 400, 401, 409, 500 |

Table 13 Authentication API documentation

1. User

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Service | Method | Route (Server based URL) | Request Components | | | Response payload | Possible status codes |
| Header | Query | Body |
| Create User | POST | /user | None | None | Name, email, age, gender, YOB, password | User | 201, 400, 401, 409, 500 |
| Get User | GET | /user/:id | JWT Token | None | None | User | 200, 400, 401, 409, 500 |
| Update User | PUT | /user/:id | JWT Token | None | Name, email, password, age, gender, YOB | User | 200, 400, 401, 409, 500 |
| Delete User | DEL | /user/:id | JWT Token | None | None | None | 200, 400, 401, 409, 500 |

Table 14 User API documentation

1. Diagnose

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Service | Method | Route (Server based URL) | Request Components | | | Response payload | Possible status codes |
| Header | Query | Body |
| Diagnose | POST | /diagnose | JWT Token | None | Risk factors | Diagnosis | 201, 400, 401, 409, 500 |
| Get user Diagnosis | GET | /diagnose/user/:id | JWT Token | None | None | Diagnosis | 200, 400, 401, 409, 500 |
| Get Diagnosis | GET | /diagnose/userDiagnosis/:diagnosed | JWT Token | None | None | Diagnosis | 200, 400, 401, 409, 500 |

Table 15 Diagnose API Documentation

1. CT Diagnose

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Service | Method | Route (Server based URL) | Request Components | | | Response payload | Possible status codes |
| Header | Query | Body |
| Ct Diagnose | POST | /ctDiagnose/upload/:userId | JWT Token | None | Image file | CT Diagnosis | 201, 400, 401, 409, 500 |
| Get CT Diagnosis | GET | /ctDiagnose/user/:userId | JWT Token | None | None | CT Diagnoses | 200, 400, 401, 409, 500 |

Table 16 Ct Diagnose API Documentation

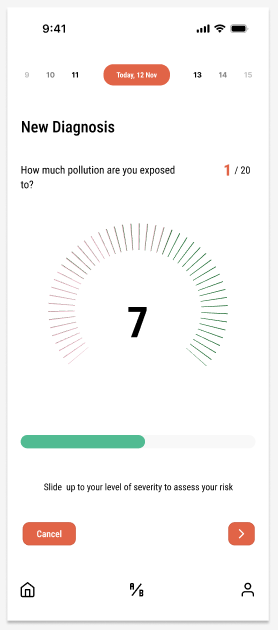
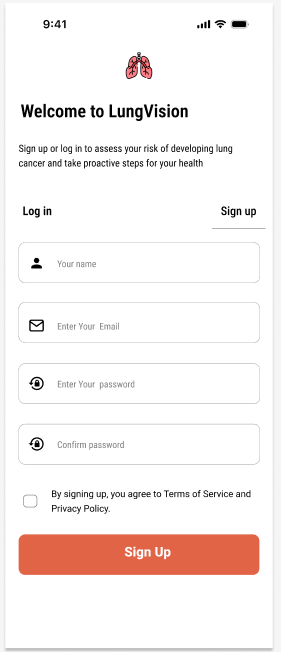
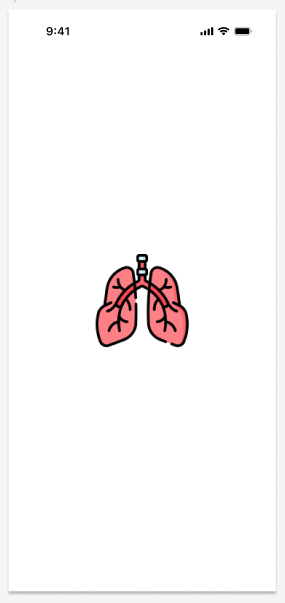
|  |  |
| --- | --- |
| STATUS CODES | |
| 200 | OK |
| 201 | Created |
| 204 | No Content |
| 400 | Bad Request |
| 401 | Unauthorized |
| 404 | Not found |
| 409 | Conflict |
| 500 | Internal Server Error |

Table 17 Status codes

**MOBILE APPLICATION DEVELOPMENT**

**Interface Implementation**

The mobile application interface was designed with key components to ensure a smooth and functional user experience. Input forms were developed to collect essential risk factor data such as age, smoking history, and symptoms. Result display screens were integrated to present the lung cancer diagnosis outcomes clearly to the user. Additionally, intuitive navigation elements were implemented to enable seamless transitions between different sections of the application.



**All the above designs are maintained on the Figma project account.** [Visit here in case you need to preview the system design system](https://www.figma.com/design/vnaEsiffj8bUSugoI3HAHG/Lung-Vision?node-id=0-1&p=f&t=uDV3SHE1Put48XTM-0)

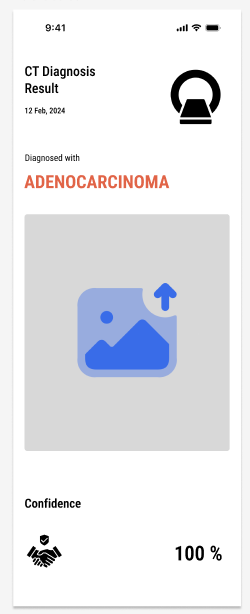
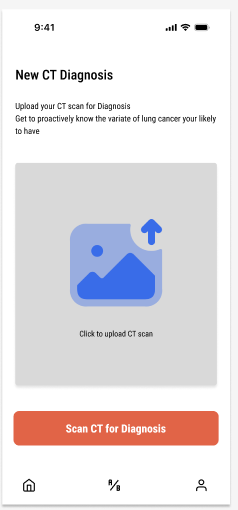
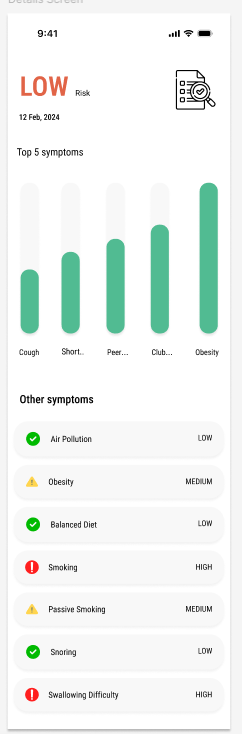


Figure 21 Wireframes for the frontend application

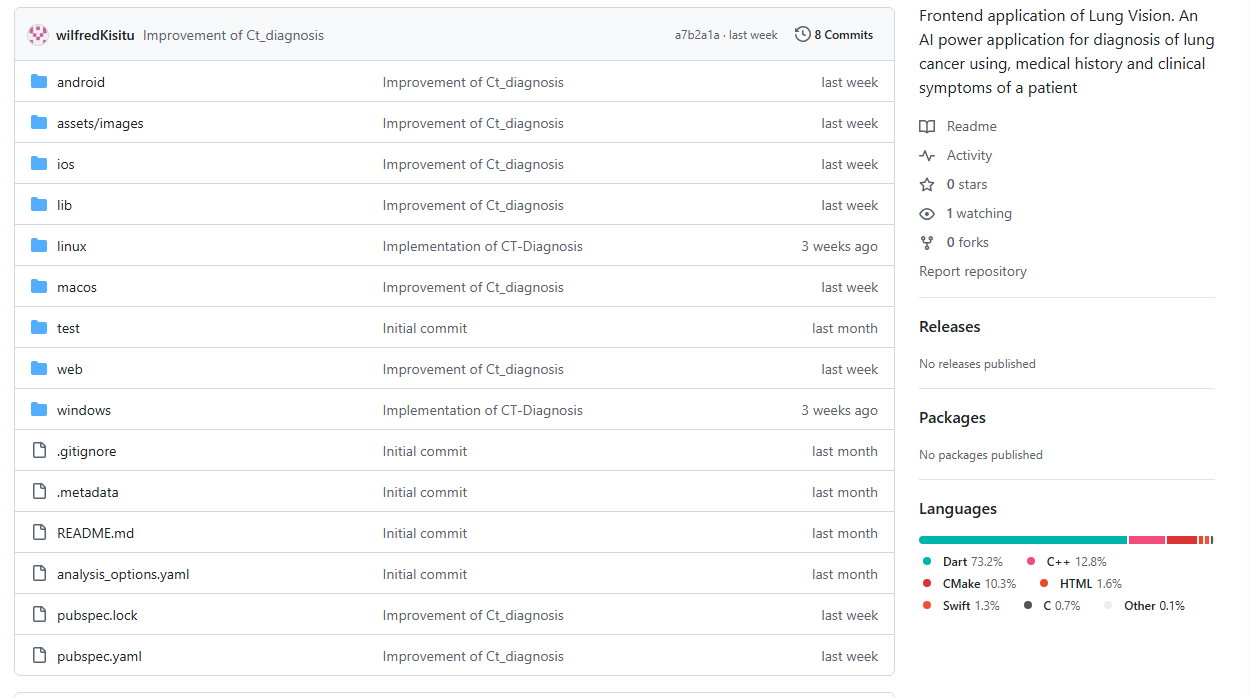
[Visit here to view the project source codes on GitHub](https://github.com/wilfredKisitu/Lung_vision_frontend)

Table 19 GitHub repository for project source codes

## ****4.2 TESTING****

**API Testing**

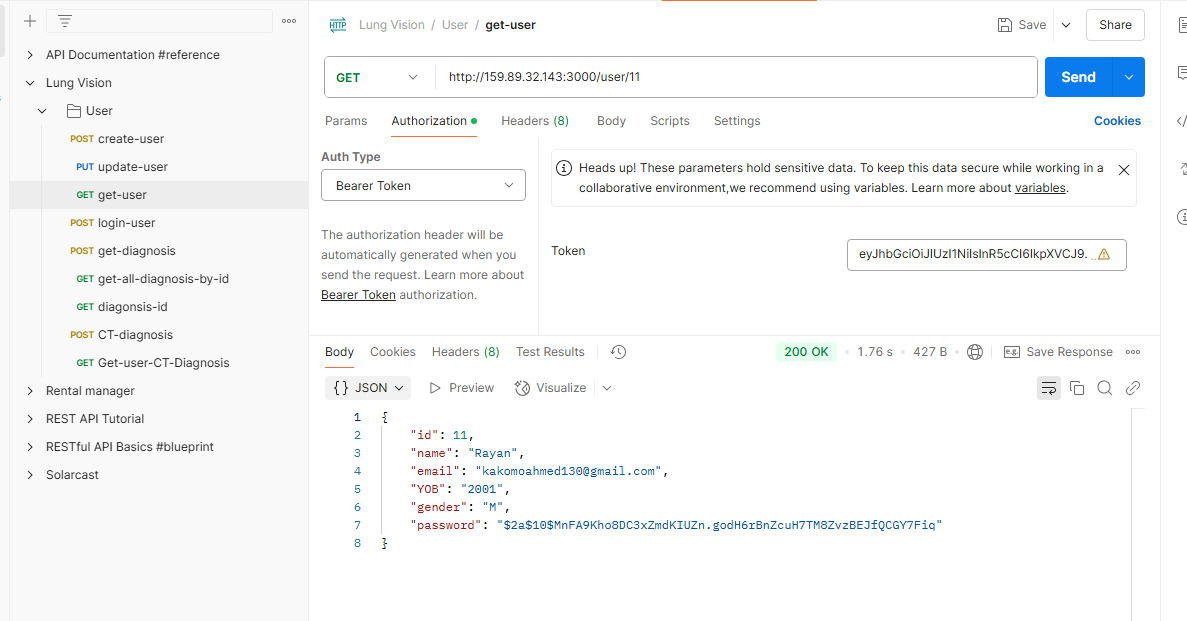
The APIs were thoroughly tested using tools like **HTTP client** and **Postman** to verify seamless communication between the mobile interface and the machine learning model, ensuring accurate data exchange and reliable system performance.

Figure 14 Postman showing the API endpoints developed

**Mobile Application Testing**

User interface components were tested to ensure smooth interaction. The following illustrates the tested features, expected behavior and their respective outcome

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Expected Behaviour** | **Outcome** | **Status** |
| Input form validation | Rejects incomplete/incorrect data | Works correctly | Passed |
| API connectivity | Fetches and sends data to API | Works correctly | Passed |
| Display of results | Shows predicted diagnosis | Works correctly | Passed |

**User Acceptance Testing (UAT)**

**Users** tested the mobile application to assess usability. They provided feedback on:

1. **Ease of Use.** Users found the app intuitive and easy to navigate.
2. **Performance.** The app responded quickly, with minimal delays in retrieving results.

**Key Findings**

* The API responded accurately to model prediction requests.
* The mobile app displayed results effectively and managed user authentication securely.
* Integration tests confirmed seamless interaction between the mobile interface and the API.

This structured approach ensured the successful development and integration of a mobile application with the lung cancer diagnosis model, providing users with an intuitive and reliable diagnostic tool.

# CHAPTER FIVE: RECOMMENDATIONS AND CONCLUSION

## 5.1 ****Challenges****

In the study we faced various challenges ranging from; data identification, methodology, system development and integration;

* In the data collection phase, we opted to identify a combined record of patient’s risk factors and CT-scan taken during the same clinical diagnosis. We didn’t manage to identify such a dataset and opted for two separate dataset “[Lung Cancer Prediction](https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link)” for risk factors and “[Lung PET CT Dx](https://www.kaggle.com/datasets/rangan2510/lung-pet-ct-dx)” for CT-scans. This implied we couldn’t perform a consistency test between the risk factor based models and the imaging model
* The “[Lung PET CT Dx](https://www.kaggle.com/datasets/rangan2510/lung-pet-ct-dx)” was imbalanced dataset with majority representation of “Adenocarcinoma” cases, and minority representation of the other cancer types; “small cell carcinoma” and “squamous cell carcinoma”. We had to improvise an approach of generating more synthetic images of the minority classes through image augmentation. Though this provided a work round to solve model biasing, better results or generic predictive weights can be obtained if a balanced dataset can be identified or collected in any future studies collateral to this study.
* Some Image processing procedures such as Image segmentation through k-means couldn’t be achieved because of the heavy computing resources required to work with a large CT-dataset of over **13,494 images**, including CT, PET, and fused scans with resolution of 1024 x 1024 pixels. The training process alone had to be batched into two and the second training was achieved by retraining the model using the weights learnt from first batch. This also accounts for the lower validation accuracy of 84% achieved by the Imaging model
* Local hosting of the models on the user’s device couldn’t be achieved. This implies the system can only be accessed online which is a limitation to communities in low resource settings. This is because of the model’s sizes with the minimum model occupying 303 Megabytes which wouldn’t support development of light weight application
* The Imaging diagnosis model required more computing resources on the remote server as compared to the risk factor based model. Faster performance could only be obtained if we purchased a GPU-based droplet with a capacity of at least 15GB of GPU memory. This therefore implies we paid a incurred a higher cost of $ 0.48 / hour of usage based on [Digital Ocean’s](https://www.digitalocean.com/?utm_campaign=search_emea_en_brand&utm_adgroup=digitalocean_misspellings&utm_keyword=digital-ocean&utm_matchtype=e&utm_adposition=&utm_creative=691040276618&utm_placement=&utm_device=c&utm_location=&utm_location=9069988&utm_term=digital-ocean&utm_source=google&utm_medium=cpc&gad_source=1&gbraid=0AAAAADw9jcsnjh6a_J2xL1dxEA36vmTK6&gclid=Cj0KCQjw2ZfABhDBARIsAHFTxGzGsQDo2LEZFLbDT21vyN6WcGXcbNOVAtXtJ0ihm8Z5MLoOPg08YgIaAk_DEALw_wcB) catalog
* Due to constrain of time, feedback such as expanding to multiple types of cancer, designing a feedback tool, and providing features for cancer patient management in case of malignancy couldn’t be implemented in this study because of the scope of the study, time and budget constrain for implementing such features
* Although, the framework, Flutter we used for mobile application design supports cross platform development. We needed access to a MacBook with X-code and an iPhone to compile an iOS version of the application which we couldn’t accomplish. This implies the available application is only available to Android Users and limited to iOS based users

## 5.2 Feedback

To gather both user and expert feedback on the system, this work was exhibited at the National Council Exhibitions held at **Busitema University on 27th of March, 2025** as well as at **Mbarara University** from **27th to 29th March 2025**. The following feedback was collected:

* 1. AI practitioners appreciated the precision of the risk factor based model and the availability of symptom based medical tool for easy early assessment as compared to traditional procedure of assessment that utilizes CT-scans, or biopsy
  2. A National Council Executive recommended expanding future studies to include other cancer variants such as **skin** and **breast cancer**, in order to broaden the system's applicability and enhance its impact across multiple cancer types.
  3. Interning medical students at **Mbarara University** suggested incorporating features for **cancer patient management**, particularly for cases where a diagnosis confirms malignancy. They emphasized that such additions could support clinical decision-making and improve the **survival rates of lung cancer patients** through timely and guided interventions.
  4. Medical Practitioner from Busitema University, Mbale Campus recommended a feedback tool where people can rate our framework of clinical diagnosis. This will help build trust in our system as well as support adoption of the system among masses

## ****5.3 Recommendations****

Throughout the course of this study, several challenges were encountered during data identification, model development, system integration, and deployment. Based on these experiences, the following recommendations are proposed to guide future work:

1. **Unified Dataset Collection**: During the data collection phase, efforts were made to identify a dataset that combined both patient risk factors and corresponding CT-scan images taken during the same clinical diagnosis. Unfortunately, such a dataset was not available, leading to the use of two separate datasets—"Lung Cancer Prediction" for structured risk factors and "Lung PET CT Dx" for CT images. This limitation hindered our ability to directly test for consistency between the risk factor-based models and the imaging model. Future studies should aim to collect or identify multi-modal datasets that combine clinical and imaging data for the same patients to enable cross-validation and consistency testing.
2. **Addressing Data Imbalance:** The imaging dataset was highly imbalanced, with a significant overrepresentation of adenocarcinoma cases compared to small cell and squamous cell carcinoma. Image augmentation techniques were used to generate synthetic samples for underrepresented classes, partially addressing the imbalance. However, we recommend the use of balanced datasets in future studies to improve generalization and reduce bias in model predictions.
3. **Computational Resource Constraints**: Image processing techniques such as K-means clustering were not feasible due to the computational demands of working with over 13,494 high-resolution images (1024 x 1024 pixels). Additionally, the imaging model had to be trained in two batches due to hardware limitations, which may have affected its final accuracy (84%). We recommend that future projects allocate access to higher-performance computing infrastructure, such as GPUs with ≥15GB memory, to allow for more efficient training and advanced image preprocessing.
4. **Offline Access and Lightweight Models**: Due to the large model sizes (the smallest being 303 MB), local hosting on user devices was not achievable. Consequently, the application is only usable with an internet connection—posing a challenge for low-resource or rural settings. Future work should explore model compression techniques or lightweight neural architectures like MobileNet or EfficientNet to enable offline diagnosis and better accessibility.
5. **Server Cost Considerations**: Deploying the imaging model on the cloud incurred higher operational costs due to the need for GPU-powered virtual machines. For instance, GPU droplets on Digital Ocean with sufficient capacity cost $0.48 per hour. Future implementations should consider cost-efficient hosting platforms or on-device inference for sustainable deployment, especially for NGOs or public health use.
6. **Feature Expansion and Feedback Integration**: Feedback collected from users and professionals recommended expanding the tool to include support for other cancer types (e.g., skin and breast cancer) and features for cancer patient management in cases of malignancy. While these suggestions could not be implemented in the current version due to time and budget constraints, they remain valuable directions for future system enhancements.
7. **Cross-Platform Deployment**: Although the mobile application was built using Flutter, which supports cross-platform development, an iOS version could not be compiled due to lack of access to a MacBook with Xcode and an iOS device. As a result, the application is currently only available to Android users. Future deployments should prioritize iOS compatibility to broaden user reach and inclusivity.
8. **Expand to Other Cancer Types**: A National Council Executive recommended extending the scope of the system to cover other forms of cancer, such as skin cancer, and breast cancer. This would significantly enhance the applicability and impact of the system across diverse medical domains. Future studies should explore datasets and models for additional cancer types, positioning the tool as a generalizable cancer risk assessment and diagnosis platform.
9. **Integrate Patient Management Features**: Interning medical students at Mbarara University suggested incorporating patient management functionalities, particularly for patients diagnosed with malignancy. Features such as treatment recommendations, referral alerts, follow-up tracking, and lifestyle adjustment suggestions could greatly improve clinical outcomes and support physicians in patient care.

## ****5.3 Conclusion****

This study successfully utilized two key datasets: the dataset “[Lung Cancer Prediction](https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link)” dataset comprising records of 1,000 patients with structured risk factors, and the “[Lung PET CT Dx](https://www.kaggle.com/datasets/rangan2510/lung-pet-ct-dx)” dataset containing 13,494 CT scan images. From the structured dataset, key risk features were extracted and used to train and validate an early lung cancer risk assessment model. In parallel, the CT scan images were used to develop a lung cancer type classification model for consistency testing. The study explored and assessed the performance of both traditional and deep learning models, including K-Nearest Neighbors (KNN), Decision Trees, Random Forest, AdaBoost, and an Artificial Neural Network (ANN). Based on 10-fold cross-validation, the ANN, Random Forest, and AdaBoost classifiers achieved perfect accuracy scores of 100%, while the Decision Tree model recorded the lowest performance at 98.6%. Additionally, the CT scan-based classifier for cancer type identification achieved a validation accuracy of 84%, indicating promising performance in image-based diagnostics.

The study also developed and implemented RESTful APIs for remote diagnosis and patient data management. These APIs support features such as data security, historical diagnosis tracking, and were successfully integrated into a user-friendly mobile application for Android devices. Comprehensive system testing was conducted to identify and resolve errors, usability issues, and security vulnerabilities, ensuring the protection of user privacy and enhancing the overall user experience. Overall, this study presents a functional and scalable framework for both early risk assessment of lung cancer and classification of cancer types using CT imaging—offering a step forward in intelligent, accessible, and data-driven healthcare solutions.

# **References**

|  |  |
| --- | --- |
| [1] | s. MB, "cote ML cancer progress and priorities," *lung cancer Epidemiol Biomakers,* vol. 28(10), pp. 1563-86, 2019. |
| [2] | J.chen, "Comparative analysis of machine learning models for lung cancer pridiction," *2023 IEEE international conference on image processing and computer applications,* vol. 10, pp. 242-246, 2023. |
| [3] | "https://data.world/cancerdatahp/lungcancer-data". |
| [4] | W. R., Dai W, Gong, Huang M, Hu T, Li H, Lin K, Tong T and Cai G, "development of a novel combined monogram model intergrating deep learning-pathomics, radomics and immunoscore to predic postoperative outcome of coleteral cancer lung metasis patients.," vol. 15(1), p. 11, 2022. |
| [5] | Ni J, Xu L, Li W, Zheng and Wu L, "targeted metabolomics for srum amino acids and acylcarnitinies in patients with lung cancer," *Exp Ther Med,* vol. 18, pp. 1888-98, 2019. |
| [6] | Peiffer-smadja N, Rowson Tn, Ahmad R, Buchard A and Holmes Ah, "machine learning for clinical decision support in infectious disease," *a narrative review of current applications,* vol. 26(5), pp. 584-95, 2020. |
| [7] | Dalal V, Carmicheal, Dhaliwal, Jain M and Batra SA, "Radiomics in stratification of pancreatic cystic lesions," *machine learning in action,* vol. 469, pp. 228-37, 2020. |
| [8] | HageChehada A, Abdallah A, marion Jn and Oueidat M, "Lung and colon cancer classificatioj using medical imaging," *a feature eengineering approach,* vol. 54(3), pp. 729-46, 2022. |
| [9] | Heuvelmans MA, van Ooijen PMA, Ather S and et al, "lung cancer prediction by deep learning to identify benign lung nodules," *lung cancer,* vol. 154, pp. 1-4, 2021. |
| [10] | g. sRUTHI, l Ram, M.K, Sai and P Sing, "cancer prediction using machine learning," in *innovative practices in technology and management*, 217-221, 2022. |
| [11] | A. Sarica, A. Cerasa and A Quattone, "random forest algorithm for the classification of neuroimaging data in alzheimers disease," *a systematic review,* p. 10, 2017. |
| [12] | M Minoor and V Baths, "diagnosis of breast cancer using random forests," *procedia Comp,* vol. 218, pp. 429-437, 2023. |
| [13] | Spiro, S.G and Silvestri, "one hundred years of lung cancer," *american journal of respiratory and critical care medicine,* vol. 172(5), pp. 523-529, 2005. |
| [14] | Bade, B. C, Dcela Cruz and C. S, "lung cancer 2020," *epidemiology, ethiology and prevention,* vol. 41(1), pp. 1-24, 2020. |
| [15] | N. Pudjihartono et al., "a review of feature selection methods for machine learning-based disease risk prediction," *front Bioinform,* vol. 2, p. 1, 2022. |
| [16] | S Mitra and S Ganguli, "cancer and nocoding RNAS," *introduction,* vol. 1, pp. 1-23, 2018. |
| [17] | E Dritsas and M Trigka, "lung cancer risk prediction with machine learning models," *big data,* vol. 6(4), p. 139, 2022. |
| [18] | D Endalie and W.T Adebe, "analysis of lung cancer risk factors from medical records in ethopia using machine learning," *digital health,* vol. 2(7), p. 10, 2023. |

# APPENDIX

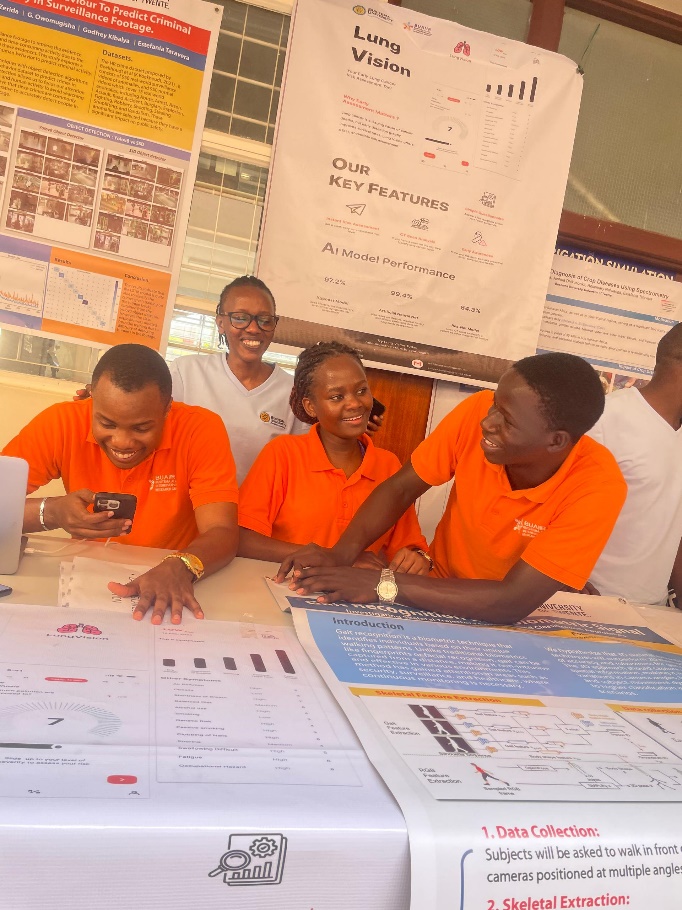
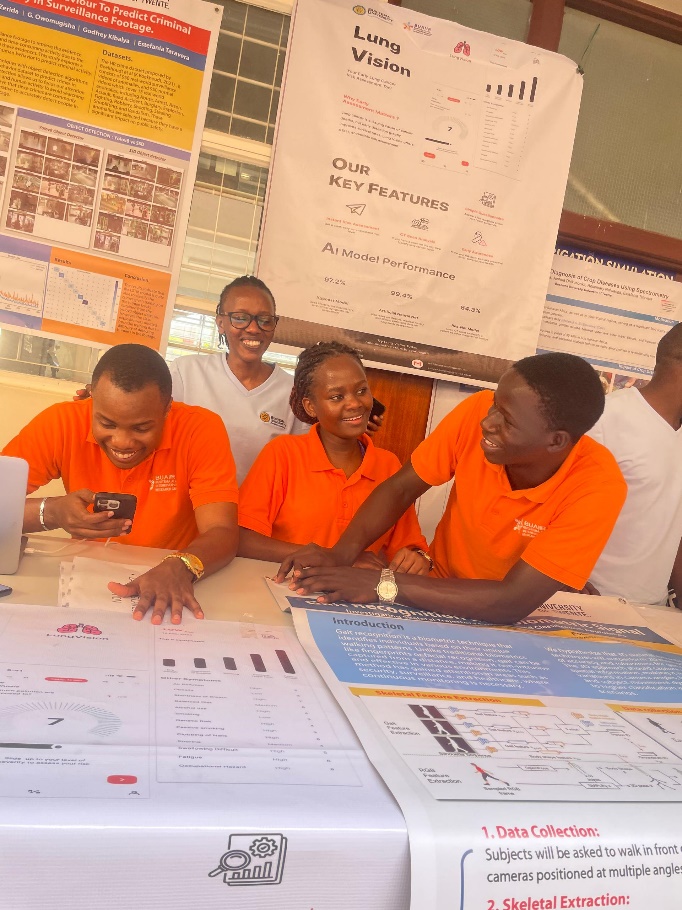
The following table provides URLs to all work for this entire project

|  |  |
| --- | --- |
| Resource | Description |
| [Lung Cancer Diagnosis Using Multi Risk Factor Analysis](https://colab.research.google.com/drive/1imLgVcEuuysN74F54nv9AVd9RU_zyczY?usp=sharing) | - Provides an in-depth account of the dataset characteristics, pre-processing strategies, model development workflow, hyper parameter optimization, training procedures, and evaluation metrics employed to assess model performance |
| [Lung-PET-CT-Dx | A Large-Scale CT and PET/CT Dataset Second Training](https://colab.research.google.com/drive/1EO91n7KFdOiJDgoeiQAjdSpt2Fhcw2gQ?usp=sharing) | -Provides an in-depth description of image dataset, pre-processing strategies, model development, hyper parameter optimization, training and evaluation of the model |
| [Mobile application mockups](https://www.figma.com/design/vnaEsiffj8bUSugoI3HAHG/Lung-Vision?node-id=0-1&p=f&t=uDV3SHE1Put48XTM-0) | - Presents mockups of the mobile application designed and implemented to facilitate end-user interaction |
| [Implementation of Restful APIs](https://github.com/wilfredKisitu/Lung_vision_backend_application) | -Presents the implementation of the Restful APIs for diagnosis, historical medical tracking and user security |
| [Implementation of mobile application](https://github.com/wilfredKisitu/Lung_vision_frontend) | -Presents the implementation of the mobile interfaces for end-to-end user interaction and API integration |

Table 20 Resources to entire project

EXHIBITION PRESENTATIONS OF THE WORK

****

****

****

****

The above represent presentation of the work for user testing to students, academic professionals from diverse backgrounds, fellow exhibitors and community members of Busitema University and Mbarara during the National Council Exhibition hosted in Busitema University and Mbarara University between 26th and 27th of March, 2025.