

**BUSITEMA
UNIVERSITY**
Pursuing Excellence

**FACULTY OF ENGINEERING AND TECHNOLOGY
DEPARTMENT OF COMPUTER ENGINEERING AND INFORMATICS**

**FINAL YEAR PROJECT REPORT
MULTIVARIATE RISK PREDICTORS FOR LUNG CANCER
DIAGNOSIS.**

By

NAME	REGISTRATION NUMBER
KISITU WILFRED	BU/UG/2021/0018
KAKOMO RAYAN AHMED	BU/UP/2021/1285

SUPERVISOR's NAME

Dr. OWOMUGISHA GODLIVER

**This Final Year Project Report submitted to the Department of
Computer Engineering and Informatics in a partial fulfillment of
the requirement for the award of the degree of Bachelor of Science
in Computer Engineering of Busitema University**

May, 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:

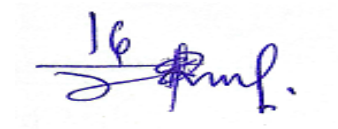
A handwritten signature in blue ink, appearing to read 'Kakomo Rayan Ahmed', with a stylized flourish at the end.

22nd May 2025

KAKOMO RAYAN AHMED

Signature:

Date:

A handwritten signature in blue ink, appearing to read 'Kisitu Wilfred', with a stylized flourish at the end.

22nd May 2025

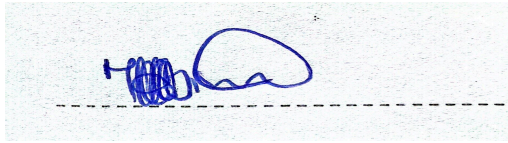
KISITU WILFRED

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



22nd May 2025

Dr. OWOMUGISHA GODLIVER

Department of Computer Engineering and Informatics

Faculty of Engineering and technology

Busitema University

DEDICATION.

We dedicate this report to our beloved parents Ms. Kajumba Sauya, Mr. Baagala Ahmed Luyirika, Mama Nabukera Matilda and Ms. Katusiime Shakina for the unrelenting financial, material and emotional support offered.

ACKNOWLEDGEMENT

We 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
Table 2 Label encoding mapping of categorical with ordinal relationship	15
Table 3 Hyper parameters of KNN Classifier	19
Table 4 Hyper parameters of Decision tree classifier	19
Table 5 Hyper parameters of Random forest classifier	20
Table 6 Hyper parameters of AdaBoost Classifier	21
Table 7 Hyper parameters of ANN Classifier	22
Table 8 Model Summary of Image Classification model.....	27
Table 9 Tools and libraries for API development.....	31
Table 10 Packages for frontend development using flutter	37
Table 11 Description of attributes	41
Table 12 Category of risk factors	41
Table 13 Comparison of model Performance	49
Table 14 Authentication API documentation	51
Table 15 User API documentation	51
Table 16 Diagnose API Documentation	52
Table 17 Ct Diagnose API Documentation	52
Table 18 Status codes	53
Table 19 GitHub repository for project source codes.....	56
Table 20 Resources to entire project.....	66

LIST OF FIGURES.

Figure 1visualization of missing values	14
Figure 2 Visualization of missing values.....	14
Figure 3 Distribution of targets.....	16
Figure 4 ANN model Summary.....	22
Figure 5 ANN Training and validation accuracy	23
Figure 6 CT, PET scans of Lung Cancer patients.....	25
Figure 7 Class distribution of Imagery dataset	26
Figure 8 Total number of model parameters.....	27
Figure 9 Training parameters of CT classification image model	28
Figure 10 Training and validation accuracy of the Cancer Variant Classification model	29
Figure 11 Entity Relation Diagram of the Backend service	32
Figure 12 Postman showing the API endpoints developed	33
Figure 13: logging into the remote server	33
Figure 14: successfully logged into server.	34
Figure 15: system information of the remote server.	34
Figure 16: kernel information of the remote server.....	35
Figure 17: RAM information of the remote server.....	35

Figure 18: disk information of the remote server.	35
Figure 19 Distribution of Attributes.....	42
Figure 20 Mutual information scores	43
Figure 21 Correlation Matrix.....	44
Figure 22 Performance report of KNN Classifier.....	45
Figure 23 Performance of the Decision tree classifier	46
Figure 24 Performance report of Random Forest classifier.....	46
Figure 25 Performance report of AdaBoost Classifier	47
Figure 26 Confusion matrix of the four machine learning models	48
Figure 27 Performance report of ANN Classifier	48
Figure 28 Wireframes for the frontend application	55
Figure 29 Postman showing the API endpoints developed	57

TABLE OF CONTENTS.

Contents

DECLARATION	i
APPROVAL.....	ii
DEDICATION.....	iii
ACKNOWLEDGEMENT.....	iv
LIST OF ACRONYMS.....	v
LIST OF TABLES.....	vi
LIST OF FIGURES.....	vi
TABLE OF CONTENTS.....	viii
CHAPTER ONE: INTRODUCTION.....	1
1.1 Background.	1
1.1.1 Why machine learning?	1
1.2 Problem statement.....	2
1.3 Objectives of study.....	3
1.3.1 Main objective	3
1.3.2 Specific objectives.....	3
1.4 Justification	3
1.5 Significance	4
1.6 Scope of study.....	4
1.6.1 Content scope.....	4
1.6.2 Geographical scope.....	4
1.6.3 Time scope.....	5
1.6.4 Unit of measurement.....	5
CHAPTER TWO: LITERATURE REVIEW	6
2.1 INTRODUCTION.....	6
2.2 RELATED WORKS.....	6
2.3 EXISTING WORKS.....	8
2.4 COMPARISON OF THE EXISTING SYSTEMS.....	9
2.5 DEVELOPED SYSTEM	11
CHAPTER THREE: METHODOLOGY.....	13
3.0 INTRODUCATION.....	13

3.1 SPECIFIC OBJECTIVE ONE: TO IDENTIFY AN EXISTING DATASET THAT CAPTURES THE LIFESTYLE, DEMOGRAPHICS, MEDICAL HISTORY AND CLINICAL SYMPTOMS OF A PATIENT.	13
3.1.1 Exploration of Dataset Repositories	13
3.2 SPECIFIC OBJECTIVE TWO: TO EXTRACT FEATURES FOR EARLY LUNG CANCER DIAGNOSIS	14
3.2.1 Data Processing and Preparation.....	14
3.2.2 Mutual Information Analysis.....	17
3.2.3 Correlation Matrix Analysis.....	17
3.3 SPECIFIC OBJECTIVE THREE: MODEL TRAINING EVALUATION OF MACHINE LEARNING MODELS ...	18
3.3.1 Development of Various Machine Learning Models	18
3.3.2 Hyper parameter Tuning.....	23
3.3.3 Model Evaluation	23
3.3.4 Model Validation Using Image-Based Analysis	24
3.4 SPECIFIC OBJECTIVE FOUR: TO DEVELOP A MOBILE APPLICATION INTERFACE AND INTEGRATE IT WITH MACHINE LEARNING MODEL	30
3.4.2 Mobile User Interface Development	35
3.4.4 Unit and Integration Testing	37
CHAPTER FOUR: DISCUSSIONS OF RESULTS AND TESTING.....	39
4.1 DISCUSSION OF RESULTS.	39
4.1.1 TO IDENTIFY AN EXISTING DATASET THAT CAPTURES THE LIFESTYLE, DEMOGRAPHICS, MEDICAL HISTORY, AND CLINICAL SYMPTOMS OF LUNG CANCER PATIENTS.....	39
Most Suitable Data Selection	39
Dataset Description and structure.....	39
4.1.2 SPECIFIC OBJECTIVE TWO: TO EXTRACT FEATURES FOR EARLY LUNG CANCER DIAGNOSIS	43
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	45
4.1.4 OBJECTIVE FOUR: TO DEVELOP A MOBILE APPLICATION INTERFACE AND INTEGRATE IT WITH MACHINE LEARNING MODEL.....	51
4.2 TESTING.....	57
CHAPTER FIVE: RECOMMENDATIONS AND CONCLUSION.....	59
5.1 Challenges	59
5.2 Feedback	60
5.3 Recommendations	61
5.3 Conclusion.....	63
References	64

APPENDIX.....	66
Lung Cancer Diagnosis Using Multi Risk Factor Analysis	66
Lung-PET-CT-Dx A Large-Scale CT and PET/CT Dataset Second Training	66
Mobile application mockups.....	66
Implementation of Restful APIs	66
Implementation of mobile application	66

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.