MULTIVARIATE RISK PREDICTORS FOR LUNG CANCER DIAGNOSIS.

KAKOMO RAYAN AHMED, KISITU WILFRED, DR. OWOMUGISHA GODLIVER.

Department Of Computer Engineering and Informatics, Faculty of Engineering, Busitema University.

ABSTRACT

Lung cancer remains one of the leading causes of cancer-related mortality, with a five-year survival rate as low as 8%, primarily due to over 70% of cases being diagnosed at advanced stages (III and IV) when surgical intervention is no longer viable [4] , [5]. Early detection significantly improves outcomes, with survival rates reaching up to 55% [4]. Recent advancements in machine learning have shown promise in aiding early diagnosis through the analysis of complex medical data. However, existing models often fall short due to limited input features or reliance on biomarkers requiring further clinical validation [9]. For instance, models based solely on metabolic indices or those omitting critical factors such as demographics and clinical symptoms fail to capture the full spectrum of diagnostic indicators, leading to late detection, false positives, and low sensitivity. This study proposes a machine learning framework that integrates lifestyle factors, demographics, medical history, and clinical symptoms to enhance the sensitivity and accuracy of early lung cancer detection. By addressing key limitations in prior research, the proposed model aims to support more effective clinical decision-making and improve patient outcomes.

I. INTRODUCTION.

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].

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.

II. LITERATURE REVIEW.

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.

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.

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.

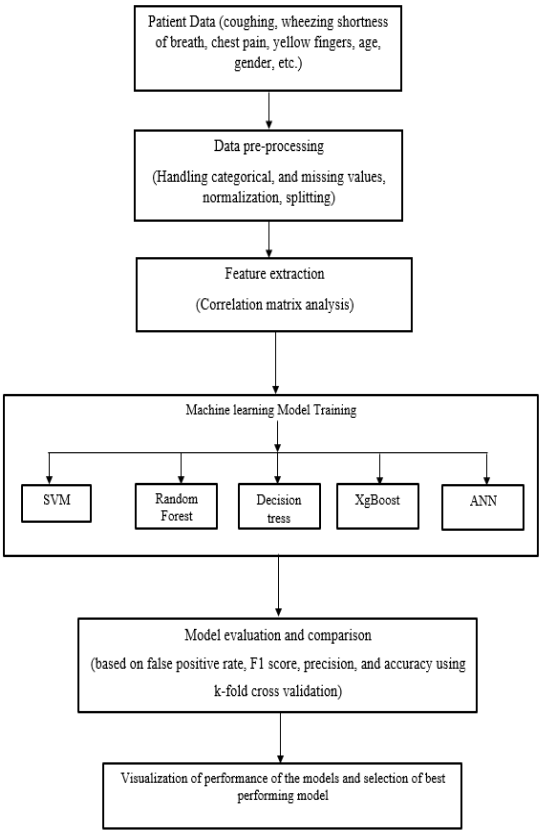
III. DEVELOPED SYSTEM.

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

IV. SYSTEM DESIGN



V. RESULTS AND DISCUSSIONS.

DATASET SELECTION AND DESCRIPTION

For this study, the “Lung Cancer Prediction” dataset from Kaggle was selected due to its comprehensive coverage of four key categories relevant to lung cancer diagnosis: demographics, lifestyle factors, medical history, and clinical symptoms. The dataset consists of 1,000 patient records with 21 attributes, making it suitable for building machine learning models aimed at early detection of lung cancer.

The attributes include:

- Demographics: Age, Gender
- Lifestyle factors: Air pollution exposure, Alcohol consumption, Occupational hazards, Diet quality, Obesity, Smoking, Passive smoking
- Medical history: Dust allergy, Genetic predisposition, Chronic lung disease
- Clinical symptoms: Chest pain, Coughing of blood, Fatigue, Weight loss, Shortness of breath, Wheezing, Difficulty swallowing, Clubbing of fingernails, and Snoring

The dataset supports key analytical objectives including:

1. Predicting the likelihood of lung cancer development,
2. Identifying key risk factors, and
3. Supporting treatment decision-making.

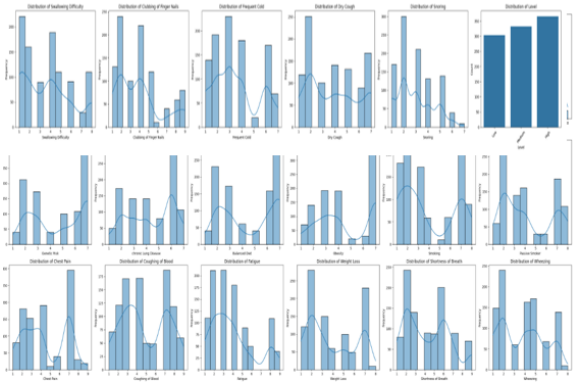
Data Exploration and Structure

The dataset includes both categorical and numerical attributes. Most features are categorical, representing levels or severity of exposure/symptoms, while age is the only numerical attribute.

A structured exploration revealed balanced representation across features, enabling robust categorization into the predefined risk factor groups. This categorization is critical in supporting a multifactorial machine learning framework that mirrors real-world diagnostic settings.

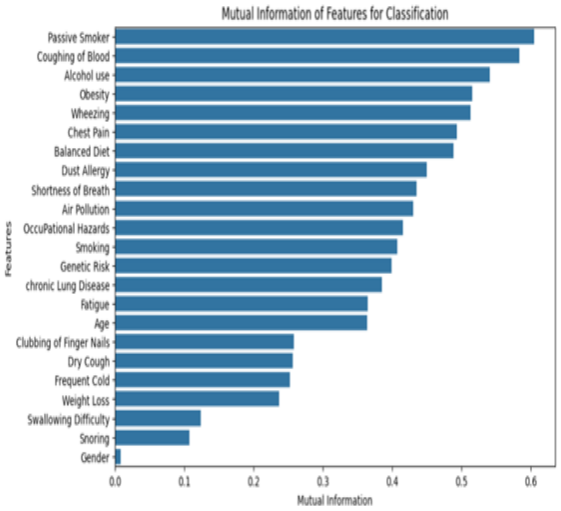
Key Findings

- The dataset effectively captures a broad spectrum of risk indicators relevant to lung cancer, filling gaps observed in previous studies that either lacked demographic, lifestyle, or symptomatic data.
- It aligns with the study's goal of integrating multiple patient factors to enhance model sensitivity and diagnostic accuracy.
- The structure and diversity of the dataset offer a solid foundation for training predictive models that can better differentiate between high-risk and low-risk individuals.

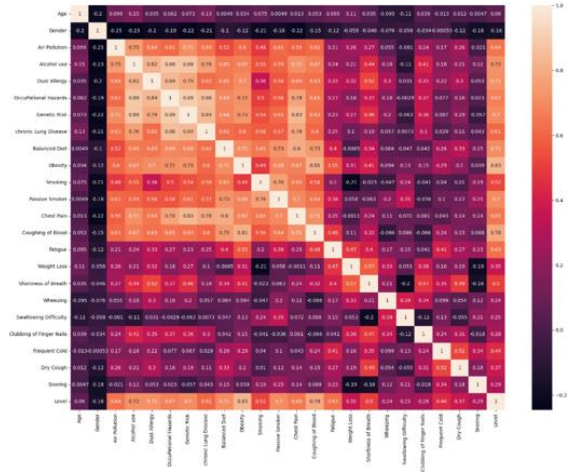


EXTRACTING THE KEY FEATURES FOR LUNG CANCER.

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.

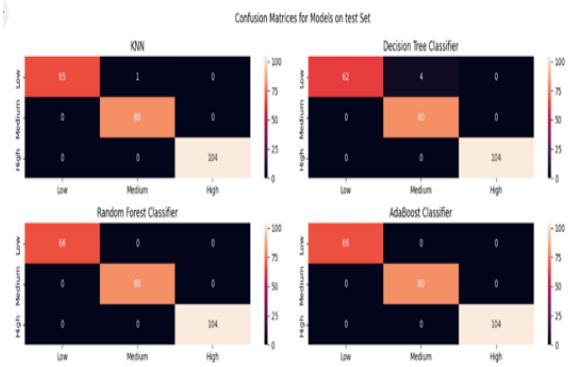


The correlation matrix analysis showed that **smoking history, age, chronic disease history, coughing, shortness of breath, and chest pain** had strong positive correlations with lung cancer diagnosis. These features were the most statistically significant and clinically relevant, confirming their importance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



MODEL PERFORMANCE REPORTS.

Five models were trained on the structured dataset and to better visualize and evaluate the performance of the 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.



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.

MOBILE APPLICATION DEVELOPMENT AND INTERGRATION WITH MACHINE LEARNING MODEL.

API DEVELOPMENT

We successfully developed four (4) APIs to facilitate seamless end-user interaction. These APIs support key functionalities,

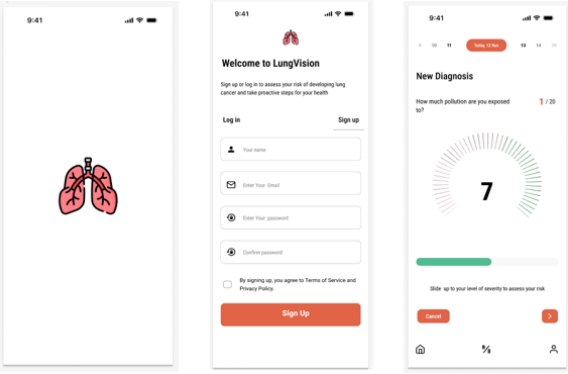
including user authentication, risk factor–based diagnosis, CT image–based diagnosis, and historical medical record tracking.

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.

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.



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:

a) **Ease of Use.** Users found the app intuitive and easy to navigate.
b) **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.

VI. 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" for risk factors and "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" 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 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

VII. 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:

A. **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.

B. **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.

C. **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.

D. **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.

E. **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.

F. **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.

G. **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.

H. **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.

I. **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.

VIII. CONCLUSSION

This study successfully utilized two key datasets: the dataset "Lung Cancer Prediction" dataset comprising records of 1,000 patients with structured risk factors, and the "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.

IX. 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.