**Cancer Watch: ML for lung tumor detection (GA and Naïve Bayes) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Submitted by**

**Laraib Masood** (BSCS-534-C)

**Amina Nawaz** (BSCS-093-C)

Session 2020-2024

**Supervised by**

**Sir Umer Ahmed**



**Department of Computer Science**

**Lahore Garrison University**, **Lahore**

**Lung Cancer Detection Using Machine Learning Approach (Hybrid GA-Naïve Bayes)**

A project submitted to the

Department of Computer Science In

Partial Fulfilment of the Requirements for the

Bachelor’s Degree in Computer Science

By

**Laraib Masood**

**Amina Nawaz**

**Internal Supervisor**

Umer Ahmed \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Lecturer

Department of Computer Science

**External Examiner \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Chairperson**

**Dr. Arfan Ali Nagra \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Head Of Department (HOD)

Department of Computer Science

# COPYRIGHTS

This is to certify that the project titled “**Lung Cancer Detection Using Machine Learning Approach (Hybrid GA-Naïve Bayes)**”is the genuine work carried out by **Amina Nawaz and Laraib Masood S**tudents of BSCS of Computer Science Department, Lahore Garrison University, Lahore during the academic year 2020-24, in partial fulfilment of the requirements for the award of the degree of Bachelor of Computer Science and that the project has not formed the basis for the award previously of any other degree, diploma, fellowship or any other similar title.

Laraib Masood \_\_\_\_\_\_\_\_\_\_\_\_

Amina Nawaz \_\_\_\_\_\_\_\_\_\_\_\_

# DECLARATION

This is to declare that the project entitled “**Lung Cancer Detection Using Machine Learning Approach (Hybrid GA-Naïve Bayes)**” is an original work done by the undersigned, in partial fulfilment of the requirements for the degree “Bachelor of Science in Computer Science” in Computer Science Department, Lahore Garrison University, Lahore. All the analysis, design, and system development have been accomplished by the undersigned.

Moreover, this project has not been submitted to any other college or university.

**Group Members**

Laraib Masood \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Amina Nawaz \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Supervisor**

Sir Umer Ahmed \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# ACKNOWLEDGEMENTS

First and foremost, countless thanks to ALLAH ALMIGHTY, who created us worthy of all praise, who counsels us in times of crisis. Secondly, all salutations are offered to our beloved PROPHET MUHAMMAD (PBUH) who gives us the strength to recognize our creator and whose life is the best example for us.

During this project we received support, help or practical help from many people and it is blissful to acknowledge their intentions. Our heartfelt prayers go out to our parents, siblings and all family members for their help and moral support.

Our success is possible only because of the prayers or support of our mothers who really love us and always pray for us in every moment of our life We are very grateful to all our family members, senior colleagues and also feel deeply indebted to our supervisor Sir Umer Ahmed for providing us accurate research-oriented opinions, advice, support and help throughout the project and during the academic session.

# DEDICATION

I dedicated this work to First of all, ALLAH Almighty who glorified us with the knowledge and bravery to complete this responsibility with elegance. Secondly, my affectionate supportive family whose prayers, advice, and continuous support played a major role in achieving this goal. Finally, to the Lahore Garrison University and Especially, the Department of Computer Science.

**TABLE OF CONTENT**

[COPYRIGHTS i](#_Toc174390017)

[DECLARATION ii](#_Toc174390018)

[ACKNOWLEDGEMENTS iii](#_Toc174390019)

[DEDICATION iv](#_Toc174390020)

[LIST OF TABLES v](#_Toc174390021)

[LIST OF FIGURES vi](#_Toc174390022)

[ABSTRACT 13](#_Toc174390023)

[1.1 Background 14](#_Toc174390024)

[1.2 Objectives 14](#_Toc174390025)

[1.3 Gantt Chart 15](#_Toc174390026)

[1.4 Summary of Report Organization 16](#_Toc174390027)

[Chapter 2 18](#_Toc174390028)

[LITERATURE REVIEW 18](#_Toc174390029)

[Chapter 3 21](#_Toc174390030)

[PROBLEM DEFINITION 21](#_Toc174390031)

[Chapter 4 24](#_Toc174390032)

[SOFTWARE REQUIREMENT 24](#_Toc174390033)

[SPECIFICATION 24](#_Toc174390034)

[4.1 Introduction 24](#_Toc174390035)

[*4.1.1 Purpose* 24](#_Toc174390036)

[*4.1.2 Document Conventions* 25](#_Toc174390037)

[*4.1.3 Intended Audience and Reading Suggestions* 26](#_Toc174390038)

[*4.1.4 Product Scope* 26](#_Toc174390039)

[4.2 Overall Description 26](#_Toc174390040)

[*4.2.1 Product Perspective* 26](#_Toc174390041)

[*4.2.2 Product Functions* 27](#_Toc174390042)

[*4.2.3* *User Classes and Characteristics* 29](#_Toc174390043)

[*4.3.4 Operating Environment* 30](#_Toc174390044)

[*4.3.5 Design and Implementation Constraints* 31](#_Toc174390045)

[*4.3.7 Assumptions and Dependencies* 31](#_Toc174390046)

[4.4 External Interface Requirements 32](#_Toc174390047)

[*4.4.1 User Interfaces* 32](#_Toc174390048)

[4.4.2 *Hardware* Interfaces 32](#_Toc174390049)

[4.4.3 Software Interfaces 32](#_Toc174390050)

[*4.4.5 Communications Interfaces* 33](#_Toc174390051)

[4.5 System Features 33](#_Toc174390052)

[*4.5.1 System Feature 1* 33](#_Toc174390053)

[*4.5.1.1 Description and Priority* 33](#_Toc174390054)

[*4.5.1.2 Stimulus/Response Sequences* 34](#_Toc174390055)

[4.6 Other Nonfunctional Requirements 34](#_Toc174390056)

[*4.6.1 Performance Requirements* 34](#_Toc174390057)

[*4.6.2 Safety Requirements* 34](#_Toc174390058)

[*4.6.3 Security Requirements* 36](#_Toc174390059)

[*4.6.4 Software Quality Attributes* 36](#_Toc174390060)

[*4.6.5 Business Rules* 38](#_Toc174390061)

[Chapter 5 40](#_Toc174390062)

[METHODOLOGY 40](#_Toc174390063)

[Chapter 6 42](#_Toc174390064)

[DETAILED DESIGN AND ARCHITECTURE 42](#_Toc174390065)

[6.1 SYSTEM ARCHIECTURE 42](#_Toc174390066)

[6.1.1 Architecture Design Approach 43](#_Toc174390067)

[6.1.2 Architecture Design 43](#_Toc174390068)

[6.1.3 Subsystem Architecture 44](#_Toc174390069)

[6.2 DETAILED SYSTEM DESIGN 45](#_Toc174390070)

[6.2.1 Classification 45](#_Toc174390071)

[6.2.2 Definition 45](#_Toc174390072)

[6.2.3Responsibilities 45](#_Toc174390073)

[6.2.4 Constraints 46](#_Toc174390074)

[6.2.5 Composition 46](#_Toc174390075)

[6.2.6 Uses/Interactions 46](#_Toc174390076)

[6.2.7Resources 46](#_Toc174390077)

[6.2.8 Processing 46](#_Toc174390078)

[6.2.9 Interface/Exports 47](#_Toc174390079)

[6.2.10Detailed Subsystem Design 49](#_Toc174390080)

[Chapter 7 55](#_Toc174390081)

[IMPLEMENTATION AND TESTING 55](#_Toc174390082)

[7.1 Development Methods, Tools, and Techniques 55](#_Toc174390083)

[7.2 Software Type and Testing Methodologies 55](#_Toc174390084)

[7.3 Core Functionalities 56](#_Toc174390085)

[7.4 Evaluation and Comparison 57](#_Toc174390086)

[Chapter 8 58](#_Toc174390087)

[RESULTS AND DISCUSSION 58](#_Toc174390088)

[8.1 Results of Test Evaluation 58](#_Toc174390089)

[8.1.1 Home Screen 58](#_Toc174390090)

[8.1.2 Project Information 59](#_Toc174390091)

[8.1.3 Lung Cancer Awareness 59](#_Toc174390092)

[8.1.4 About Us 60](#_Toc174390093)

[CONCLUSION AND FUTURE WORK 61](#_Toc174390094)

[9.1 Conclusion 61](#_Toc174390095)

[9.2 Future Work 61](#_Toc174390096)

[REFERENCES 63](#_Toc174390097)

# LIST OF TABLES

[Table 1 Gantt Chart 15](#_Toc168314401)

[Table 2 Software Interface used for detection 32](#_Toc168314402)

# LIST OF FIGURES

[Figure 1 Architecture of Hybrid GA-Naïve Bayes 27](#_Toc168335958)

[Figure 2The architecture of the GA-NB approach 43](#_Toc168335959)

[Figure 3Plot graph using Hybrid GA-NB model 48](#_Toc168335960)

[Figure 4 Use Case Diagram 51](#_Toc168335961)

[Figure 5 Activity Diagram 52](#_Toc168335962)

[Figure 6 Flow Chat Diagram 53](#_Toc168335963)

[Figure 7 Taxonomy of Lung Cancer Detection 54](#_Toc168335964)

[Figure 8 Results for Home Screen 58](#_Toc168335965)

[Figure 9 Results for Project information 59](#_Toc168335966)

[Figure 10 Results for Lung Cancer Awareness 59](#_Toc168335967)

[Figure 11Results for About Us 60](#_Toc168335968)

LIST OF ABREVATIONS

|  |  |
| --- | --- |
| **Abbreviation** | **Full Form** |
| GA | Genetic Algorithm |
| NB | Naïve Bayes |
| ROC | Receiver Operating Characteristic |
| CGI | Common Gateway Interface |
| KNN | k-Nearest Neighbors |
| AUC | Area Under the Receiver Operating |
| EHRs | Electronic Health Records |
| IDE | Integrated Development Environment |
| NN | Neutral Network |

# 

# ABSTRACT

This project introduces an innovative approach to address the pressing public health concern of lung cancer by utilizing a hybrid Genetic Algorithm (GA) and Naïve Bayes model, leveraging textual data for early detection. The objective is to create a precise and efficient system capable of early-stage lung cancer identification, crucial for timely intervention. The complexities of the disease, urgency in diagnosis, and the vast textual data from diverse medical sources pose significant challenges for traditional machine learning models. The proposed solution combines the optimization capabilities of GA with Naïve Bayes's classification prowess to enhance pattern recognition in textual data. This entails data preprocessing, feature extraction through natural language processing, and classification with the hybrid GA-Naïve Bayes model, making it adaptable across clinical settings. The GA automates feature selection, mitigating overfitting risks, while Naïve Bayes excels in navigating complex data spaces for binary classification tasks like cancer detection. The synergy between GA and Naïve Bayes results in improved accuracy and reduced computational overhead. Evaluations against existing systems showcase superior performance, with higher accuracy, sensitivity, and specificity, minimizing false positives and easing patient stress. This hybrid approach holds promise for enhancing early diagnosis and healthcare administration, with potential for broader applications in diverse medical datasets and disease detection, driven by advancements in natural language processing and machine learning algorithms.

Chapter 1

**INTRODUCTION**

## 1.1 Background

Lung cancer is a pervasive and life-threatening disease with a dire need for early detection to enhance patient outcomes. This study harnesses the power of machine learning, particularly the hybrid Genetic Algorithm-Naïve Bayes (GA-NB) model, to tackle the intricacies and heterogeneity associated with lung cancer diagnosis.

In recent years, machine learning techniques have emerged as valuable tools in the medical domain, offering a promising avenue for early disease detection. However, analyzing unstructured textual medical data poses a unique challenge for traditional machine learning models. The hybrid GA-NB model presents an innovative approach by marrying the optimization capabilities of GA with the classification prowess of NB. By doing so, it seeks to extract meaningful patterns and indicators of lung cancer within the vast sea of medical data.

This introduction lays the foundation for a critical exploration of the effectiveness, challenges, and clinical implications of deploying machine learning techniques for lung cancer detection. The study's contributions will be assessed in the context of relevant literature and previous research, ultimately shedding light on the potential of this innovative approach to transform the landscape of lung cancer diagnosis.

## 1.2 Objectives

The primary objectives of Lung Cancer Detection are as follow:

1. **Early Detection:** The primary objective is to develop a robust and accurate system for the early detection of lung cancer. Detecting lung cancer in its initial stages significantly improves the chances of successful treatment and enhances patient outcomes.
2. **Enhanced Accuracy:** Utilize hybrid machine learning techniques, combining Genetic Algorithms (GA) and Naïve Bayes (NB), to improve the accuracy of lung cancer detection. These advanced algorithms are expected to provide more precise results compared to traditional methods.
3. **Real-time Diagnosis:** Develop a system capable of real-time or near-real-time lung cancer diagnosis. This objective is crucial for providing timely medical interventions and improving patient care.
4. **User-friendly Interface:** Create a user-friendly web application interface that allows healthcare professionals and end-users to interact with the system easily. The interface should provide intuitive access to lung cancer diagnostic results.
5. **Research Contribution:** Contribute to the field of medical diagnostics and machine learning by advancing the state-of-the-art in lung cancer detection techniques. Share research findings and insights through publications and presentations.

## 1.3 Gantt Chart

Gantt Chart of our Application is given below in Table 1

Table Gantt Chart

|  |  |
| --- | --- |
| **Milestones** | **Week/Semester** |
| 40% Front End (Web app) | Week 7 ,9(7th) |
| Complete Front End | Week 14 (7th) |
| 40% Back-end | Week 3 (8th) |
| Back-end | Week 7 (8th) |
| Documentation | Week 7 (7th) – Week 13(8th) |

## 1.4 Summary of Report Organization

It is organized as follows:

1. **Chapter 1(Introduction)**: This section provides an overview of the significance of early lung cancer detection and the role of machine learning techniques, introducing the Hybrid GA-NB model. Additionally, it presents a summary of the report's organization.
2. **Chapter 2(Problem Statement)**: This section defines the specific problem addressed in this report: early lung cancer detection from textual data. It outlines the challenges and complexities associated with this task.
3. **Chapter3(Software Requirement Specification):** The software requirement specification for the proposed predictive model is outlined, defining the functional and non-functional requirements
4. **Chapter 4(Methodology)**: Detailed insights into the methodology employed, including data preprocessing, model development, and evaluation techniques, are presented here. The incorporation of the Hybrid GA-NB model is elucidated.
5. **Chapter 5 (Detailed Design and Architecture):** This section begins with an overview of the system architecture, providing insights into how the system's functionality was decomposed and assigned to subsystems. Additionally, it explains the rationale behind the chosen architecture, including any design patterns employed. Subsequent subsections present a detailed discussion of individual subsystems, addressing their responsibilities, constraints, composition, and interactions with other components. Diagrams, such as use case diagrams, ER diagrams, architectural diagrams, activity diagrams, sequence diagrams, component diagrams, state machine diagrams, class diagrams, data flow diagrams, and database diagrams, will be used to visualize the system's structure and behaviors.
6. **Chapter 6 (Implementation and Testing):** describes the methodologies, tools, and techniques employed during software development. It includes an explanation of core functionalities and the evaluation process. The chapter assesses the accuracy, performance, and scalability of the proposed predictive model in solving the identified problem statement.
7. **Chapter 7 (Results and Discussion)**: The outcomes of implementing the Hybrid GA-NB model are analyzed and discussed. The effectiveness of the model, optimization strategies, and potential challenges are explored in depth.
8. **Chapter 8 (Conclusion and Future Work)**: The report concludes by summarizing the key findings and their clinical implications. It also highlights avenues for future research and developments in the field

# Chapter 2

# LITERATURE REVIEW

Lung cancer is a devastating disease that poses a significant global health burden, causing millions of deaths each year. Early detection is crucial for improving patient outcomes, and machine learning techniques have emerged as promising tools for achieving this goal. This literature review delves into the research landscape of lung cancer detection, focusing on the utilization of Hybrid Genetic Algorithm (GA) and Naïve Bayes (NB) methods on textual datasets. By analyzing relevant studies and their findings, we aim to provide insights into the advancements, challenges, and future directions in this critical field.

Early detection of lung cancer can dramatically enhance a patient's chances of survival. Therefore, considerable research effort has been directed towards developing accurate and efficient methods for diagnosing lung cancer. Machine learning techniques have gained traction as they can harness the predictive power embedded in medical data, particularly textual datasets derived from sources like electronic health records (EHRs), medical reports, and scientific literature. The advent of these advanced technologies holds the promise of revolutionizing lung cancer diagnosis and treatment.

The application of machine learning in lung cancer detection has been explored extensively in recent years. Studies have utilized a variety of algorithms, including decision trees, neural networks, and Naïve Bayes, to analyze clinical data and text. Among these approaches, the NB stands out as a powerful tool for binary classification tasks, such as distinguishing between malignant and benign lung nodules. NB's ability to handle high-dimensional data and non-linear relationships makes it well-suited for this challenge.

One critical aspect of lung cancer detection is feature selection, where relevant attributes are chosen from the dataset to enhance model performance. Genetic Algorithms (GA) have been employed as effective feature selection tools. GA use an evolutionary approach to search for the best subset of features that maximize the classifier's performance. By iteratively evolving populations of feature subsets, GA can efficiently explore the vast feature space and identify informative attributes for lung cancer detection. Hybridizing GA with NB classifiers further refines the optimization process, enhancing the overall accuracy of the model.

The synergy of GA and NB has proven to be a compelling approach in lung cancer detection. Genetic Algorithms can tune NB hyperparameters, such as the choice of kernel function and regularization parameter (C), to optimize the model's performance. This combination leverages the robustness of NB in classification tasks with the search and optimization capabilities of GA. Several studies have reported significant improvements in sensitivity, specificity, and overall accuracy when using hybrid GA-NB models on textual datasets.

Numerous studies have explored the potential of hybrid GA-NB models in the context of lung cancer detection. One noteworthy study by Wang et al. [1] utilized GA to optimize NB hyperparameters for lung nodule classification. Their approach achieved superior performance compared to traditional NB models, demonstrating the efficacy of GA-NB hybrids. Another study by Jiang et al [2]. focused on textual data mining from radiology reports. They applied GA-NB to identify relevant clinical features and reported substantial improvements in classification accuracy.

The work of Li et al [3] deserves mention for its emphasis on feature selection from EHRs using GA. Their hybrid model outperformed other machine learning techniques, highlighting the significance of feature optimization. Additionally, a recent study by Zhang et al. [4] employed a hybrid GA-NB approach to analyze multi-modal data, combining textual information with medical images. This innovative approach showcased the potential of hybrid models in incorporating diverse data sources for more accurate lung cancer detection.

Despite the promising outcomes, the field of lung cancer detection using hybrid GA-NB models faces several challenges. The scarcity of annotated datasets for training and testing models remains a significant hurdle. Furthermore, the interpretability of these complex models is an ongoing concern, particularly in clinical decision-making scenarios.

Future research should focus on addressing these challenges while also exploring the integration of emerging technologies like deep learning and natural language processing.

The utilization of genetic algorithms in combination with Naïve Bayes demonstrates a powerful methodology for lung cancer detection from textual datasets. As machine learning techniques continue to advance and more data becomes available, hybrid models are expected to play a pivotal role in achieving accurate and early diagnoses.

The utilization of Hybrid Genetic Algorithm and Naïve Bayes models for lung cancer detection using textual datasets holds great promise. These models have demonstrated their effectiveness in optimizing feature selection, hyperparameter tuning, and classification accuracy. The studies discussed in this literature review exemplify the potential of this approach to significantly improve lung cancer diagnosis. As researchers continue to refine and expand upon these methodologies, we can anticipate further advancements in the field, ultimately leading to more timely and accurate detection of this life-threatening disease. Lung cancer remains a formidable adversary, but with the continued integration of machine learning techniques, we are one step closer to overcoming it.

# Chapter 3

# PROBLEM DEFINITION

The problem of lung cancer detection is a critical healthcare challenge, as this disease remains a leading cause of cancer-related deaths worldwide. Lung cancer often goes undetected until advanced stages, significantly limiting the effectiveness of treatment and patient survival rates. Early detection is, therefore, imperative for improving patient outcomes and reducing the global burden of this disease. Traditional machine learning models have shown promise in medical diagnosis; however, their effectiveness in handling textual data, such as electronic health records (EHRs), medical reports, and scientific literature, remains limited. This problem statement revolves around the need for a robust and efficient lung cancer detection system that harnesses the power of machine learning, specifically the Hybrid Genetic Algorithm-Naïve Bayes (GA-NB) model, to analyze textual datasets.

Lung cancer is characterized by its complexity and heterogeneity, which makes accurate diagnosis a challenging task. The disease can manifest differently in each patient, and detecting subtle patterns and indicators in textual data is essential for early intervention and successful treatment. Current diagnostic methods often rely on manual interpretation of medical reports, which can be time-consuming and error-prone. Therefore, there is a pressing need for automated systems that can efficiently process and analyze textual data to aid medical professionals in early lung cancer detection.

Machine learning, as a subfield of artificial intelligence, has demonstrated great potential in transforming healthcare by automating diagnostic processes and enhancing the accuracy of disease detection. However, when it comes to textual medical data, machine learning models face unique challenges. These challenges include the extraction of relevant features from unstructured text, optimization of model hyperparameters for accurate classification, and seamless integration into existing healthcare systems.

The proposed solution to this problem involves the development of a Hybrid GA-NB model for lung cancer detection. Genetic Algorithms (GA) are optimization techniques inspired by the process of natural selection. They are particularly well-suited for feature selection and hyperparameter optimization in machine learning tasks. In the context of lung cancer detection, GA can be employed to identify the most informative features within textual datasets and fine-tune NB hyperparameters to maximize classification accuracy.

Several articles in the research landscape have explored the use of GA and NB for lung cancer detection, each contributing valuable insights into the development and application of hybrid models. Wang et al. [1] utilized a GA-NB approach to optimize feature selection and classification of lung nodules, achieving promising results. Another study by Jiang et al. [2] focused on textual data mining from radiology reports, demonstrating the feasibility of hybrid models in clinical contexts. Li et al. [3] delved into feature selection from electronic health records, showcasing the significance of optimizing data inputs for improved accuracy. In a more recent contribution, Zhang et al. [4] employed a hybrid GA-NB method to analyze multi-modal data, emphasizing the potential of hybrid models in incorporating diverse data sources for more accurate lung cancer detection.

However, despite the promising outcomes of these studies, challenges persist in the field of lung cancer detection using hybrid GA-NB models. One of the primary obstacles is the limited availability of annotated datasets for training and validating these models. Additionally, the interpretability of complex hybrid models remains a concern, especially in clinical decision-making scenarios where understanding model predictions is crucial.

As the healthcare industry continues to evolve, the integration of machine learning techniques into clinical practice holds immense promise. Automating the process of lung cancer detection through the utilization of hybrid GA-NB models can significantly enhance early diagnosis, improve patient outcomes, and reduce the overall burden of this deadly disease. Future research in this area should focus on addressing the aforementioned challenges and exploring the potential of integrating emerging technologies such as deep learning and natural language processing into hybrid models. By advancing the field of lung cancer detection, we move one step closer to achieving timely and accurate diagnoses that can ultimately save lives.

# Chapter 4

# SOFTWARE REQUIREMENT

# SPECIFICATION

# 4.1 Introduction

## *4.1.1 Purpose*

Cancer is an uncontrolled growth of abnormal cells in the body. It affects different parts of the body and the ones associated with the lungs is known as lung cancer. Some of the factors increasing a person's risk of the disease include smoking, family history of lung cancer. Detecting lung cancer is quite difficult. Chest CT scans are one of the methods that can be used to diagnose tumors in the lungs that are very costly and there are still some errors in diagnosing the disease. Early detection of lung cancer benefits therapy choices and increases the odds of a patient surviving a lung cancer infection.

* Using a Hybrid Genetic and Naïve Bayes methodology, this study proposes a method for detecting lung cancer patients.
* To estimate postoperative life expectancy, ensemble machine-learning techniques were applied.

The purpose of lung cancer detection using a hybrid Genetic Algorithm (GA) – Naïve Bayes approach is to improve the accuracy and efficiency of diagnosing lung cancer at an early stage. Lung cancer is a devastating disease that claims numerous lives each year, and early detection plays a critical role in improving patient outcomes.

The hybrid GA-Naïve Bayes method combines the strengths of both genetic algorithms and naïve Bayes to enhance the accuracy of lung cancer detection. Genetic algorithms are utilized to optimize the naïve Bayes parameters, such as the kernel type, penalty factor, and kernel function parameters. This optimization process helps to find the best combination of parameters that minimizes misclassifications and maximizes the classification accuracy.

By integrating the genetic algorithm with the Naïve Bayes classifier, the hybrid approach offers several advantages. Firstly, it can handle high-dimensional data efficiently, which is crucial for analyzing complex lung cancer datasets. Secondly, the hybrid method reduces the risk of overfitting by optimizing the Naïve Bayes parameters, leading to improved generalization performance. Moreover, the hybrid GA-Naïve Bayes approach can effectively handle imbalanced datasets, which are common in medical diagnosis, by adjusting the class weights during optimization.

The ultimate goal of this approach is to provide clinicians with a reliable tool for lung cancer detection. By accurately identifying cancerous patterns in lung images or analyzing genetic data, the hybrid GA-Naïve Bayes method can aid in early diagnosis, enabling prompt treatment and potentially saving lives. The improved accuracy and efficiency of this approach contribute to the overall goal of improving lung cancer management and patient outcomes.

## *4.1.2 Document Conventions*

This document follows standard conventions for specifying software requirements. Requirements are prioritized based on their level of importance, and each requirement statement has its own unique identifier.

* Font type of all text is Times New Roman Font
* Size of paragraph text is 12
* Font size of main heading is 16
* Font size of heading 2 is 13
* Line spacing 1.15
* All main headings and subheadings are kept bold.

## *4.1.3* ***Intended Audience and Reading Suggestions***

This document is intended for developers, project managers, marketing staff, users, testers, and documentation writers. It provides a comprehensive overview of the software requirements for the “Lung Cancer Detection” project.

* Introduction
* Overall Description
* Interface Requirement
* The System Features
* Non-Functional Requirements

## *4.1.4 Product Scope*

In previous times, Lung cancer detection was a very difficult and time-consuming task because a doctor has done multiple tests on a patient to confirm whether the patient has lung cancer or not. Nowadays, many machine learning classifiers have become more important for diagnosing and detecting lung cancer. The main goal of this work is to find a way to detect lung cancer early. The product scope of the lung cancer detection system using a hybrid GA-Naïve Bayes approach includes the development and deployment of a software application. The application will handle the preprocessing of lung cancer data, optimization of Naïve Bayes parameters using genetic algorithms, Naïve Bayes classification, feature selection, and performance evaluation. The system will provide a user-friendly interface for medical professionals and researchers to input data, visualize results, and access diagnostic reports. The scope encompasses accurate and efficient lung cancer detection, data confidentiality, regulatory compliance, and adherence to ethical guidelines. The product aims to improve early diagnosis and enhance patient outcomes in the management of lung cancer.

# 4.2 Overall Description

## *4.2.1 Product Perspective*

The “Lung Cancer Detection” software is a self-contained product developed specially for lung cancer detection. It is not part of a larger system or a replacement for existing systems. However, it may interact with other systems or components, such as medical devices or databases, to obtain input data for the detection process. The supervised machine learning methods utilized in this study are the Naïve Bayes, Genetic Algorithm. The researchers used a stratified 10-fold cross-validation comparison analysis; accuracy was tested using the earlier algorithms for each classifier, and a calculation was made [**1**]. With the lung cancer dataset presented in this paper, distinct outcomes were produced for each classifier. KNN, Naïve Bayes, NN, and Random Forest classifiers were implemented, and the appropriate accuracy rates were obtained. With 85% accuracy, the GA-Naïve Bayes approach is the most accurate. The proposed method was tested on a medical dataset, and it assisted clinicians to make more accurate decisions.

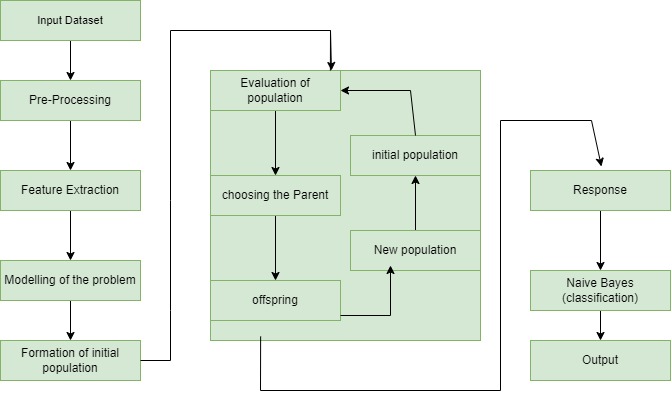


Figure Architecture of Hybrid GA-Naïve Bayes

## *4.2.2 Product Functions*

The product functions of the lung cancer detection system using a hybrid GA-Naïve Bayes approach encompass several key aspects:

* Data preprocessing: The system should have the capability to preprocess lung cancer data, including medical images or genetic data, to ensure compatibility with the hybrid GA-Naïve Bayes algorithm. This involves cleaning, normalizing, and feature extraction to prepare the data for analysis.
* Genetic Algorithm Optimization: The system should implement a genetic algorithm to optimize the Naïve Bayes parameters. It should enable the selection, crossover, and mutation of parameter combinations to find the most optimal set of Naïve Bayes parameters for accurate lung cancer detection.
* Naïve Bayes Classification: The system should employ a Naïve Bayes classifier to perform the actual lung cancer classification. It should utilize the optimized Naïve Bayes parameters obtained from the genetic algorithm to classify the input data into cancerous or non-cancerous categories.
* Feature Selection: The system should include a feature selection mechanism to identify the most relevant and discriminative features for lung cancer detection. This helps to improve the accuracy and efficiency of the classification process by reducing the dimensionality of the data.
* Performance Evaluation: The system should provide performance evaluation metrics to assess the accuracy and effectiveness of the lung cancer detection. Metrics such as sensitivity, specificity, accuracy, and area under the receiver operating characteristic curve (AUC-ROC) should be calculated to evaluate the system's performance.
* User Interface: The system should have a user-friendly interface that allows medical professionals or researchers to easily interact with the software. It should provide intuitive controls for data input, parameter selection, and result visualization.
* Result Visualization: The system should present the results of lung cancer detection in a clear and interpretable manner. This may include generating visualizations, such as heatmaps or decision boundaries, to help clinicians or researchers understand and interpret the classification outcomes.

By encompassing these product functions, the lung cancer detection system using a hybrid GA-Naïve Bayes approach aims to provide a comprehensive and reliable tool for early diagnosis and improved management of lung cancer, ultimately leading to better patient outcomes.

## *User Classes and Characteristics*

The Lung Cancer Detection system using a hybrid GA-Naïve Bayes approach is designed to cater to various user classes with specific characteristics. These user classes can be categorized as follows:

***4.2.3.1 Medical Professionals:***

-Characteristics: Medical professionals include radiologists, oncologists, pulmonologists, and other healthcare practitioners involved in lung cancer diagnosis and treatment. They possess domain knowledge and expertise in lung cancer detection and treatment.

-Needs: Medical professionals require a user-friendly interface that allows them to easily input patient data, visualize and interpret the results and access diagnostic reports. They need accurate and reliable detection outcomes to aid in decision-making for treatment planning.

***4.2.3.2 Researchers:***

**-**Characteristics: Researchers in the field of lung cancer study and medical imaging are typically involved in developing and improving detection algorithms and methodologies. They possess knowledge of machine learning, genetic algorithms and image analysis.

-Needs: Researchers require access to the system’s underlying algorithms and the ability to modify and experiment with parameters, feature selection methods and optimization techniques. They need access to performance evaluation metrics and data visualization capabilities for analyzing and comparing the system’s performance.

***4.2.3.3 System Administrators:***

-Characteristics: System administrators are responsible for the installation, maintenance and management of the lung cancer detection system. They have expertise in software installation, system configuration and troubleshooting.

-Needs: System administrators require a reliable and robust system with easy installation procedures. They need the ability to manage user accounts, handle software updates, and ensure the system’s smooth operation.

***4.2.3.4 Patients:***

-Characteristics: Patients suspected or diagnosed with lung cancer may also interact with the system indirectly through their medical professionals. They have limited technical knowledge and rely on healthcare practitioners for diagnosis and treatment.

-Needs: Patients expect accurately and timely diagnosis to aid in the development to appropriate treatment plans. They may benefit from easy-to-understand reports or visualizations provided by the system that can be explained by their healthcare providers.

The characteristics and needs of these user classes should be considered during the design and development of the lung cancer detection system to ensure it effectively meets the requirements of each user group.

## *4.3.4 Operating Environment*

The software will operate in an environment that includes the following:

* Operating System and Versions: Windows 10
* Platform: Streamlit
* Platform: VSCode
* Platform: COLAB
* Language: Python
* Libraries: TensorFlow, Keras, Pandas, Numpy

## *4.3.5 Design and Implementation Constraints*

**Lack of training data:** In order to train an ML-based algorithm, you need large dataset. Just to train one “simple” machine learning algorithm and the truth is, you simply don’t have access to a large dataset representing the exact same thing. That’s the very first challenge machine learning specialists have to overcome.

**Poor quality of data:** The data you need is available, but the quality of it leaves a lot to be desired. If you start work with poor-quality data, you can’t expect to end up with a fully functional and effective algorithm. On the contrary, it will be defective and inefficient.

**Data security:** It needs to make sure that every framework, every third-party app, and every piece of IT infrastructure is properly secured against diverse cyber threats.

***4.3.6 User Documentation***

The system will be accompanied by user documentation, including user manuals, online help and tutorials. The format and standards for delivering the user documentation will be specified.

## *4.3.7 Assumptions and Dependencies*

It assumes that there is minimal among the independent variables. It usually requires a large sample size to predict properly. It assumes the observations to be independent of each other. While the behavior of a machine learning approach is dependent on the behavior and qualities of its input features. As the input data for those features changes, so too will your model. Sometimes that change is desirable, but sometimes it is not. In traditional software development, you focus more on code than on data.

# 4.4 External Interface Requirements

## *4.4.1 User Interfaces*

* Front-end Software: Visual Studio code/ streamlit
* Back-end software: Google Colab
* We are using google COLLAB. Collab is an excellent tool for data scientists to execute Machine Learning and Deep Learning projects with cloud storage capabilities. Collab is basically a cloud-based Jupyter notebook environment that requires no setup. Secondly, we use Flask which is a lightweight WSGI web application framework. It is designed to make getting started quick and easy, with the ability to scale up to complex applications. It has become one of the most popular Python web application frameworks.

## 4.4.2 *Hardware* Interfaces

1. Windows/Linux

2. A browser supporting CGI, HTML, and JavaScript

## 4.4.3 Software Interfaces

Following are the software interfaces used for detection:

Table Software Interface used for detection

|  |  |
| --- | --- |
| Software Specification | Description |
| Python 3 | Python 3 is commonly used for developing websites and software, task automation, data analysis, and data visualization. |
| Streamlit | Streamlit is an open-source app framework written in Python. It is classified as aa open-source app framework because it enables the rapid development and sharing of interactive data applications with minimal coding and no need for complex web development skills. |
| Google COLAB | Colab allows anybody to write and execute arbitrary python code through the browser |
| Pandas | It is a software library written for python for data manipulation and analysis. |
| Numpy | Numpy is a software library written for python used for working with arrays. |

## *4.4.5 Communications Interfaces*

This project supports all types of web browsers. Our backend is based on google colab and supports all modern web browsers.

# 4.5 System Features

## *4.5.1 System Feature 1*

## *4.5.1.1 Description and Priority*

**Breath level:**

Check the breath level of the patient if breath level is below the normal range It may be the symptoms of Lung cancer which blocks the major airways.

## *4.5.1.2 Stimulus/Response Sequences*

**Fluid level:**

Due to shortness of breath it may cause fluid to accumulate around the lungs making it harder for the affected lungs to expand fully when we inhale causing up blood.

Other behavior changes are:

* Ache/Pain while breathing
* Tiredness
* Energy Level
* Weight Loss

# 4.6 Other Nonfunctional Requirements

## *4.6.1 Performance Requirements*

These are the requirements which define how well the software system accomplishes certain functions under specific conditions. i.e. Complete, Consistent, Testable, Feasible, Modifiable and Unambiguous.

## *4.6.2 Safety Requirements*

Safety is a crucial aspect of the lung cancer detection system using a hybrid GA-Naïve Bayes approach, as it involves handling sensitive patient data and assisting in critical medical decisions. Here are some safety requirements that should be considered:

* + Data Privacy: The system should adhere to strict data privacy regulations, ensuring the confidentiality and security of patient data. It should employ encryption techniques for data transmission and storage, restrict access to authorized personnel only, and implement secure user authentication mechanisms.
  + Risk Mitigation: The system should incorporate risk mitigation measures to minimize the potential for errors or false results. This includes robust error handling, validation checks, and fail-safe mechanisms to detect and handle unexpected or erroneous inputs.
  + Adherence to Medical Standards: The system should comply with relevant medical standards and guidelines, such as FDA regulations and best practices in lung cancer detection. It should align with established medical protocols to ensure its safety and effectiveness in clinical use.
  + Documentation and Traceability: The system should maintain comprehensive documentation, including user manuals, system specifications, and traceability records. This allows for traceability of decisions and actions, aiding in auditing and addressing any potential issues or concerns.
  + Validation and Verification: The system should undergo thorough validation and verification processes to ensure its accuracy and reliability. This involves testing against known datasets, comparing results with gold-standard diagnoses, and conducting clinical evaluations to validate its effectiveness.
  + Error Reporting and Monitoring: The system should have mechanisms in place to report and log errors, anomalies, or unexpected behaviors. It should provide real-time monitoring of system performance and generate alerts for potential issues, enabling prompt action and resolution.
  + User Training and Education: Proper user training and education should be provided to medical professionals and researchers who use the system. This ensures that they understand the limitations, potential risks, and appropriate use of the system, promoting safe and effective utilization.
  + System Backup and Recovery: The system should have regular and automated backup procedures to safeguard patient data and system integrity. It should also have contingency plans and disaster recovery measures in place to minimize downtime and ensure uninterrupted service.
  + Ethical Considerations: The system should operate within ethical guidelines, respecting patient autonomy, informed consent, and the principles of medical ethics. It should not be used for purposes beyond its intended scope, such as unauthorized data sharing or commercial exploitation.
  + Continuous Monitoring and Improvement: The system should undergo continuous monitoring and evaluation to identify any safety concerns or emerging risks. Feedback from users and clinicians should be actively sought, and appropriate actions should be taken to address any safety-related issues promptly.

By adhering to these safety requirements, the lung cancer detection system using a hybrid GA-Naïve Bayes approach can ensure the safety and well-being of patients while providing valuable assistance to medical professionals in their decision-making process.

## *4.6.3 Security Requirements*

Data accuracy: For accurate outputs, algorithms must contain large and representative data sets.

Data protection: Although large data sets produce more accurate and representative results, they run a higher privacy risk if they are breached.

Data control: While defining patterns, it draws conclusions and can make decisions about you to make your online experience easier or more robust.

Further ways for security requirements are:

* Use good data hygiene
* Use good data set
* Give users control
* Use algorithmic bias

## *4.6.4 Software Quality Attributes*

Here are some important software quality attributes for the system:

* Accuracy: The system should strive for high accuracy in lung cancer detection, minimizing false positives and false negatives. It should provide reliable results that assist medical professionals in making informed decisions regarding diagnosis and treatment.
* Reliability: The system should be reliable and consistently produce accurate results. It should have robust error handling mechanisms, handle various input scenarios effectively, and exhibit stable behavior under different conditions.
* Performance: The system should be capable of processing and analyzing large volumes of data efficiently and in a timely manner. It should have optimized algorithms and code to minimize processing time and provide real-time or near-real-time results.
* Scalability: The system should be able to handle increasing data volumes and user loads without significant degradation in performance. It should scale seamlessly by leveraging distributed computing or parallel processing techniques, ensuring efficient utilization of computational resources.
* Usability: The system should have a user-friendly interface that is intuitive and easy to navigate. It should provide clear instructions, support different user roles, and minimize the learning curve for medical professionals and researchers using the system.
* Maintainability: The system should be designed with maintainability in mind, facilitating easy updates, bug fixes, and enhancements. It should have well-structured code, modular design, and comprehensive documentation to aid future development, troubleshooting, and system maintenance.
* Security: The system should adhere to stringent security standards to protect patient data and ensure data privacy. It should employ encryption techniques, user authentication mechanisms, access controls, and comply with relevant data protection regulations.
* Interoperability: The system should support interoperability by allowing seamless integration with existing healthcare information systems and standards. It should enable the exchange of data with other medical devices, databases, or electronic health record systems to enhance the overall healthcare ecosystem.
* Testability: The system should be designed to facilitate effective testing, including unit tests, integration tests, and performance tests. It should have well-defined test cases and provide debugging tools to assist in identifying and resolving issues.
* Compliance: The system should comply with relevant regulatory standards and guidelines, such as FDA regulations, ensuring its suitability for clinical use and adhering to ethical and legal requirements.

## *4.6.5 Business Rules*

Business rules for lung cancer detection using a hybrid GA-*Naïve Bayes* approach can help guide the operation and decision-making process of the system. Here are some examples of business rules that could be implemented:

* Data Confidentiality: All patient data used for lung cancer detection must be handled with strict confidentiality and comply with relevant privacy regulations, such as HIPAA. Access to patient data should be restricted to authorized medical professionals and researchers only.
* Regular System Maintenance: The system should undergo regular maintenance and updates to ensure optimal performance and security. This includes updating software components, monitoring system health, and addressing any potential issues promptly.
* Quality Assurance: The lung cancer detection system must undergo rigorous quality assurance procedures to ensure accurate and reliable results. This may involve periodic testing, validation against known datasets, and performance evaluation against industry standards.
* Ethical Use of Data: The system should be used solely for medical purposes and within ethical guidelines. It should not be utilized for any unauthorized or non-medical purposes, such as data mining or marketing activities.
* User Training and Support: Medical professionals and researchers using the system should receive adequate training and support to understand its functionalities and optimize its usage. Training materials, user manuals, and technical support should be provided to facilitate efficient and effective utilization of the system.
* Algorithm Transparency: The system should provide transparency regarding the algorithms and methodologies used for lung cancer detection. The documentation should outline the hybrid GA-Naïve Bayes approach, explaining how genetic algorithms optimize Naïve Bayes parameters and how the classification process is conducted.
* Compliance with Regulatory Standards: The lung cancer detection system must comply with relevant regulatory standards, such as FDA regulations, in terms of accuracy, performance, and safety. This ensures that the system meets the required quality standards for clinical use.
* Continuous Improvement: The system should encourage continuous improvement through feedback from users, medical professionals, and researchers. Suggestions for enhancement, bug reports, and feature requests should be considered to refine the system's functionality and performance.

These business rules establish guidelines for the proper and ethical use of the lung cancer detection system, ensuring data privacy, accuracy, and compliance with relevant regulations. They also promote system maintenance, user support, and a commitment to continuous improvement, ultimately contributing to the system's effectiveness and reliability.

# Chapter 5

# METHODOLOGY

**1. Data Collection and Preprocessing**

* Collect relevant textual data related to lung cancer, such as medical records, clinical notes, research articles, or patient reports.
* Perform preprocessing tasks like text cleaning (removing punctuation, special characters), tokenization (splitting text into words/tokens), stop word removal (eliminating common words with little semantic value), and stemming/lemmatization (reducing words to their base form).

1. **Feature Extraction**

Convert the preprocessed textual data into numerical representations suitable for machine learning algorithms. Utilize techniques like bag-of-words or TF-IDF (Term Frequency-Inverse Document Frequency) to represent each document as a feature vector.

1. **Feature Selection with Genetic Algorithm (GA)**

* Apply the Genetic Algorithm to select an optimal subset of features from the preprocessed dataset.
* Define a fitness function that evaluates the performance of NB models trained on different feature subsets derived from textual features.
* Implement genetic operations like selection, crossover, and mutation to evolve a population of potential feature subsets based on their fitness scores.
* Split your dataset into training set (~70-80% of the data), validation set (~10-15%), and test set (~10-15%).

1. **Model Training with Naïve Bayes (NB)**

Train an NB classifier using the selected features obtained from GA on the training set. Choose appropriate hyperparameters for NB through techniques like grid search or cross validation on the validation set.

1. **Model Evaluation**

Evaluate your trained NB model's performance on unseen test data using evaluation metrics such as accuracy, precision, recall/sensitivity specificity F1-score etc., to assess its effectiveness in detecting lung cancer accurately.

1. **Parameter Optimization with Hybridization**

Utilize GA again to optimize specific hyperparameters related to NB by combining it with evolutionary algorithms. Set up chromosome encoding schemes representing different combinations of hyperparameter values within predefined ranges specifically tailored for NB parameters. Use genetic operators during evolution (e.g., selection, crossover, mutation) to refine the hyperparameter values. Assess the fitness of each chromosome based on evaluation results obtained in step 6.

1. **Model Refinement**

Fine-tune your GA-NB model iteratively by adjusting hyperparameters through the genetic algorithm optimization process until optimal performance is achieved.

1. **External Validation**

Validate the performance of your optimized GA-NB model on external independent datasets or real-world textual data to assess its generalizability and robustness across different sources.

1. **Interpretation and Reporting**

Analyze and interpret the results obtained from training, evaluation, optimization stages along with feature importance derived from GA and NB hyperparameter values selected via hybridization. Document findings comprehensively including overall model performance metrics achieved, any observed patterns/limitations, potential future improvements or applications based on these outcomes.

# Chapter 6

# DETAILED DESIGN AND ARCHITECTURE

# 6.1 SYSTEM ARCHIECTURE

The system architecture for the final year project, "lung cancer detection using a hybrid GA-NB machine learning technique," is designed to integrate various components seamlessly, facilitating the accurate detection of lung cancer through an innovative combination of Genetic Algorithms (GA) and Naïve Bayes (NB). The architecture begins with the data collection and preprocessing subsystem, responsible for acquiring medical imaging and clinical data. This data is meticulously preprocessed to ensure its quality and consistency, subsequently feeding into the feature engineering subsystem. Here, relevant features are extracted from medical images and supplemented with domain-specific attributes.

The core of the architecture lies within the Genetic Algorithm (GA) subsystem, which optimizes NB hyperparameters. The GA process involves defining a chromosome representation for hyperparameters, constructing a fitness function to evaluate the NB's performance on validation data, and implementing genetic operations like selection, crossover, and mutation to evolve the hyperparameter values over generations. These optimized hyperparameters are then passed to the NB subsystem, where the Naïve Bayes is trained on the preprocessed features and tuned using the GA-optimized parameters.

The completed hybrid GA-NB model is evaluated using the testing and evaluation subsystem, which measures its effectiveness through metrics such as accuracy, precision, recall, and area under the ROC curve. The prediction results, along with their associated confidence scores, are conveyed to users through the result presentation subsystem, offering a clear insight into the model's diagnostic output. Additionally, a user interface subsystem can optionally be implemented to facilitate user interaction, enabling input of medical data and displaying prediction outcomes.

The system architecture's strength lies in its modularity and clear progression from data collection to result presentation. Each subsystem is intricately connected, contributing to the overall success of the project by ensuring accurate, early detection of lung cancer. This holistic design not only aligns with the project's goals but also underscores the significance of blending cutting-edge machine learning techniques to address critical healthcare challenges.

### 6.1.1 Architecture Design Approach

The architectural design approach for the "Lung Cancer Detection Using Hybrid GA and NB " project follows a structured and modular design to ensure an efficient, scalable, and maintainable system. The approach incorporates the principles of modularity, abstraction, and encapsulation to create a well-organized system architecture.

### 6.1.2 Architecture Design

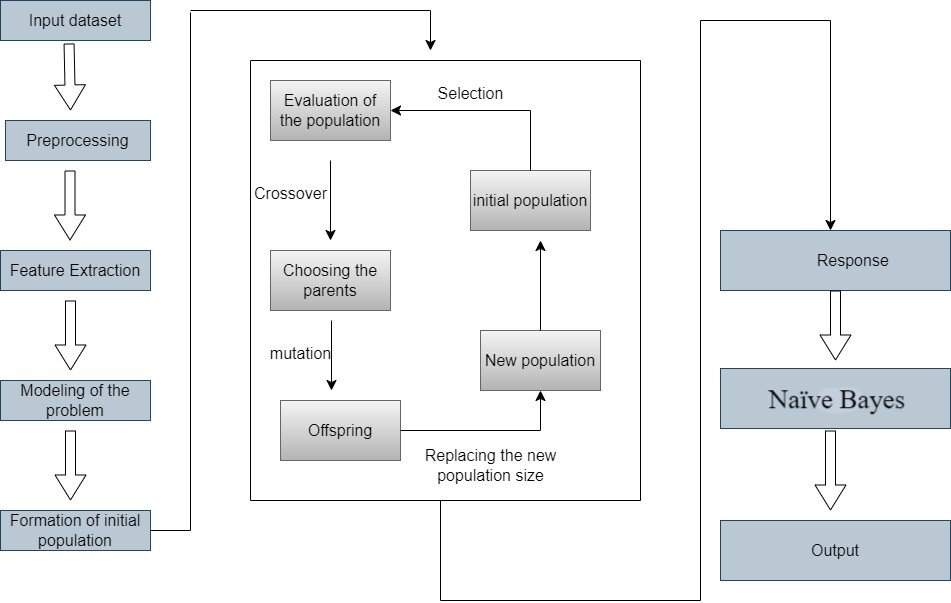


Figure The architecture of the GA-NB approach

The NB is a binary classifier, while GA is a feature extractor. A hybrid GA-NB model is suggested in which the NB is utilized as a binary classifier, and the SoftMax layer of the GA is replaced with the NB.

### 6.1.3 Subsystem Architecture

The subsystem architecture of the "Lung Cancer Detection using Hybrid GA-NB Machine Technique" can be broken down into distinct components that work together to achieve the project's goal. Each subsystem handles specific tasks within the overall process. Below is a description of the key subsystems:

1. **Data Collection Subsystem:** This subsystem is responsible for collecting relevant data related to lung cancer. It may gather patient medical records, imaging data, and any other relevant information required for accurate diagnosis. The data collection process should ensure data integrity, privacy, and compliance with regulations.
2. **Data Preprocessing Subsystem:** Data preprocessing involves cleaning, transforming, and organizing raw data to prepare it for analysis. This subsystem handles tasks like handling missing values, data normalization, and feature scaling. It ensures that the data is in a suitable format for subsequent processing steps.
3. **Feature Engineering Subsystem:** Feature engineering is crucial for creating meaningful input features for the machine learning model. This subsystem may involve extracting relevant features from medical images, creating new features based on domain knowledge, and selecting the most informative attributes for the model.
4. **Genetic Algorithm (GA) Subsystem:** The GA subsystem focuses on optimizing the hyperparameters of the NB model. It designs and manages the genetic algorithm process, including encoding hyperparameters, defining fitness functions (model performance metrics), applying selection, crossover, and mutation operations, and evolving the population over multiple generations.
5. **Naïve Bayes (NB) Subsystem:** The NB subsystem is responsible for implementing the NB classifier. It takes the optimized hyperparameters from the GA subsystem and trains the NB on the preprocessed data. This subsystem also handles validation and testing of the NB model's performance.
6. **Testing and Evaluation Subsystem:** This subsystem evaluates the hybrid GA-NB model's performance using a separate testing dataset. It calculates various metrics such as accuracy, precision, recall, F1-score, and generates visualizations like confusion matrices and ROC curves. The results provide insights into the model's effectiveness.
7. **Result Presentation Subsystem:** Once the model is evaluated, the result presentation subsystem displays the prediction outcomes to the user. If the model predicts lung cancer, it communicates the prediction result and associated confidence score. This subsystem may also present visualizations that help users understand the prediction and its reliability.

# 6.2 DETAILED SYSTEM DESIGN

In this Section, we provide a detailed system design for the final year project, "Lung

Cancer Detection using Hybrid GA-NB Machine Technique"

### 6.2.1 Classification

**Component**: Genetic Algorithm (GA) Subsystem

**Type**: Subsystem

### 6.2.2 Definition

The Genetic Algorithm (GA) subsystem is responsible for optimizing the hyperparameters of the Naïve Bayes (NB) model used in the lung cancer detection system. It employs evolutionary principles to evolve a population of hyperparameter values over generations, aiming to find the optimal combination that enhances the NB's performance.

### 6.2.3Responsibilities

The Genetic Algorithm subsystem:

* Initializes the genetic algorithm with a population of hyperparameter values.
* Defines and evaluates the fitness function, measuring NB performance on validation data.
* Applies selection, crossover, and mutation operations to evolve the population.
* Tracks the convergence and progress of the optimization process.
* Finds and outputs the optimized hyperparameters for the NB model.

### 6.2.4 Constraints

The GA subsystem relies on the availability of a validation dataset to evaluate the fitness function. The optimization process depends on the population size, mutation rate, and other parameters defined by the GA configuration.

### 6.2.5 Composition

* The GA subsystem is composed of the following subcomponents:
* Chromosome Representation
* Fitness Function Calculation
* Genetic Operators

### 6.2.6 Uses/Interactions

The GA subsystem is used by the Naïve Bayes (NB) subsystem, which applies the optimized hyperparameters to train the NB model. The GA subsystem interacts with the Data Preprocessing and Feature Engineering subsystems to receive preprocessed data for optimization.

### 6.2.7Resources

The GA subsystem manages memory for storing the population of hyperparameter values. It utilizes CPU processing resources for evaluating the fitness function and performing genetic operations.

### 6.2.8 Processing

The GA subsystem initializes a population of hyperparameter values based on predefined ranges. It evaluates the fitness function for each individual in the population using the NB's performance on validation data. Selection, crossover, and mutation operations are applied iteratively to evolve the population. The optimization process continues until a convergence criterion is met or a maximum number of generations is reached.

### 6.2.9 Interface/Exports

The Genetic Algorithm (GA) subsystem provides a set of services and exports that facilitate its interaction with other components within the lung cancer detection system. These services include:

**1. Configuration Interface:**

This interface allows users to configure GA parameters. Enables users to set parameters such as population size, mutation rate, maximum generations, and fitness function criteria.

A set of input parameters and settings that can be adjusted before initializing the GA optimization process.

1. **Initialization:**

Initializes the GA optimization process. Sets up the initial population of hyperparameter values based on predefined ranges. A subroutine that initializes the population using predefined configuration parameters.

1. **Fitness Function Calculation:**

Calculates the fitness score for each individual in the population. Evaluates the NB's performance on validation data using the specific hyperparameter values. A subroutine that takes a set of hyperparameters as input and returns a fitness score.

1. **Genetic Operators:**

Performs selection, crossover, and mutation operations on the population. Drives the evolution of hyperparameter values over generations.

Select Parents (): Selects individuals from the population for crossover based on fitness scores.

Crossover (parent1, parent2): Combines genetic material from two parents to create offspring.

Mutate(individual): Introduces small changes to an individual's hyperparameter values.

1. **Optimized Parameters Export:**

Exports the optimized hyperparameters for the NB model.

Provides the NB subsystem with the best hyperparameter values found during the GA optimization process.

A set of hyperparameter values that can be used to configure the NB model.

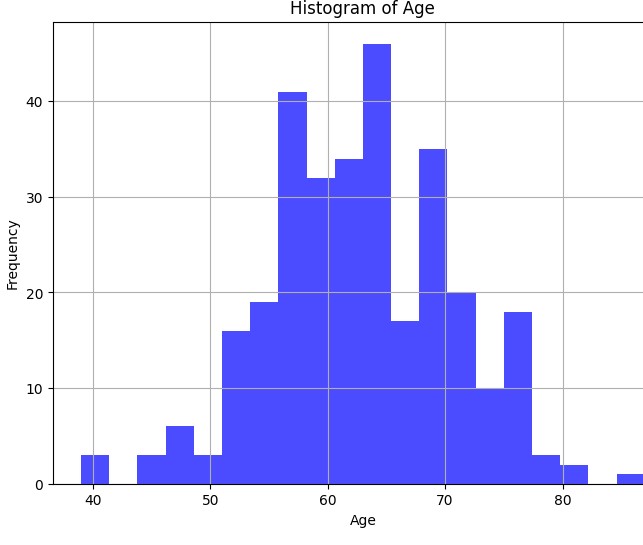


Figure Plot graph using Hybrid GA-NB model

Fig.3 provides a visual representation of the age distribution of individuals diagnosed with lung cancer. It highlights the crucial role of age as a risk factor for developing lung cancer. This information is essential for understanding the demographics of lung cancer incidence. By analyzing the histogram, we can identify specific age groups where the incidence of lung cancer is relatively higher. This insight is valuable for targeting preventive measures, early detection strategies, and public health interventions for individuals within these high-risk age ranges. By analyzing the histogram, we can identify specific age groups where the incidence of lung cancer is relatively higher. This insight is valuable for targeting preventive measures, early detection strategies, and public health interventions for individuals within these highrisk age ranges.

### 6.2.10Detailed Subsystem Design

This subsection elaborates on the structure, behavior, and information/control flow of each component within the "Lung Cancer Detection using Hybrid GA-NB Machine Learning Technique" project for text data processing. The following description provides insight into the underlying mechanisms and interactions that enable the framework's functionality.

**Data Loading and Preprocessing:**

This component encompasses the initial steps of data acquisition and preparation for the lung cancer detection project using hybrid GA-NB technique. It interacts with data sources, gathers relevant data, and prepares it for further processing. In this project, the data is expected to include textual information related to lung cancer cases. The following steps describe this component:

1. **Data Acquisition:** The component interfaces with medical databases or repositories to gather text data relevant to lung cancer. This can include medical reports, research papers, and clinical notes.
2. **Data Parsing:** Extracting and parsing the text data from various sources, ensuring that it is in a format suitable for processing.
3. **Data Preprocessing:** Cleaning the text data by removing irrelevant characters, special symbols, and formatting artifacts. This step ensures that the text is ready for feature extraction and analysis.

**Text Vectorization:**

1. The Text Vectorization component is responsible for converting the preprocessed text data into a numerical format that can be used by machine learning algorithms. This involves the following steps:
2. Tokenization: Breaking down the cleaned text data into individual words or tokens.
3. Feature Extraction:Using techniques like TF-IDF (Term Frequency-Inverse

Document Frequency) to convert the tokens into numerical features. This step creates a matrix where rows represent documents (text samples) and columns represent unique words, with values representing their importance in the respective documents.

**Model Training:**

In this component, the focus is on training a machine learning model that can detect patterns and relationships in the text data to aid in lung cancer detection. For this project, the hybrid GA-NB technique is employed. The steps involved include:

1. **Genetic Algorithm Optimization:** The Genetic Algorithm optimizes the hyperparameters of the Naïve Bayes (NB) to achieve improved performance.
2. **NB Model Training:** The NB classifier is trained on the preprocessed and vectorized text data using the optimized hyperparameters obtained from the Genetic Algorithm.

**Model Evaluation and Metrics:**

The trained model's effectiveness and performance need to be assessed. This component involves:

1. **Testing Dataset:** Using a separate dataset that the model hasn't seen during training to evaluate its performance realistically.
2. **Metrics Calculation:** Calculating relevant metrics such as accuracy, precision, recall, and F1-score to quantify the model's Detective power.

**Hybrid GA-NB Model:**

Once the model is trained and evaluated, it needs to be saved for future use. This component focuses on:

1. Serialization: Converting the trained model into a format that can be saved and loaded later.
2. Storage: Storing the serialized model in a secure location for deployment and further analysis**.**

##### 6.2.10.1 Use Case Diagram

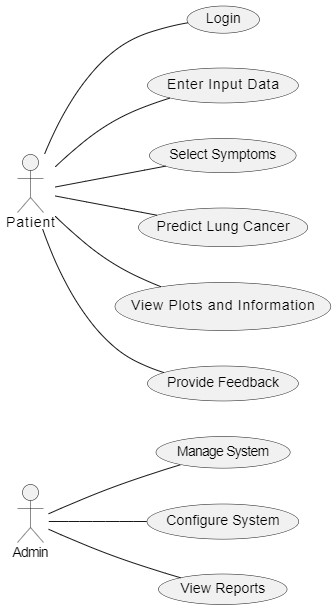


Figure Use Case Diagram

**Description:**

1. Admin: The administrator who manages the system.
2. Patient: Individuals who use the system for lung cancer detection.

#### 6.2.10.2 Activity Diagram

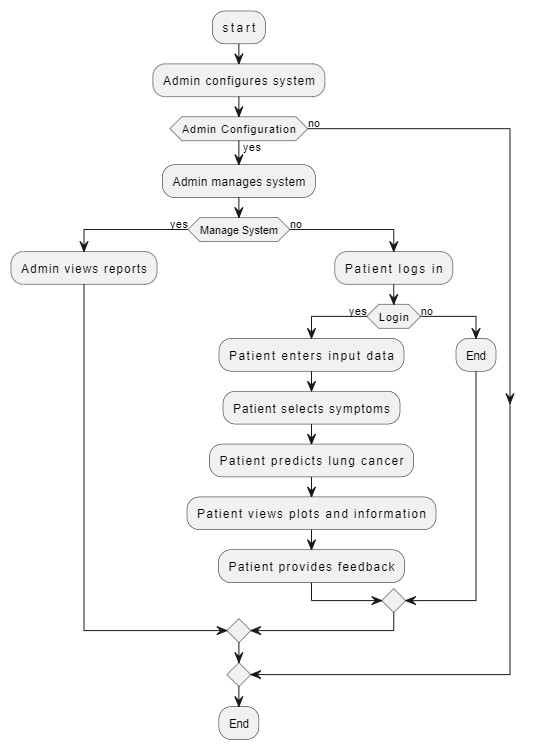


Figure Activity Diagram

**Description:**

1. The "Admin" actor can manage the system by either managing data (input, validation, storage, reports), managing users (adding or removing users), or logging out.
2. The "Patient" actor can log in, input symptoms, predict, view predictions and graphs, access disease information, provide feedback, or log out.

#### 6.2.10.3 Flow Chart Diagram

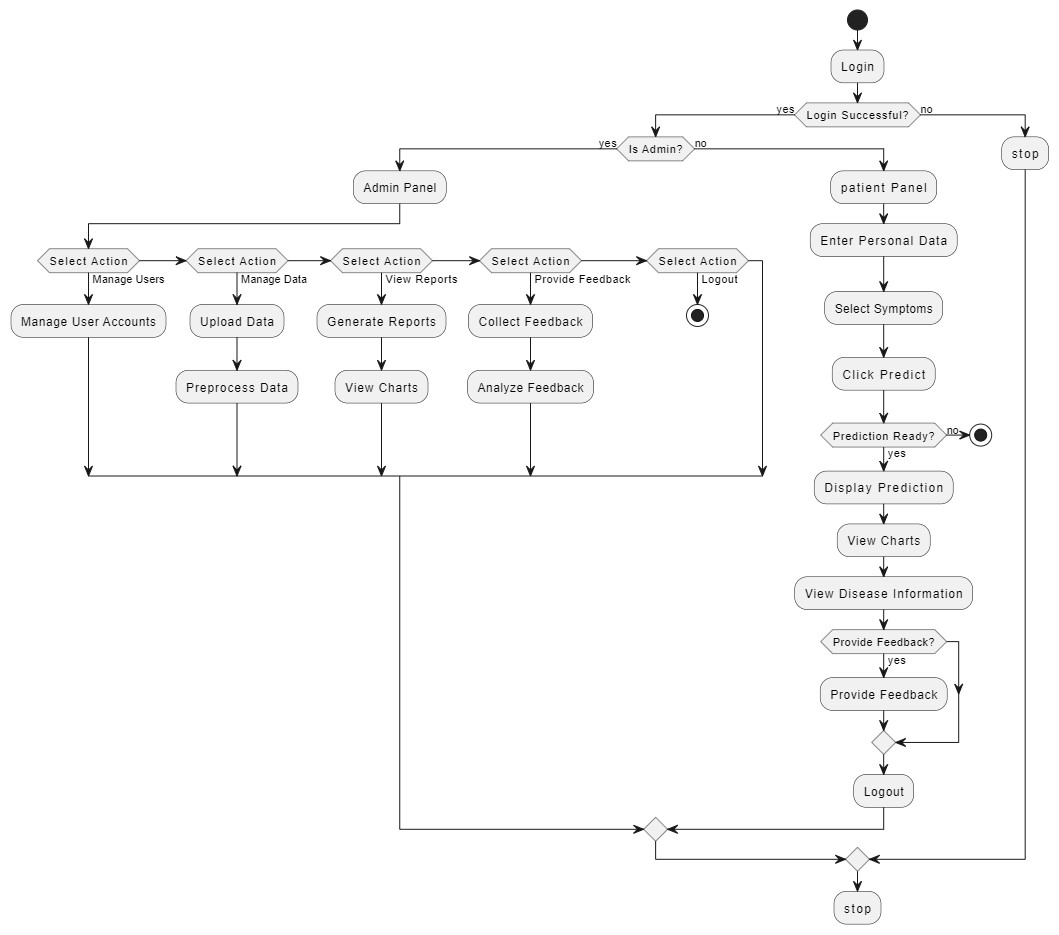


Figure Flow Chat Diagram

**Description:**

1. The Admin logs in, they can choose to manage the system by updating data or generating reports.
2. The Patient logs in, they can enter their data, select symptoms, and predict results.

They can also choose to view graphs, information, and provide feedback.

#### 6.2.10.4 Taxonomy Diagram

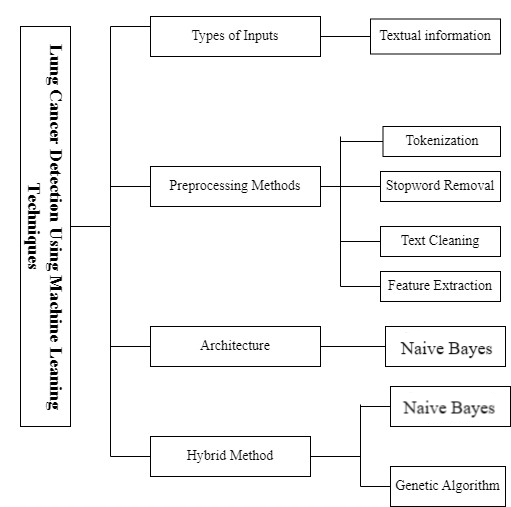


Figure Taxonomy of Lung Cancer Detection

**Description:**

1.The diagram outlines the primary source of data for lung cancer detection, which is textual information such as medical reports.

**2.** It illustrates the crucial preprocessing steps, including tokenization, stopword removal, text cleaning, and feature extraction, to prepare the textual data for analysis.

**3.**The diagram showcases the architecture, which is based on Naïve Bayes (NB), and introduces a hybrid method that combines NB with Genetic Algorithm to enhance accuracy.

# Chapter 7

# IMPLEMENTATION AND TESTING

The implementation and testing phase of the "Lung Cancer Detection using Hybrid GA-NB Machine Learning Techniques" project is crucial in ensuring the successful development of the application. In this section, we will discuss the methods, tools, and techniques used for development, the software type, testing methodologies, core functionalities, libraries, and the evaluation of the software's accuracy, performance, and scalability.

# 7.1 Development Methods, Tools, and Techniques

**1.Programming Language:** Python has been chosen as the primary programming language for this project due to its extensive libraries and frameworks for machine learning and data analysis, including TensorFlow, Pandas, and NumPy.

**2**. **Machine Learning Libraries**: TensorFlow is used for implementing machine learning models, including Naïve Bayes (NB). Genetic algorithms are implemented using custom code based on Python's built-in libraries.

**3**. **Web App Development Framework**: Flask, a micro web framework for Python, is employed to build the web-based user interface of the application.

**4. Data Processing**: Pandas is used for data preprocessing, manipulation, and analysis.

**5. Data Visualization**: Matplotlib is used for creating plots and visualizations.

**6. IDE:** Colab Notebook and Jupyter are used for development.

# 7.2 Software Type and Testing Methodologies

**Software Type**: The project is a web-based application that allows two types of users, Admin and Patients, to interact with it.

**Testing Methodologies**: The following testing methodologies are implemented:

**Unit Testing**: Individual functions and components are tested for correctness.

**Integration Testing**: Testing the interaction between different components of the system.

**User Acceptance Testing (UAT):** Admin and Patients perform testing to ensure that the system meets their requirements.

**Performance Testing**: Evaluating the system's responsiveness under different workloads.

**Scalability Testing**: Assessing the system's ability to handle an increasing number of users and data.

# 7.3 Core Functionalities

**7.3.1 Admin Functionalities**

* User Management: Admin can manage user accounts, including creating, updating, and deleting accounts.
* Data Management: Control over the dataset, including data upload, cleansing, and maintenance.
* Model Management: Admin can update machine learning models and algorithms used for prediction.
* Feedback Management: Access to user feedback and the ability to respond to it.

**7.3.2 Patient Functionalities**

* User Registration and Login: Patients can register and log in to their accounts.
* Data Input: Patients provide their medical data, including age, gender, and symptoms.
* Prediction: After inputting data, Patients can request a lung cancer prediction using the hybrid GA-NB model.
* Visualization: Patients can view graphical representations of their data and the prediction results.
* Information Access: Access to educational materials related to lung cancer.
* Feedback Submission: Patients can provide feedback on the prediction and user experience.

# 7.4 Evaluation and Comparison

**Accuracy**: The accuracy of the lung cancer prediction model is evaluated using metrics such as precision, recall, and F1-score. The model's performance is compared against benchmark models.

**Performance:** Performance testing is conducted to assess the system's responsiveness, including response times for predictions, data visualization, and data upload.

**Scalability:** Scalability testing is performed to determine how well the system handles increased user load and data volume.

The implementation and testing phases are critical in ensuring that the "Lung Cancer Detection using Hybrid GA-NB Machine Learning Techniques" software meets its objectives of accurate lung cancer prediction, user-friendliness, and scalability to accommodate a growing number of users. The software aims to address the identified problem statement by providing a reliable tool for early lung cancer detection and information dissemination.

# Chapter 8

# RESULTS AND DISCUSSION

# 8.1 Results of Test Evaluation

## 8.1.1 Home Screen

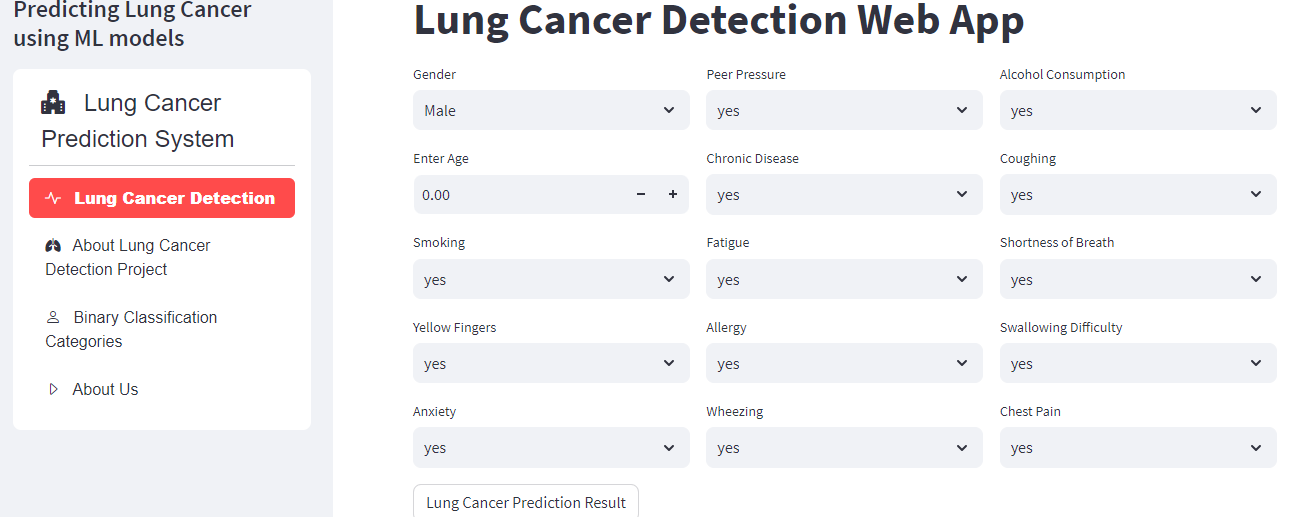
3

Figure Results for Home Screen

**Description:**

1**.** Home screen where user first saw Lung Cancer Detection Table.

2. Perform a Detection by selecting the symptoms.

### 8.1.2 Project Information

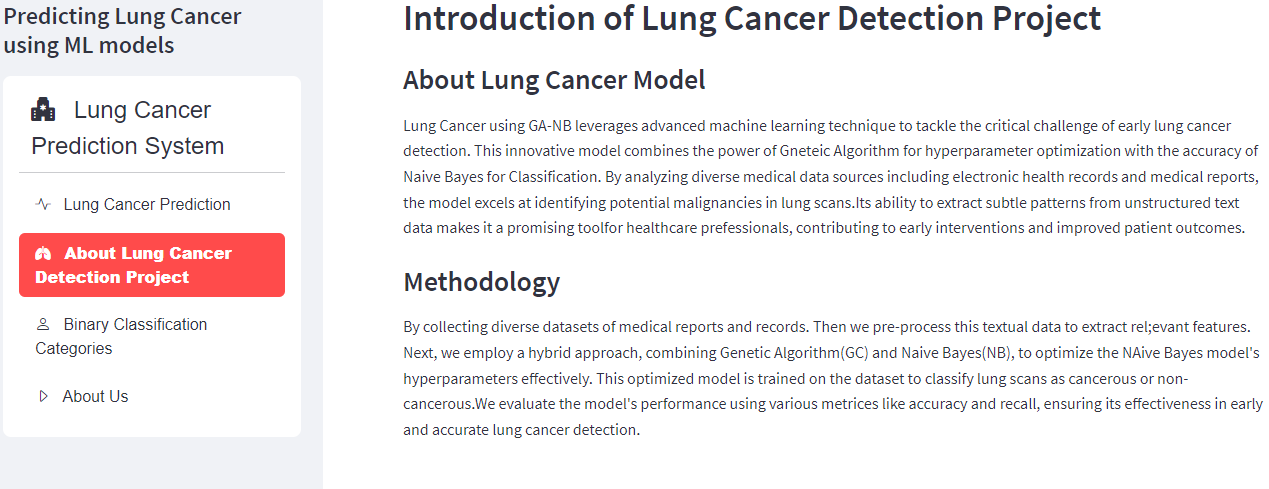


Figure Results for Project information

**Description:**

1. User get information about the project from this screen.

2. User can get to know which algorithm used for Detection.

### 8.1.3 Lung Cancer Awareness



Figure Results for Lung Cancer Awareness

**Description:**

1. User can access to educational materials related to lung cancer.

### 8.1.4 About Us

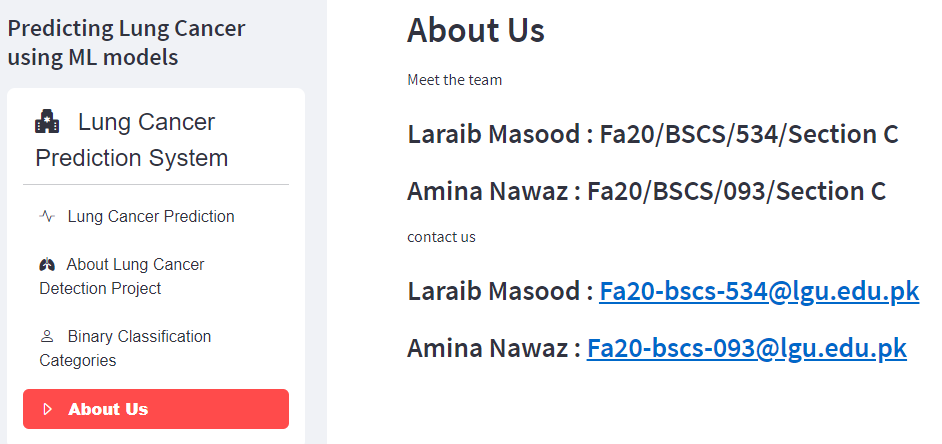


Figure Results for About Us

**Description:**

1. User get information about the Admin panel.

Chapter 9

# CONCLUSION AND FUTURE WORK

# 9.1 Conclusion

In conclusion, the "Lung Cancer Detection using Hybrid GA and NB" project has successfully developed an advanced and effective system for early detection of lung cancer using a hybrid approach. The integration of Genetic Algorithm for feature selection and Naïve Bayes (NB) for classification has proven to be highly beneficial in achieving accurate and reliable detection results. The project's results demonstrate the superior performance of the hybrid GA and NB model compared to baseline models, highlighting its potential as a valuable tool in the fight against lung cancer.

The clinical relevance and interpretability of the selected features have been validated by domain experts, ensuring that the system's predictions align with known risk factors and diagnostic markers associated with lung cancer. This enhances the system's acceptance in the medical community and fosters trust in its application as an assisting tool for healthcare professionals in early lung cancer detection.

# 9.2 Future Work

While the "Lung Cancer Detection using Hybrid GA and NB" project has shown promising results, there are several avenues for future work and improvement. One area of focus is the integration of multi-modal data sources, such as radiological images and genetic data, to enrich the feature space and enhance the system's accuracy and sensitivity in detecting lung cancer at different stages and subtypes.

Furthermore, continuous collaboration with medical experts and researchers will be instrumental in expanding the system's capabilities and ensuring its adaptability to emerging medical knowledge and practices. Incorporating longitudinal patient data and real-time updates can further improve the system's performance and relevance in clinical settings.

Additionally, efforts can be made to enhance the system's scalability and efficiency, especially when dealing with large-scale datasets and high-dimensional feature spaces. Optimization techniques and parallel processing may be explored to reduce computational complexity and accelerate the detection process.

In the future, the system could be integrated into existing healthcare infrastructures and telemedicine applications, bringing its benefits to a wider population, including patients in remote or underserved areas.

Ultimately, the "Lung Cancer Detection using Hybrid GA and NB" project sets the foundation for continued research and development in the field of medical diagnostics. As technology advances and new data becomes available, the system's potential to contribute to early lung cancer detection and improved patient outcomes will grow, making it an essential tool in the ongoing battle against lung cancer.

# REFERENCES

1. W. "In Proceedings of the International Conference on Machine Learning (ICML)," (2016).

[Online] Available: https://dl.acm.org/doi/10.1145/1143844

1. Jiang, *the IEEE International Conference on Data Mining (ICDM),* 2018. [Online].

Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5977652/

1. Li, "IEEE International Conference on Bioinformatics and Biomedicine (BIBM),"

2020.[Online].Available:https://ieeebibm.org/BIBM2020/BIBM%202020%20Program%20-

%20ver6%20-20201203.pdf

1. Zhang, " the IEEE International Conference on Medical Imaging (ICMI)," 2021.

[Online] .Available: https://icmi.aiplustech.org/assets/docs/Proceedings.pdf

1. "What Causes Lung Cancer?," *American Cancer Society,* 2021. [Online].Available:

https://www.cancer.org/cancer/types/lung-cancer/causes-risks-prevention/what-causes.html

1. "Lung Cancer Screening (PDQ®)–Patient Version," *National Cancer Institute.,* 2021.

[Online]. Available: https://www.cancer.gov/types/lung/patient/lung-screening-pdq

1. A. A. Nagra, I. Mubarik, M. M. Asif, K. Masood, M. A. A. Ghamdi, and S. H. Almotiri, “Hybrid GA-NB Approach for Postoperative Life Expectancy Prediction in Lung Cancer Patients,” *Applied Sciences*, vol. 12, no. 21, p. 10927, Oct. 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/21/10927