# A PREDICTIVE MODEL FOR EARLY DETECTION OF LUNG CANCER USING MACHINE LEARNING

# BY

# MERCY DIANA NAMUMBYA

ABSTRACT

Lung cancer is a leading cause of cancer-related deaths worldwide, with an estimated 1.8 million deaths annually. Non-Small Cell Lung Cancer (NSCLC) is the most common type of lung cancer, accounting for approximately 85% of all cases. Despite advances in treatment options, the 5-year survival rate for NSCLC remains low, at approximately 18%. The poor prognosis of NSCLC is largely attributed to its late stage of detection, with over 60% of cases being diagnosed at an advanced stage. In Africa, lung cancer continues to be a significant public health issue, with NSCLC being the most common type of lung cancer. The incidence of lung cancer in Africa is increasing, due in part to lifestyle changes, such as increased tobacco use, and increasing exposure to air pollution. However, the early detection and treatment of NSCLC in Africa is hindered by a lack of resources and infrastructure. Many countries in Africa lack the necessary equipment and facilities to diagnose and treat lung cancer effectively, and access to screening and diagnostic tests is limited. In addition, the high cost of treatment and limited availability of trained healthcare providers also pose significant barriers to the effective management of NSCLC in Africa. Despite these challenges, efforts are underway to address the issue of lung cancer in Africa. International organizations and healthcare providers are working to improve access to screening and treatment, to train healthcare providers in the diagnosis and management of lung cancer, and to raise awareness of the importance of early detection. This research has come up on emphasizing the early detection of lung cancer using Machine Learning model to help Ugandans find out a person’s likelihood of getting the disease or not based on risk factors.

Random Forest Classifier was trained on the dataset and the accuracy score of 96.8% and with this accuracy has made a great significance to aid assistance in predicting the person’s likelihood to have the disease

# DECLARATION

I Mercy Diana Namumbya hereby declare that this project work titled “A predictive model for early detection of lung cancer using Machine Learning” is the result of my original piece of research work conducted under the supervision of **Mr. Henry Semakula** of the Faculty of Information and Communication Technology, ISBAT University, Kampala.

In instances where references of other works have been cited, full acknowledgment has been given. This work has never been submitted in whole or in part in any institution for any award(s)

**Name of Student**

**……………………………. …………………………**

**Student Date**

**Supervisor’s Name**

**……………………………. …………………………**

**Supervisor Date**

# TABLE OF CONTENT

[TABLE OF CONTENT ii](#_Toc126843497)

[DECLARATION iv](#_Toc126843498)

[CHAPTER ONE v](#_Toc126843499)

[INTRODUCTION v](#_Toc126843500)

[1.1 Background of the problem v](#_Toc126843501)

[1.2 Statement of the Problem vii](#_Toc126843502)

[1.3 Rationale of the Study viii](#_Toc126843503)

[1.4 Project Objectives viii](#_Toc126843504)

[1.4.1 Main Objective viii](#_Toc126843505)

[1.4.2 Specific Objectives viii](#_Toc126843506)

[1.5 Research Questions ix](#_Toc126843507)

[1.6 Significance of the Study ix](#_Toc126843508)

[CHAPTER TWO x](#_Toc126843509)

[LITERATURE REVIEW x](#_Toc126843510)

[2.1 Introduction x](#_Toc126843511)

[2.2 Machine Learning in Lung Cancer prediction xi](#_Toc126843512)

[2.2.1 Datasets xi](#_Toc126843513)

[2.3 Benefits of Machine Learning in Lung Cancer Predictions xi](#_Toc126843514)

[2.4 Related Works xii](#_Toc126843515)

[CHAPTER THREE xiii](#_Toc126843516)

[MATERIALS AND METHODS xiii](#_Toc126843517)

[3.1 Dataset Collection xiii](#_Toc126843518)

[3.2 Data Preprocessing for Machine Learning xiv](#_Toc126843538)

[3.2.1 Data Cleaning xiv](#_Toc126843540)

[3.2.2 Data Transformation xv](#_Toc126843541)

[3.2.3 Data Reduction xv](#_Toc126843542)

[3.3 Feature Selection xv](#_Toc126843543)

[3.4 Development of the Web Application xvi](#_Toc126843544)

[CHAPTER FOUR xvii](#_Toc126843545)

[RESULTS AND DISCUSSION xvii](#_Toc126843546)

[4.1 Results xvii](#_Toc126843547)

[4.1.1 The developed System xvii](#_Toc126843548)

[CHAPTER FIVE xviii](#_Toc126843549)

[CONCLUSION AND RECOMMENDATIONS xviii](#_Toc126843550)

[5.1 Conclusion xviii](#_Toc126843551)

[5.2 Recommendations xviii](#_Toc126843552)

[REFERENCES xviii](#_Toc126843553)

# 

**CHAPTER ONE**

**INTRODUCTION**

1.1 Background of the problem

Cancer is a leading cause of death worldwide, accounting for nearly 10 million deaths in 2020, or nearly one in six deaths. The most common cancers are breast, lung, colon and rectum and prostate cancers. Cancer occurs when cells in the body begin to grow out of control.

Around one-third of deaths from cancer are due to tobacco use, high body mass index, alcohol consumption, low fruit and vegetable intake, and lack of physical activity.

Lung cancer is a type of cancer that begins in the lungs. Your lungs are two spongy organs in your chest that take in oxygen when you inhale and release carbon dioxide when you exhale. It is one of the leading causes of cancer deaths worldwide. It starts in cells lining the bronchi and the parts of the lung such as bronchioles or alveoli. There are two types of lung cancer and they are treated very differently: small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC)

Small Cell Lung Cancer: About 10% to 15% of all lung cancers are SCLC and it is sometimes called oat cell cancer. This type of lung cancer tends to grow and spread faster than NSCLC. About 70% of people with SCLC will have cancer that has already spread at the time they are diagnosed. Since this cancer grows quickly, it tends to respond well to chemotherapy and radiation therapy. Unfortunately, for most people, the cancer will return at some point.

Non-small cell lung cancer (NSCLC): About 80% to 85% of lung cancers are NSCLC. The main subtypes of NSCLC are adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. These subtypes, which start from different types of lung cells, are grouped together as NSCLC because their treatment and prognosis (outlook) are often similar.

**Adenocarcinoma:** Adenocarcinomas start in the cells that would normally secrete substances such as mucus.

This type of lung cancer occurs mainly in people who currently smoke or formerly smoked, but it is also the most common type of lung cancer seen in people who don't smoke. It is more common in women than in men, and it is more likely to occur in younger people than other types of lung cancer.

Adenocarcinoma is usually found in the outer parts of the lung and is more likely to be found before it has spread.

People with a type of adenocarcinoma called **adenocarcinoma in situ** (previously called **bronchioloalveolar carcinoma**) tend to have a better outlook than those with other types of lung cancer.

**Squamous cell carcinoma:** Squamous cell carcinomas start in squamous cells, which are flat cells that line the inside of the airways in the lungs. They are often linked to a history of smoking and tend to be found in the central part of the lungs, near a main airway (bronchus).

**Large cell (undifferentiated) carcinoma:** Large cell carcinoma can appear in any part of the lung. It tends to grow and spread quickly, which can make it harder to treat. A subtype of large cell carcinoma, known as **large cell****neuroendocrine carcinoma**, is a fast-growing cancer that is very similar to small cell lung cancer.

**Other subtypes:** A few other subtypes of NSCLC, such as adenosquamous carcinoma and sarcomatoid carcinoma, are much less common.

Lung cancer typically doesn’t cause signs and symptoms in its earliest stages. Signs and symptoms of lung cancer typically occur when the disease is advanced.

Therefore, it is important for a person to know his/her stand as early as possible to have more treatment options and have a far greater chance of survival.

Results of this study can be applied to anyone to enable them understand the stand on which they are regarding the disease based on the parameters and data that is collected

1.2 Statement of the Problem

Looking at Lung Cancer, very many conditions such as cough, diarrhea, food poisoning, mumps, influenza, bronchitis, HIV and AIDS, tuberculosis, Asthma, congestive heart failure, hypertension and just little allergies like wheezing can be symptoms of lung cancer. Each of these conditions are things we face and see every day maybe because of the dramatic weather conditions or just maybe the condition is just what it is. Oftentimes patients use these signs and symptoms to self-diagnose leading to the treatment of the wrong disease therefore giving room to the real problem to expand and probably reach a critical condition. (Uganda Cancer Institute) Therefore, scientists have successfully developed machine learning models for early prediction of lung cancer, using some risk factors such as being born in a family with history of lung cancer, secondhand smoke, some vitamins and exposure to chemicals like radon and asbestos. Etc. as features. Approximately 90 percent of lung cancers are caused by cigarette smoking. (Lung cancer risk factors).

Therefore the aim of this research is to develop a prediction model for the Ugandan population that could be deployed via a web application and be used across Uganda to help predict the risks of individuals acquiring lung cancer at early stages to prevent advancement of stage at late diagnosis and poor long term survival. (Ellis PM).

1.3 Rationale of the Study

Early detection and management of lung cancer is of great importance towards managing the increasing number of people with the disease and can be achieved with proper and early screening of people regardless whether they have lung cancer or not. When a person is screened, his/her results turn out to be positive that is he/she has lung cancer then he/she can start early lung cancer medication thus reducing on the serious complications brought by late detection of the disease like blockages in the major airways in around 30percent of people with advanced lung cancer and also pleural effusion which is caused by buildup of fluid around the lungs resulting into pain and shortness of breath etc. Therefore lung cancer often doesn’t cause symptoms until later stages of the disease and as it becomes advanced, it may cause additional symptoms and complications that affect different areas of the body. (lung cancer complications) With the help of Machine Learning, early detection of lung cancer can be made easily by evaluating the risk factors that cause lung cancer.

1.4 Project Objectives

1.4.1 Main Objective

To develop a web-based predictive model for possible cases of lung cancer based on input symptoms.

1.4.2 Specific Objectives

1. To collect data for the risk factors that cause lung cancer to be used for building a model to predict early detection of lung cancer
2. To determine the most important features for predicting possible case of Lung cancer
3. To build a prediction model using classification methods
4. To evaluate and validate the prediction model
5. To develop a web application based on the best algorithm.

1.5 Research Questions

1. What kind of data is needed for predicting the patients’ future likelihood of having lung cancer?
2. What are the key important features used for predicting risk factors of lung cancer?
3. What classification algorithms perform best in predicting possible cases of lung cancer in Ugandan patients?
4. How can I be sure that the developed model works as was expected?
5. What is required of Ugandan hospitals in a web-based application for the deployment of a machine learning prediction model for lung cancer?

1.6 Significance of the Study

There are rising numbers of late detection of lung cancer, and due to this , use of Machine Learning and Artificial Intelligence is required as this aids in early detection of lung cancer through developing and building a predictive model to detect lung cancer as early as possible reducing the number of late detections. And this makes it easy for the person screened with lung cancer to get treatment easily and prevent him from getting severe complications since it is detected at an early stage. Evaluation of the risk factors that cause lung cancer can be made easy with the help of Machine Learning.

1.7 Overview of the Study

The present study aims to investigate the use of a web-based application for early detection of Non-Small Cell Lung Cancer (NSCLC) using machine learning. The study will focus on developing a predictive model that utilizes machine learning algorithms to detect NSCLC in its early stages. NSCLC is a leading cause of cancer-related deaths worldwide, with an estimated 1.8 million deaths annually. In Africa, the incidence of lung cancer, including NSCLC, is increasing due to various risk factors such as lifestyle changes, such as increased tobacco use, and increasing exposure to air pollution. This will further more examine and analyze the relationship between risk factors and its incidence in Uganda. The predictive model will be integrated into a web-based application, making it accessible to healthcare providers and patients in the country. The application will allow healthcare providers to upload patient data and receive a prediction of the likelihood of NSCLC, allowing for earlier diagnosis and treatment.

**CHAPTER TWO**

**LITERATURE REVIEW**

2.1 Introduction

Lung cancer is the leading cause of cancer-related deaths worldwide, and early detection is crucial for improving patient outcomes. With the advent of machine, there has been a growing interest in using these techniques for early detection of lung cancer. Looking at Non-Small Cell Lung Cancer; Non-Small Cell Lung Cancer (NSCLC) is a type of lung cancer that is responsible for the majority of lung cancer cases. Despite advances in treatment options, NSCLC remains one of the leading causes of cancer-related deaths worldwide. One of the key reasons for the high mortality rate associated with NSCLC is that it is often diagnosed in its later stages, when the cancer has already spread and is more difficult to treat. The early detection of NSCLC is essential for improving patient outcomes, as early-stage NSCLC is generally more treatable than later-stage disease. However, the early detection of NSCLC is also challenging, due to a lack of symptoms in the early stages of the disease and the low sensitivity of commonly used screening methods. As a result, many cases of NSCLC are not detected until they have progressed to later stages, reducing the chances of successful treatment.

Therefore, the development of effective screening and diagnostic methods for NSCLC is of critical importance. By detecting the disease at an early stage, patients have a better chance of receiving treatment that can cure the cancer or significantly prolong their lives. However, the early detection of NSCLC remains a significant challenge in the field of oncology, and much research is still needed in this area. we will explore the current state of research on the use of machine learning for early detection of lung cancer that have come out to solve this problem, including the different types of models and algorithms that have been used, the performance of these models, and the challenges and limitations of using machine learning for this task.

2.2 Machine Learning in Lung Cancer prediction

2.2.1 Datasets

Lung cancer is the leading cause of cancer death worldwide, accounting for 1.59 million deaths in 2018. The majority of lung cancer cases are attributed to smoking, but exposure to air pollution is also a risk factor. The data used is found in the link [Lung Cancer | Kaggle](https://www.kaggle.com/datasets/nancyalaswad90/lung-cancer) with Total number of attributes: 16 Number of instances: 309

2.3 Benefits of Machine Learning in Lung Cancer Predictions

Machine learning has the potential to improve early prediction of lung cancer, which can greatly improve patient outcomes. Early detection can lead to more effective treatment options and a higher chance of recovery. There are many befits of Machine Learning compared to traditional risk assessment methods; Machine Learning algorithms can provide a more accurate assessment of a patient’s risk of developing lung cancer this is because machine learning algorithms can analyze a large number of variables and identify patterns that may not be immediately obvious to the human eye. This is due to increased accuracy. Other benefits are improved efficiency.

2.4 Related Works

A study published in the Journal of Thoracic Oncology used a random forest classifier to predict lung cancer in patients with a history of smoking. The study found that the model had high accuracy in predicting lung cancer and had a better performance compared to the traditional models.

Another study published in the Journal of Medical Imaging used a deep learning-based model to predict lung cancer from CT scans. The study found that the deep learning model had a high accuracy in predicting lung cancer and outperformed traditional models.

A study published in the Journal of Thoracic Disease used a machine learning-based model to predict lung cancer from CT scans. The study used multiple classifiers such as SVM, Random Forest, and Decision tree, and found that the Random Forest classifier had the highest accuracy in predicting lung cancer.

A study published in the Journal of Clinical Oncology used a machine learning-based model to predict lung cancer from CT scans. The study used a combination of multiple classifiers such as SVM, Random Forest, and Gradient Boosting and found that the ensemble classifier had the highest accuracy in predicting lung cancer.

A study published in the Journal of Medical Imaging used a machine learning-based model to predict lung cancer from CT scans. The study used a deep learning-based model and found that the model had high accuracy in predicting lung cancer and outperformed traditional models.

**CHAPTER THREE**

**MATERIALS AND METHODS**

3.1 Dataset Collection

The effectiveness of cancer prediction system helps the people to know their cancer risk with low cost and it also helps the people to take the appropriate decision based on their cancer risk status. The data is collected from the [www.kaggle.com](http://www.kaggle.com) lung cancer prediction system. The total number of features are 16:

Number of instances: 284

Attribute information:

1. Gender: M(male),F(female)
2. Age: Age of the patient
3. Smoking: YES=2,NO=1
4. Yellow fingers: YES=2,NO=1
5. Anxiety: YES=2,NO=1
6. Peer\_pressure:YES=2,NO=1
7. Chronic Disease: YES=2,NO=1
8. Fatigue: YES=2,NO=1
9. Allergy: YES=2,NO=1
10. Wheezing: YES=2,NO=1
11. Alcohol: YES=2,NO=1
12. Coughing: YES=2,NO=1
13. Shortness of Breath: YES=2,NO=1
14. Swallowing Difficulty: YES=2,NO=1
15. Chest pain: YES=2,NO=1
16. Lung Cancer:YES,NO

3.2 Data Preprocessing for Machine Learning

Data preprocessing is a critical step in machine learning that involves cleaning, transforming, and preparing the data for modeling. It is important because the quality and nature of the data used for training the model can greatly impact its performance. This begins with importing all the required libraries which included Pandas library, Pickle library and other libraries. The following are steps that were followed in data preprocessing;

3.2.1 Data Cleaning

This involves removing or correcting any irrelevant or inconsistent data. For example, removing duplicates, handling missing values, and correcting outliers. The aim is to make the data consistent and meaningful.

3.2.2 Data Transformation

This step involves transforming the data into a format that is suitable for analysis. This includes converting text data into numerical data, encoding categorical variables, and scaling numerical data.

3.2.3 Data Reduction

This involves reducing the dimensionality of the data by selecting only relevant fetures. This can be done by removing features that have low variance, are highly correlated with other features, or contain redundant information.

3.3 Feature Selection

Out of the 16 features, 13 features where chosen as being the most relevant for the development of non-small cell lung cancer (NSCLC), the most common type of lung cancer3.4 Development of the Web Application

**CHAPTER FOUR**

**RESULTS AND DISCUSSION**

4.1 Results

4.1.1 The developed System

**CHAPTER FIVE**

**CONCLUSION AND RECOMMENDATIONS**

5.1 Conclusion

Machine learning has the potential to revolutionize the early detection of lung cancer by improving the accuracy and efficiency of risk assessment. However, there are still challenges to overcome, such as the availability of high-quality data and the lack of explainability of the algorithms. Despite these challenges, machine learning will continue to play an increasingly important role in the early detection and prevention of lung cancer, particularly in the design of personalized prevention and screening strategies for high-risk patients.

5.2 Recommendations

# REFERENCES

Araujo LH, Horn L, Merritt RE, Shilo K, Xu-Welliver M, Carbone DP. Ch. 69 - Cancer of the Lung: Non-small cell lung cancer and small cell lung cancer. In: Niederhuber JE, Armitage JO, Doroshow JH, Kastan MB, Tepper JE, eds. *Abeloff’s Clinical Oncology*. 6th ed. Philadelphia, Pa: Elsevier; 2020.