A FIELD PROJECT REPORT

on

“Enhancing Lung Cancer Detection With Machine Learning”

Submitted

221FA04212

BH. Veda Prakash

221FA04403

M. Ujwal

221FA04615

A. Niharika

221FA04674

M. Sai Pushpak

Under the guidance of

Dr. S. Deva Kumar

Associate Professor

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

VIGNAN'S FOUNDATION FOR SCIENCE, TECHNOLOGY AND RESEARCH Deemed to be UNIVERSITY

Vadlamudi, Guntur.

ANDHRA PRADESH, INDIA, PIN-522213.

VIGNAN'S
FOUNDATION FOR SCIENCE, TECHNOLOGY & RESEARCH
(Deemed to be University) - Estd. u/s 3 of UGC Act 1956

CERTIFICATE

This is to certify that the Field Project entitled **"Enhancing Lung Cancer Detection With Machine Learning"** that is being submitted by 221FA04212 (Veda Prakash), 221FA04403(Ujwal), 221FA04615(Niharika) and 221FA04674(Sai Pushpak) for partial fulfilment of Field Project is a bonafide work carried out under the supervision of Dr. S. Deva Kumar., Associate Professor, Department of CSE.

Dr.S. Deva Kumar
Associate Professor

Dr.S.V. Phani Kumar
HOD,CSE

Dr.K.V. Krishna Kishore
Dean, SoCI

VIGNAN'S
FOUNDATION FOR SCIENCE, TECHNOLOGY & RESEARCH
(Deemed to be University) - Estd. u/s 3 of UGC Act 1956

DECLARATION

We hereby declare that the Field Project entitled "**Enhancing Lung Cancer Detection With Machine Learning"** that is being submitted by 221FA04212 (Veda Prakash), 221FA04403(Ujwal), 221FA04615(Niharika) and 221FA04674(Sai Pushpak) in partial fulfilment of Field Project course work. This is our original work, and this project has not formed the basis for the award of any degree. We have worked under the supervision of Ms. Dr. Dr. S. Deva Kumar, Associate Professor, Department of CSE.

By

221FA04212 (Veda Prakash),
221FA04403(Ujwal),
221FA04615(Niharika),
221FA04674(Sai Pushpak)

Date: 15/10/2024

ABSTRACT

The clinical investigation of lung cancer has been much improved by recent developments in imaging and sequencing technology; yet, the enormous volumes of data produced beyond human capacity for efficient interpretation. For analyzing and combining these massive, intricate datasets, machine learning (ML) has become an essential tool that offers fresh insights into the diagnosis and treatment of lung cancer. ML models, particularly deep learning algorithms like convolutional neural networks (CNNs), have demonstrated considerable potential in early detection by accurately identifying lung lesions in medical imagery like CT scans. These models can enhance diagnostic accuracy and produce more individualized screening programs when paired with clinical data unique to each patient. By examining past patient data, machine learning is also revolutionizing prognosis prediction and auxiliary diagnosis by helping to identify subtypes of lung cancer and forecast the course of the disease. By examining immunological and genetic profiles, machine learning (ML) is being utilized in the field of immunotherapy to forecast patient reactions and better customize immunotherapy therapies. Furthermore, ML plays a critical role in precision medicine by directing the creation of targeted medicines through the analysis of molecular data unique to each patient, resulting in individualized treatment regimens. The availability of high-quality data, privacy issues, and the interpretability of ML models in healthcare contexts are some of the obstacles that still exist despite its enormous potential. In order to make ML models more transparent and reliable, future prospects include combining multi-omics and real-world data with developments in explainable AI. Machine learning has the potential to greatly enhance lung cancer detection, prognosis, and therapy as it develops further, which will ultimately result in improved patient outcomes and more individualized cancer care.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER-1

INTRODUCTION

1. INTRODUCTION

1.1 Background and Significance of Lung Cancer

When cells in the lungs grow abnormally, they can create tumors that can spread to other parts of the body, which is how lung cancer starts. It is one of the most prevalent and deadly types of cancer in the world, contributing significantly to cancer-related fatalities. Although second hand smoke, environmental pollutants, genetic predispositions, and occupational hazards (such as asbestos exposure) can also cause lung cancer in non-smokers, smoking is the main risk factor for the illness, accounting for about 85% of cases.

Lung cancer comes in two primary varieties:

About 85% of cases are non-small cell lung cancer (NSCLC), making it the most prevalent type. Adenocarcinoma, squamous cell carcinoma, and giant cell carcinoma are some of its subtypes.

A more severe type of lung cancer that frequently spreads swiftly and is typically associated with chronic smoking is small cell lung cancer (SCLC).

Significance of Lung Cancer

Global Health Burden: With more than 2 million new cases diagnosed each year, lung cancer is the primary cause of cancer-related deaths worldwide. Because of its high death rate and the challenge of early detection, it presents a serious health burden.

Economic Impact: Long-term treatment, hospital stays, and the financial loss from lower workforce participation are some of the ways that lung cancer raises healthcare expenses. Families and caregivers are also burdened financially and emotionally.

Difficulties in Detection and Treatment: Because early-stage lung cancer sometimes exhibits no symptoms, it is diagnosed at an advanced stage when there are few available treatments. Despite the widespread use of immunotherapy, radiation therapy, chemotherapy, and surgery, survival statistics are still poor, particularly for late-stage lung cancer.

Prevention and Awareness: Public health campaigns place a strong emphasis on quitting smoking, lowering exposure to environmental pollutants, and implementing screening programs (such low-dose CT scans) to detect lung cancer early. Targeted treatments may become possible as our knowledge of the genetic alterations linked to lung cancer expands.

1.2 Overview of Machine Learning in Medical Diagnosis

Medical diagnosis is being transformed by machine learning (ML), which gives computers the ability to examine enormous volumes of medical data, identify trends, and forecast outcomes. It gives medical personnel effective tools to increase the speed, accuracy, and efficiency of diagnosis for a range of illnesses.

Machine Learning Applications in Medical Imaging and Diagnosis:

Medical Imaging:

Radiology: X-rays, MRIs, CT scans, and ultrasounds are frequently analyzed using machine learning (ML) models, particularly deep learning approaches like convolutional neural networks (CNNs). These models help radiologists diagnose diseases like cancer early because they can accurately identify abnormalities like tumors, fractures, and blemishes.

Ophthalmology: By analyzing retinal images, ML is used to identify glaucoma and diabetic retinopathy.

Predicting and diagnosing diseases:

Cancer Diagnosis: To find biomarkers and forecast the likelihood or existence of cancer early on, machine learning models examine trends in genetic data, medical histories, and test findings. For instance, ML can detect cancers in mammograms early on in the identification of breast cancer.

Cardiovascular Disease: By examining clinical data, like blood pressure and cholesterol levels, as well as patient history and lifestyle, machine learning algorithms assist in predicting the risk of heart disease.

Neurological Disorders: By examining brain scans, behavioral data, and cognitive test results, machine learning is utilized to diagnose Alzheimer's disease and other neurodegenerative diseases.

Genomics and Pathology:

Histopathology: By using machine learning models to examine tissue samples (biopsies) and identify malignant cells, pathologists can make precise diagnoses faster.

Genomic Data Analysis: Using a patient's genetic profile, precision medicine uses machine learning (ML) to analyze complicated genomic data, find mutations connected to diseases, and create individualized treatment strategies.

Analytics for Prediction:

Chronic Disease Management: By tracking and forecasting the course of chronic conditions including diabetes, high blood pressure, and renal disease, machine learning models assist physicians in taking early action and modifying treatment plans.

Sepsis Detection: ML models can examine clinical information and vital signs to forecast when sepsis will strike hospitalized patients, thereby lowering mortality through prompt treatment.

NLP, or natural language processing:

Medical Records: By summarizing patient histories, diagnoses, and treatment plans, NLP approaches are utilized to extract information from unstructured medical records, facilitating better clinical decision-making.

Symptom Analysis: Machine learning (ML)-based chatbots and software can now comprehend symptoms reported by patients and offer potential diagnosis or medical advice.

1.3 Research Objectives and Scope

Research on machine learning in medical diagnostics may have the following goals:

Boost Diagnostic Accuracy: By examining clinical and imaging data, machine learning models are created to improve the precision of early diagnosis for conditions like cancer, heart disease, diabetes, and neurological disorders.

In order to enable proactive medical interventions, predictive models that evaluate a patient's likelihood of contracting specific diseases based on genetics, environmental factors, and patient history should be developed.

Cut Down on Diagnostic Time: To investigate how machine learning can speed up diagnostic results by cutting down on the amount of time needed to analyze complicated medical data (such as radiological pictures or pathology reports).

Improve Personalized Medicine: Examine how machine learning can be applied to create individualized treatment regimens using patient information, including genetic markers, lifestyle choices, and medical background.

Expand Access to Diagnostic Tools: To investigate the possibilities of machine learning-based diagnostic tools that can be applied in rural or low-resource environments where access to medical specialists is limited.

Reduce Bias and Enhance Generalization of the Model:

To improve diagnostic accuracy across a range of populations by identifying and reducing biases in machine learning models through training on representative and diverse datasets.

Integrate with Clinical Workflow: To investigate how machine learning tools can be easily included into the current clinical workflow, guaranteeing that medical practitioners can make efficient use of them without interfering with accepted procedures.

Research Scope

1. Machine Learning Algorithms:

Examining various machine learning methods, including deep learning (e.g., convolutional neural networks for image processing), unsupervised learning (e.g., clustering algorithms), and supervised learning (e.g., support vector machines, random forests, neural networks).

2. Application in Various Medical Fields:

Oncology: Using imaging and molecular data to diagnose malignancies (prostate, breast, and lung) early.

Cardiology: Using ECGs, echocardiograms, and other clinical measures to diagnose and forecast cardiac diseases.Automated analysis of radiological images, including MRIs, CT scans, and X-rays, is known as radiology.

Pathology: Machine learning is used to analyze histopathology images and find abnormalities, such as malignant cells.

Genetic data analysis for disease susceptibility and individualized treatment strategies is known as genomics.

3.Sources of Data:

using data from a variety of sources, such as wearable technology, lab test results, clinical data, electronic health records (EHR), medical imaging, and genetic data. Natural language processing (NLP) may also be used to process unstructured data from patient records, doctor's notes, and online health forums.

4.Legal and Ethical Aspects to Consider:

addressing moral issues such as patient permission, data privacy, and the duty of healthcare providers when using machine learning algorithms to make decisions. It also addresses adherence to regulatory frameworks such as GDPR and HIPAA.

5.Obstacles and Restrictions:

recognizing the difficulties in using machine learning systems for medical diagnosis, such as possible biases in predictions, interpretability of models, and data quality.

6.Assessment of the Model:

In order to guarantee dependability in practical medical applications, the study will also assess machine learning models using measures including accuracy, precision, recall, F1 score, and ROC-AUC.

7.Effect on Medical Systems:

The study will evaluate the potential effects of machine learning in medical diagnosis on the larger healthcare system, with a focus on enhancing patient outcomes, lowering diagnostic errors, and lowering healthcare expenses.

8.Integration of Technology:

investigating the integration of machine learning tools with current healthcare technology, including cloud-based healthcare platforms, AI-driven diagnostic tools, and electronic health record (EHR) systems.

1.4 Current Challenges in Lung Cancer Detection

The high death rate of lung cancer is a result of numerous major obstacles to its identification. The disease's characteristics, screening technology constraints, and a number of clinical, biological, and logistical aspects all contribute to these difficulties.

1. Late-Stage Diagnosis: Early identification of lung cancer is challenging because the disease frequently exhibits no symptoms in its early stages. The cancer has frequently evolved to a more advanced stage by the time symptoms (such as a chronic cough, chest pain, or shortness of breath) manifest.

Low Uptake of Screening: Low-dose CT (LDCT) scans and other screening programs are advised solely for high-risk groups (such as heavy smokers), yet many eligible people may not get regular screening because they are reluctant, unaware, or lack access.

2. Diagnostic Procedures That Invade

Surgical procedures and biopsies: Biopsies and other invasive procedures are frequently necessary to confirm a diagnosis of lung cancer. These are not appropriate for everyone and come with dangers, especially for elderly or pre-existing patients.

False Positives and Overdiagnosis: Non-cancerous lesions may be mistakenly diagnosed as cancer as a result of screening techniques such as LDCT. This may cause patients to become anxious and undergo needless invasive procedures.

3. High Variability in the Features of Tumors

Heterogeneity of Lung Cancer: There are several subtypes of lung cancer, such as small cell lung cancer and non-small cell lung cancer, each of which behaves differently and needs a different course of treatment. It is challenging to create a detection technique that works for everyone because of this heterogeneity.

Quick Development of Certain Subtypes: There is a limited window for early detection and treatment of some types of lung cancer, especially small cell lung cancer (SCLC), because of their rapid progression.

4. The Existing Screening Tools' Drawbacks

Imaging with radiography: The most popular imaging methods, chest X-rays and CT scans, often overlook tiny nodules or lesions, particularly when they are still in the early stages. They might also find it difficult to distinguish between benign and malignant cancers.

Radiation Exposure: CT scans are inappropriate for some people, particularly for long-term surveillance, because repeated radiation exposure can raise the risk of secondary malignancies.

5. Absence of Useful Biomarkers

Need for Reliable Biomarkers: Although several genetic alterations (such EGFR and ALK) are linked to lung cancer, there are currently insufficiently accessible and trustworthy biomarkers for regular early detection. Although "liquid biopsies," or blood-based biomarkers, are being developed, they are not yet generally accessible or completely verified.

Complex Molecular and Genetic Environment: Numerous genetic mutations and molecular abnormalities can cause lung cancer, and finding useful biomarkers that can be applied to the wide variety of lung cancer subtypes can be challenging.

6. Healthcare Inequalities and Screening Access

Socioeconomic and Geographic Barriers: In low-income or rural locations, access to lung cancer screening, particularly LDCT scans, is frequently restricted. This leads to inequalities in early identification and Outcomes.

expense of Screening and Follow-Up: Even in nations where screening for lung cancer is available, some people may find the expense of follow-up diagnostic tests and treatments to be unaffordable, which could cause them to postpone or forego therapy.

7. Human error in misinterpreting imaging results: When radiologists read lung cancer screenings, they may overlook tiny nodules or mistake benign growths for malignant ones, which can result in incorrect diagnoses. This emphasizes the necessity of more accurate, automated instruments to aid in detection, including AI-driven imaging systems.

Interobserver Variability: Different radiologists may have different interpretations of the same imaging results, which can cause diagnostic variability and make detection even more difficult.

8. Early Non-Smoking Detection

Growing Incidence among Non-Smokers: Although smoking is still the leading cause of lung cancer, non-smokers are increasingly developing the disease, especially women. Because smoking history is usually emphasized as a risk factor in screening programs, these cases are more difficult to identify.

Lack of Known Risk Factors: It can be challenging to identify individuals who are at risk and develop screening procedures for them since nonsmokers who have been diagnosed with lung cancer may not have known risk factors, such as exposure to secondhand smoke, environmental pollution, or genetic susceptibility.

9. Limited Deployment of AI and Machine Learning Integration: Although machine learning models have the potential to improve the detection of lung cancer by analyzing imaging data (e.g., detecting nodules in CT images), clinical workflows have not yet made extensive use of these models.

Generalization Issues: AI models that have been trained on small datasets may not generalize well across a variety of patient demographics or healthcare environments, which could result in bias and incorrect diagnoses.

10. Opposition to Screening Initiatives

Patient Reluctance: Many patients are reluctant to take part in lung cancer screening programs, particularly those who are asymptomatic or unaware of their risk level. This reluctance is frequently fueled by misperceptions that screening is only required for symptomatic people, fear of the results, or a lack of faith in medical systems.

Public Awareness: The importance of early lung cancer identification is still generally not well understood, particularly by high-risk groups like former smokers or those exposed to occupational dangers.

1.5 Applications of ML to Lung Cancer Detection

Machine learning (ML) has demonstrated significant promise in improving the identification of lung cancer by increasing diagnostic accuracy, decreasing processing time, and facilitating early detection. Large datasets, such as clinical records, genetic information, and medical imaging, are used in the integration of machine learning (ML) in lung cancer diagnosis to help medical practitioners make quicker, better judgments.

Important Uses of Machine Learning in the Identification of Lung Cancer

1. Analysis of Medical Imaging

Finding Lung Nodules in CT Images:

Lung nodules in CT images are automatically detected and classified using machine learning (ML) models, especially deep learning models (such as convolutional neural networks, or CNNs). These nodules are tiny lumps of lung tissue that may or may not be cancerous. Early detection is essential for the diagnosis of lung cancer.

By helping radiologists spot minute details that conventional imaging analysis might overlook, AI-based technologies might increase the precision of diagnoses.

Nodule Characterization: In addition to detecting nodules, machine learning algorithms are able to distinguish between benign and malignant nodules by examining their size, shape, texture, and density. This lessens the need for intrusive diagnostic techniques and needless biopsies.

Computer-Aided Detection (CAD): CAD systems are made to identify any anomalies in CT and chest X-ray images so that radiologists can examine them further. By serving as a "second opinion," these devices lessen human error in the early detection of lung cancer.

2. Predictive Modeling for Early Detection Risk Stratification: Machine learning algorithms are able to estimate a person's risk of lung cancer by analyzing patient data such as age, smoking history, family history, and environmental exposures. This aids in identifying high-risk patients who, even before symptoms manifest, should have routine screening. Additionally, risk prediction models aid in screening program prioritization, directing scarce resources toward patients who stand to gain the most from early detection.

Personalized Screening Recommendations: By focusing on people who might not fall into conventional high-risk groups, such non-smokers, machine learning algorithms can tailor screening recommendations according to a person's health profile, increasing the efficacy of lung cancer screening.

3. Automated Histopathological Analysis Analysis of Biopsy Samples: Biopsies of lung tissue are frequently examined by pathologists in order to confirm a diagnosis of lung cancer. Histopathological image analysis can be automated with machine learning (ML) techniques, especially deep learning models, which can accurately identify malignant cells in lung tissue.

By lowering human interpretation variability of biopsy results, machine learning models improve the accuracy of lung cancer diagnosis.

The connection between cancer cells and the surrounding healthy tissues is one aspect of the tumor microenvironment that machine learning can examine in histopathology images. Treatment choices can be guided by this information, which can also be used to evaluate the tumor's aggressiveness.

4. Liquid Biopsies and the Identification of Biomarkers

Non-Invasive Biomarker Detection: Machine learning models are being created to examine blood samples (liquid biopsies) for exosomes, microRNAs, and circulating tumor DNA (ctDNA), which are biomarkers for lung cancer. By providing a non-invasive substitute for tissue biopsies, these biomarkers can detect lung cancer early on.

In order to find genetic abnormalities and changes unique to lung cancer, including EGFR mutations, ML algorithms can assist in finding patterns in genomic data. These patterns are crucial for individualized therapy regimens.

Omics Data Integration: Using machine learning models in conjunction with several biological data types (genomics, proteomics, and transcriptomics) can enhance the capacity to identify and categorize subtypes of lung cancer, resulting in more precise diagnosis and individualized treatment plans.

5. Forecasting Treatment Outcomes and Prognoses

In order to forecast how a patient will react to different treatments, including chemotherapy, radiation therapy, immunotherapy, and targeted therapies, machine learning algorithms can examine clinical and genomic data. This enables individualized treatment programs that enhance results.

In the event that the cancer recurs, predictive models can also estimate the chance of recurrence following therapy, allowing for earlier action and closer surveillance.

Probability of Survival: By examining variables such as tumor features, genetic mutations, treatment plans, and general health, machine learning models are able to forecast patient survival rates. Physicians can use these forecasts to inform their choices regarding the course of treatment and aftercare.

6. Using Natural Language Processing (NLP) to Extract Diagnostic Information from Medical Records: NLP methods can be used to extract diagnostic information from unstructured data in electronic health records (EHRs), including radiography, pathology, and doctor's notes. Better patient management is made possible by ML models that use natural language processing (NLP) to extract pertinent data on lung cancer diagnoses, symptoms, therapies, and outcomes.

Automated Report Generation: Doctors can more easily record diagnostic findings and treatment plans in a consistent way by using ML-powered natural language processing (NLP) technologies to automate the creation of structured reports from medical data.

7. Clinical Decision Support Systems (CDSS): These systems use machine learning (ML) to make suggestions to doctors in real time based on patient information. Based on the most recent clinical recommendations and research, these systems assist doctors in selecting the right diagnostic tests, interpreting imaging results, and suggesting treatment alternatives.

Reducing Diagnostic Errors: By identifying discrepancies or anomalies in a patient's diagnostic workup, CDSS helps to prevent lung cancer from being overlooked or misdiagnosed.

Benefits of ML in Lung Cancer Detection

Improved Accuracy: Machine learning models, particularly those using deep learning, have shown higher sensitivity and specificity in detecting lung cancer than traditional methods, leading to fewer false positives and false negatives.

Early Detection: By identifying lung cancer at an earlier stage, ML-driven screening tools can increase survival rates by enabling timely intervention.

Personalized Medicine: ML algorithms help in identifying unique genetic and molecular profiles of lung cancer patients, enabling personalized screening and treatment plans tailored to the individual.

Cost-Effective and Scalable: ML tools can process large amounts of data quickly, reducing the time and cost associated with manual analysis. This scalability is crucial for implementing widespread lung cancer screening programs.

Reduction of Human Error: Automated analysis reduces the chances of human error, ensuring that important diagnostic information is not overlooked.

Challenges of ML in Lung Cancer Detection

Data Availability and Quality: To train, machine learning algorithms need substantial, superior datasets. Inaccurate forecasts may result from model performance being impacted by partial or biased data.

Interpretability: A lot of machine learning models, particularly deep learning algorithms, are "black boxes" that don't reveal much about how they make a diagnosis. Clinical adoption may be hampered by this lack of transparency.

Concerns regarding patient privacy, data security, and the regulatory licensing of AI-driven diagnostic tools are raised by the application of machine learning (ML) in healthcare, especially in the identification of lung cancer.

Generalization: It may be difficult to extrapolate the outcomes of models trained on certain datasets to a variety of clinical contexts or demographics.

By increasing diagnosis speed and accuracy, facilitating early detection, and bolstering tailored care, machine learning is revolutionizing the detection of lung cancer. Lung cancer mortality could be considerably decreased by using machine learning (ML) in clinical decision support, biomarker analysis, and medical imaging. To fully reap its benefits in clinical practice, however, issues including data quality, model interpretability, and ethical considerations must be resolved.

CHAPTER-2

LITERATURE SURVEY

2. LITERATURE SURVEY

2.1 Literature review

One of the main causes of death for both men and women is lung cancer, which is frequently associated with smoking and exposure to toxic substances. Improved survival rates depend on early detection, and a number of machine learning techniques have been used to diagnose it. Promising outcomes have been demonstrated by approaches such as SVM, KNN, decision trees, and deep learning models like UNet and ResNet, particularly when paired with image processing techniques like CT scans. The significance of metabolic analysis and biomarkers for early detection is emphasized by recent studies. In contrast, machine learning algorithms that classify lung cancer with high accuracy rates include SVM and decision trees[1].

Machine learning approaches such as Support Vector Machines (SVM), Decision Trees, and Logistic Regression have been widely explored for lung cancer prediction; SVM has consistently demonstrated higher accuracy. Research highlights the need of comparing algorithms using a variety of datasets and cross-validation. Data mining techniques like artificial neural networks (ANN) and Naive Bayes are frequently assessed for post-operative life expectancy. Different classification techniques, such as SVM and KNN, have been compared in brain tumor identification; the results vary depending on the dataset[2].

There has been a great deal of research done on the use of machine learning for lung cancer identification, and a number of studies have shown how CAD systems can help radiologists. A seminal work presented a Support Vector Machine (SVM) classifier with encouraging outcomes for the identification of lung cancer. The efficacy of various techniques, including Random Forest, Naive Bayes, and K-Nearest Neighbors (KNN), in identifying cancers has also been compared. Preprocessing and feature extraction are crucial for improving accuracy in various models. Subsequent investigations will prioritize enhancing these algorithms and incorporating deep learning to achieve better forecasting and early identification[3].

Advanced approaches and procedures have been used in considerable study on lung cancer detection. Many strategies concentrate on enhancing early detection through segmentation and augmentation methods for more accurate identification. Techniques such as breath analysis using the electronic nose (e-nose) have also been developed; these methods use breath odor analysis to help detect respiratory disorders, including lung cancer. Finding anomalies in cancer nodules requires a critical process called feature extraction and classification. Research highlights the significance of early identification in enhancing survival rates[4].

Machine learning techniques have been investigated in a number of research to enhance lung cancer detection. Faisal evaluated a number of classifiers, and the Gradient-Boosted Tree model produced 90% accurate results. Patra used several classifiers to categorize data on lung cancer and discovered that the RBF classifier worked best, achieving an accuracy rate of 81.25%. Hussein demonstrated notable advancements in tumor classification for CAD systems by implementing deep learning and transfer learning approaches. Krishnaiah suggested using data mining methods for early lung cancer detection, such as Naive Bayes. Using Rotation Forest, Dritsas was able to identify high-risk lung cancer cases with a 99.3% AUC[5].

Prostate, lung, and breast cancer prediction techniques are covered in the paper "Cancer Prediction using Machine Learning". It focuses on using data-driven models to examine factors such as smoking habits for lung cancer and clump thickness in breast cancer. These techniques make it possible to categorize cancer as benign or malignant, which promotes early identification and raises survival rates. The method illustrates how, given patient data, machine learning-based algorithms can help predict the possibility of cancer[6].

The review of the literature investigates a range of cancer prognostic approaches, with an emphasis on the combination of genetic profiling with cutting-edge patient data processing tools. Because machine learning can handle big datasets and intricate patterns, it has become a popular method. Research elucidates its function in discerning high-risk patients by means of data mining, clinical data, and demographic variables. Many models have been put into practice, such as those that use radiological imaging and genetic data to diagnose cancer early. Together, these strategies seek to increase forecast accuracy and direct focused treatments[7].

Machine learning methods for predicting lung cancer, with a focus on analyzing risk variables such genetic predispositions and smoking patterns. Clinical datasets have been subjected to a variety of models, such as Support Vector Machines (SVM) and Artificial Neural Networks (ANN), which have shown encouraging results in early identification. Deep learning methods have also been applied to medical picture analysis, assisting in precise diagnosis. When combined, these developments allow for more focused and prompt interventions, which ultimately enhance patient outcomes[8].

It analyzes fractal analysis to identify malignant from healthy cells and focuses on early diagnosis through medical imaging, mainly using computed tomography (CT) and X-rays. The study highlights how machine learning algorithms can enhance diagnostic precision, particularly when deep learning models are used. Understanding cell morphology using fractal dimension analysis helps determine the stages of cancer and the efficacy of treatment[9].

Deep learning models such as ResNet50 and InceptionV3, which achieve high accuracy rates of 100% and 99.92%, respectively, have improved lung cancer detection . To understand how the models make decisions, explainable artificial intelligence (XAI) techniques like Grad-CAM and LIME are used. Radiologists can benefit from transparency as these techniques assist emphasize important picture features utilized in categorization [10] Nonetheless, differences in expert diagnoses and model focus regions highlight the necessity of expert validation in clinical contexts. More XAI models will be incorporated in future research to improve medical confidence.

The application of evolutionary methods to enhance deep learning and machine learning models for lung cancer diagnosis is highlighted in the literature review. Lu et al. optimized a CNN architecture with the use of the Marine Predators Algorithm, attaining an accuracy rate of 93.4%. Rahmani et al. selected and classified features with 98.65% accuracy using the KNN and Grasshopper Optimization Algorithm. Senthil et al. improved lung cancer prediction by combining Back Propagation Networks with Ant Lion Optimization. Vijh et al. used CNN to apply Particle Swarm Optimization and Whale Optimization, with a 97.18% accuracy rate. Lastly, Guo et al. accurately diagnosed lung cancer using CNN and Harris Hawk Optimizer[11].

The role of several machine learning algorithms in early diagnosis is covered in the literature review on lung cancer detection. Lung nodules can be identified from CT scans using methods such as Computer-Aided Detection and Diagnosis (CADe and CADx). Significant accuracy has been attained by techniques like 3D deep CNNs and electronic nose technology, which uses breath analysis to identify lung cancer . Early detection may also be possible using hybrid approaches that include machine learning, metabolomics, and biomarkers . Prediction accuracy has been further improved by SVM-based techniques and cutting-edge algorithms like the Water Cycle Bat Algorithm [12].

The breakthroughs in machine learning and deep learning for lung cancer diagnosis are highlighted in the literature review. For early diagnosis, feature extraction, and classification, methods including convolutional neural networks (CNNs), support vector machines (SVMs), and deep neural networks are used. Numerous studies concentrate on enhancing the sensitivity, specificity, and accuracy of lung cancer detection through the application of evolutionary algorithms, hybrid models, and image processing techniques. Optimizing algorithms and early detection techniques based on the Internet of Things also seem promising for improving diagnostic performance. Numerous techniques make use of datasets that are openly accessible and exhibit notable advancements over conventional ways[13].

Outlines various methods that make use of deep learning and machine learning to identify lung cancer. Deep learning was the main technique utilized by Sefat et al. to identify lung cancer in chest X-rays. CAD systems for lung cancer detection, including feature extraction and image preprocessing, were covered by Kumar et al. Machine learning techniques for automated lung cancer diagnosis were reviewed by Gupta et al. Vignesh et al. discussed techniques for detecting lung cancer, such as feature extraction and classification. Ahmadi et al. concentrated on employing image processing methods to identify lung nodules[14].

Significant gains in diagnosis accuracy have been observed when using artificial intelligence (AI) to diagnose lung cancer, especially when it comes to categorizing subtypes of the disease based on histological information. By analyzing medical imaging such as CT scans, AI-driven systems can identify problems earlier and improve patient outcomes. By offering comprehensive insights and facilitating individualized therapy, AI also supports well-informed treatment planning. Furthermore, AI models predict patient prognoses effectively, which improves care even more. By simplifying procedures and cutting down on time-consuming chores, integrating AI into diagnostic workflows improves productivity[15].

Diverse approaches are presented in the literature on the use of deep learning and machine learning for lung cancer screening. DenseNet121 was utilized by Ausawalaithong et al. to solve CAD-x problems with 74.43% accuracy. Baraa et al. combined metaheuristic techniques with image processing, demonstrating excellent early detection accuracy. When Sasikumar et al. compared RNN, KNN, and SVM, RNN produced an accuracy of 92.75%. An attention-based neural network with 97.40% accuracy was introduced by Mahaska et al. CNN and RNN-LSTM were integrated by Wang et al. to create a strong CAD system that performed well in the diagnosis of lung diseases[16].

The methods used to identify lung cancer have changed throughout time, moving from non-invasive approaches to sophisticated imaging. A noteworthy advancement in cancer diagnosis is the use of biomarkers, including NSE and CEA, whose presence in bodily tissues, blood, or urine can help detect cancer early. Furthermore, breath analysis, which measures volatile organic compounds (VOCs), has demonstrated potential as a non-invasive, economical approach of detecting lung cancer. Even though CT scans are quite successful, there are still issues with identifying tumors. These issues can be resolved with computer-assisted methods that make use of AI and image processing. All things considered, integrating biomarkers, breath analysis, and cutting-edge imaging methods offers promise for precise and early lung cancer identification[17].

The goal of the study is to use CT scan pictures to detect lung cancer. Processing the photos is required to extract important information like area, perimeter, and shape as well as to detect edges. These characteristics aid in determining the tumor's malignancy or benignity. The tumors are then categorized using a machine learning model called Support Vector Machine (SVM) based on these attributes. This technique increases lung cancer detection accuracy[18].

Using cutting edge techniques to improve lung cancer detection has been the subject of numerous studies. In one study, deep learning based on gene expression outperformed models like SVM (94%) and Random Forest (95%), with an accuracy of 99%. Another study found that combining SVM with CNN improved the speed and efficacy of early-stage lung cancer diagnosis with CT images, achieving 94% accuracy. A third study examined the effects of depression, social support, and resilience on lung cancer patients and concluded that mental health is vital to their overall wellbeing. Last but not least, a fuzzy soft set system identified lung cancer with 100% accuracy by analyzing symptoms including weight loss and chest pain, demonstrating the possibility of more precise detection using creative approaches[19].

Numerous studies have explored the application of AI techniques to detect lung cancer early and improve diagnostic accuracy. Machine learning (ML) algorithms such as decision trees (DTs), support vector machines (SVMs), and deep learning models like convolutional neural networks (CNNs) are widely used. These models help in recognizing patterns in medical images like CT scans to identify potential cancerous regions.

A key step in this process is image preprocessing, where techniques like Gaussian filtering and median filtering are used to reduce noise and enhance image quality. This ensures that the AI systems can process clearer images, improving their ability to detect lung nodules.

For image segmentation, methods such as thresholding and region growing are used to divide an image into different sections to highlight areas that may contain cancerous cells. Advanced techniques like U-Net and transformer networks have been highly effective in segmenting lung nodules from medical images, allowing for better identification of tumors.

Overall, the combination of preprocessing, segmentation, and AI algorithms has significantly improved the accuracy of computer-aided diagnosis (CAD) systems, helping doctors make faster and more accurate decisions in lung cancer detection[20].

2.2 Motivation

The urgent need for early and precise lung cancer detection—lung cancer is still one of the world's top causes of death—motivates this literature review. The correlation between smoking, exposure to chemicals, and environmental variables and lung cancer underscores the need for sophisticated diagnostic instruments. Since lung cancer survival rates increase dramatically when the disease is discovered early, early detection is essential. However, the inaccuracy and inefficiency of conventional diagnostic techniques make them inadequate for early detection, underscoring the significance of embracing new innovations.

The diagnosis of lung cancer could be revolutionized by machine learning (ML) techniques. When it comes to helping radiologists identify malignant nodules from medical images such as CT scans, machine learning (ML) offers quick, precise, and scalable solutions. When paired with image processing techniques, models like Support Vector Machines (SVM), Decision Trees, and other classification algorithms perform well, indicating the increasing importance of these technologies in improving diagnostic accuracy. Furthermore, early interventions—which are critical for improving patient outcomes—are made possible by computer-aided diagnosis (CAD) systems, which also lessen the strain of radiologists and increase accuracy.

The significance of feature extraction, segmentation, and image preprocessing methods for medical image analysis is further emphasized by this survey. Techniques like median and Gaussian filtering enhance image quality and facilitate lung nodule identification for machine learning algorithms. Furthermore, segmentation methods like thresholding and region expanding aid in the division of images into discrete parts, facilitating the identification of suspicious regions more accurately.

The overall goal of this literature review is to present a thorough summary of the state-of-the-art ML-driven techniques and stimulate new research to enhance lung cancer detection. Researchers can create more dependable, quick, and accurate diagnostic solutions by combining machine learning techniques with advances in medical imaging. This will ultimately improve patient outcomes and increase survival rates.

CHAPTER-3 PROPOSED SYSTEM

3. PROPOSED SYSTEM

A. Dataset, The lung cancer dataset includes 24 features covering demographic, environmental, lifestyle, and medical factors, such as age, gender, air pollution, smoking, genetic risk, and symptoms. The target variable, "Level," categorizes lung cancer risk as "Low," "Medium," or "High." All features are numeric, representing either ordinal or categorical values.

B. Data Preprocessing, The dataset was pre- processed to handle missing data using imputation techniques like mean, median, or k-nearest neighbours. Feature scaling methods such as Min-Max scaling or standardization were applied to normalize numerical values. Class imbalance was addressed using techniques like SMOTE or class-weighted models.

C. Exploratory Data Analysis (EDA) Correlation analysis identified relationships between features and the target variable, using heatmaps and scatter plots for visualization. Feature selection was guided by correlation matrices and feature importance, with methods like Recursive Feature Elimination (RFE) or Principal Component Analysis (PCA) employed to reduce dimensionality.

D. Model Development Several supervised learning algorithms were tested for lung cancer level prediction:

Logistic Regression: An interpretable model that provides insights into feature significance.

Random Forest: An ensemble method that handles categorical and continuous features well, offering feature importance scores.

Gradient Boosting (XGBoost, LightGBM): Enhances prediction accuracy by iteratively building weak learners.

Support Vector Machines (SVM): Useful for classifying non-linearly separable data, employing kernel functions if necessary.

Neural Networks (MLP): Applied for complex patterns, especially when feature interactions are significant.

E. Model Training The dataset was split into training (70%), validation (15%), and test (15%) sets. K-fold cross-validation (e.g., k = 5) was used to ensure generalizability and prevent overfitting. Hyperparameter tuning was performed using grid search or random search techniques.

F. Model Evaluation Model performance was evaluated using metrics like accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC curve. Special focus was placed on sensitivity and specificity, particularly for the "High" cancer level category, to minimize false negatives.

G. Model Interpretation Feature importance scores were analysed for models like Random Forest and Gradient Boosting. SHAP or LIME was used to interpret model predictions, providing transparency and clinical trust in the decision-making process.

H. Final Model Selection and Testing The best-performing model was chosen based on validation metrics, ensuring balanced sensitivity and specificity across all cancer levels. The model was then tested on unseen data to verify its generalization performance.

I. Deployment and Continuous Improvement The model was deployed as a decision-support tool in a clinical setting, with a web-based interface for inputting patient data and receiving cancer level predictions. Continuous model monitoring and updates were planned to refine predictions with new patient data.

J. Ethical Considerations Data privacy was ensured in compliance with regulations like HIPAA or GDPR. Regular checks were performed to mitigate biases, ensuring fair and equitable model performance across different demographic groups.

3.1 Input dataset

The dataset contains a number of characteristics that could affect or suggest health outcomes, and it seems to concentrate on aspects related to cancer patients. The collection contains patient-level information with a range of characteristics that could suggest symptoms or increase the risk of cancer. A distinct "Patient Id" is used to identify each patient in each row. The 26 columns in the dataset describe various lifestyle, genetic, and environmental factors as well as specific health outcomes and symptoms.

3.1.1 Detailed Features of the Dataset

Patient Id: A special number assigned to every patient.

Age: The patient's age.

Gender: The patient's gender (encoded; probably 1 for a man and 2 for a woman).

Air Pollution: Exposure level to air pollution (perhaps on a scale of 1 to 10).

Alcohol use: The frequency of alcohol intake (based on a scale).

Dust Allergy: Probability of having a dust allergy (with a scale).

Occupational Hazards: Workplace exposure to risks (scale-based).

Genetic Risk: A scale-based assessment of a person's genetic susceptibility to cancer.

Chronic Lung Disease: The existence of long-term lung conditions (based on a scale).

Balanced Diet: If the patient follows a scale-based balanced diet.

Obesity: Body mass index (scale-based) or degree of obesity.

Smoking: Smoking behaviours (based on a scale).

Exposure to second hand smoke (scale-based) is known as passive smoking.

Chest Pain: A scale-based measure of chest pain.

Blood Coughing: The occurrence of blood coughing (scale-based).

Fatigue: The patient's degree of weariness (scale-based).

Weight Loss: The degree of weight loss (based on a scale).

Breathlessness: The sensation of being out of breath (scale-based).

Wheezing: The existence of wheezing (based on a scale).

Consuming Difficulty: Scale-based swallowing difficulty.

Clubbing of Finger Nails: A scale-based clinical indication of fingernail malformation.

Frequent Cold: The scale-based frequency of cold symptoms.

Dry Cough: A dry cough episode (scale-based).

Snoring: A scale-based experience of snoring.

Level: Most likely the Low, Medium, or High degree of cancer severity.

3.2 Data Pre-processing

Data pre-processing is the essential process of preparing raw data for analysis and modelling by cleaning, transforming, and structuring it to enhance data quality and utility. It involves tasks like handling missing values, correcting errors, encoding features, and scaling data to ensure it's in an optimal form for further analysis. It encompasses a range of operations and transformations designed to refine raw data, ensuring that it is clean, structured, and amenity subsequent analysis. This process is driven by its manifold significance in data science and analysis.

Through meticulous data cleaning, transformation, feature engineering, dimensionality reduction, outlier handling, scaling, and data splitting, it prepares raw data for more accurate and reliable analysis and modelling. Ultimately, the goal is to obtain more meaningful insights, make informed decisions, and optimize predictive models for a wide range of applications in data science and analysis.

Dropping Unnecessary Columns

Index, patient ID, clubbing of fingernails, air pollution, swallowing difficulty, and gender were among the columns that were eliminated.

Reason: In order to simplify the dataset and lower noise for the model, these columns were judged unnecessary or unhelpful for predicting the severity of the malignancy.

Encoding the Target Variable:

'Low', 'Medium', and 'High' were the values of the categorical target variable Level, which LabelEncoder() converted into numeric values.

As a result, the numerical designations Low = 0, Medium = 1, and High = 2 were created from the 'Level' column. For machine learning models to properly process the target variable, this is necessary.

With pertinent features and a target variable that has been correctly encoded, the cleaned dataset is now prepared for model training. These preparation procedures guarantee that the data is formatted appropriately for machine learning algorithms to efficiently classify the severity of malignancy.

3.3 Model Building

Using the cleaned dataset, the model development portion of the study aimed to predict cancer severity (Low, Medium, High). The Naive Bayes classifier was the model chosen for this challenge because it is easy to use and effective at solving classification issues, especially when the features are independent.

Preparing Data

The dataset was first divided into two parts: characteristics (X) and the goal variable (y). X contained all of the pertinent patient features, while y stood for the target variable, "Level," which indicates the severity of the malignancy. Using standardization procedures, feature scaling was used to make sure the features were on the same scale. In order to keep features with higher values from overpowering those with lower values during model training, this step was essential.

Data Division

A training set (70%) and a testing set (30%) were created from the data. A trustworthy indicator of the model's performance is provided by this separation, which guarantees that it can learn from the training data and be assessed on test data that hasn't been seen yet.

Training of Models

The training data was used to train the Gaussian Naive Bayes model. Each of the three classes (Low, Medium, and High) has its probability determined by this model, which then chooses the class with the highest probability to be the prediction. To prevent any problems with zero probability when specific feature values are missing from the training data, a smoothing parameter was used.

Forecasting and Assessment

The model was used to forecast the test set's cancer severity after it had been trained. The model's fit to the data was evaluated by calculating both training and testing accuracies. While the training accuracy gauges how well the model learned from the training data, the testing accuracy offers information about how well the model performs on fresh, unseen data.

Important metrics including accuracy, precision, recall, and F1-score were calculated in order to assess the model further. A thorough understanding of the model's performance is offered by these metrics:

Accuracy gauges how accurate the model is overall.

The number of projected positive cases (such as high severity) that were actually true is known as **precision**.

The **model's recall** indicates how effectively it represented every real positive instance.

The **F1-score** is helpful when the dataset is unbalanced since it offers a balance between precision and recall.

The number of accurate and inaccurate predictions for each class (Low, Medium, and High) was displayed in a confusion matrix that was also created to represent the categorization findings. This made it easier to identify the model's strong points and areas for improvement. With a balance between training and testing accuracy, the Naive Bayes classifier produced encouraging results. According to the evaluation criteria (accuracy, precision, recall, and F1-score), the model demonstrated a respectable level of accuracy in classifying the severity of the malignancy. The confusion matrix also pointed out areas that can use improvement, like incorrectly classifying nearby severity levels (e.g., Medium vs. High).

3.4 Methodology of the system

A. Architecture of the System

Data collection, preprocessing, feature extraction, model training, and classification are some of the interrelated steps in the suggested system architecture for determining the severity of cancer based on patient data. The structure is made up of:

Input layer: Gathering patient information with a range of environmental and health-related characteristics.

Data transformation and cleaning for model training is done in the preprocessing layer.

Layer of feature extraction: obtaining pertinent features for efficient classification.

Classifier: Predicting the degree of malignancy by using a machine learning algorithm.

Output layer: Showing the classification outcome (High, Medium, or Low) according to the input data.

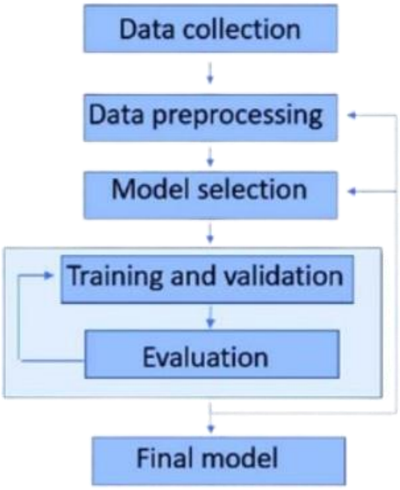
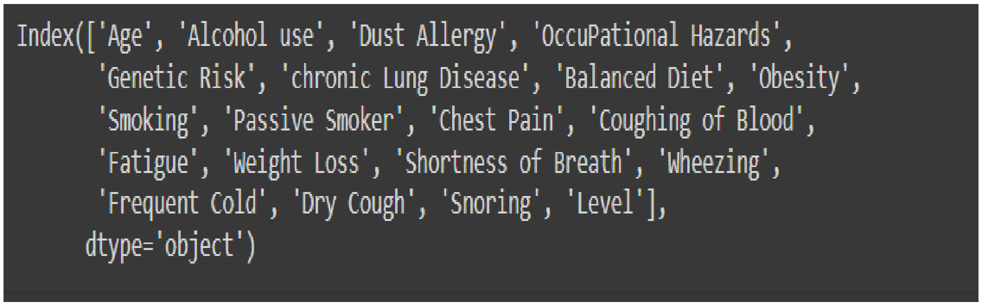


Figure 1. Architecture of the proposed system

B. Training and Preprocessing of Data

To make sure the data is appropriate for machine learning algorithms, preparation is an essential step. The preprocessing methods listed below were used:

Data cleaning is the process of eliminating columns like "Patient Id," "Clubbing of Finger Nails," and "Air Pollution" that are superfluous and do not substantially add to the classification.



```
Index(['Age', 'Alcohol use', 'Dust Allergy', 'OccuPational Hazards',
       'Genetic Risk', 'chronic Lung Disease', 'Balanced Diet', 'Obesity',
       'Smoking', 'Passive Smoker', 'Chest Pain', 'Coughing of Blood',
       'Fatigue', 'Weight Loss', 'Shortness of Breath', 'Wheezing',
       'Frequent Cold', 'Dry Cough', 'Snoring', 'Level'],
      dtype='object')
```

Figure 2. Various features in the dataset after Pre-Processing

Label Encoding: To make the target variable "Level" (Low, Medium, High) compatible with machine learning models, it is converted into numerical form.

Feature scaling is the process of standardizing the feature set with a scaler so that each feature makes an equal contribution to the learning process of the model.

Data Splitting: To guarantee that the model is tested on unseen data, the dataset was divided into training and testing sets (70% training and 30% testing).

C. Extraction of Features

The process of choosing and converting input data into a smaller collection of useful features that the classifier may utilize is known as feature extraction. After eliminating less important characteristics, pertinent characteristics like age, genetic risk, obesity, smoking, and alcohol use were kept in this study. By concentrating on variables most pertinent to the severity of cancer, feature extraction enhances model performance.

D. Bayes's Naive

Because of its ease of use and efficiency for classification tasks, the Naive Bayes classifier was selected as the main machine learning model. In order to compute probabilities for every class and generate predictions based on maximum likelihood estimation, Naive Bayes relies on the premise that features are conditionally independent. In this study, the Gaussian Naive Bayes variant was employed, which performs well with continuous data such as patient attributes.

E. Classification

The classification challenge is predicting the cancer severity (Low, Medium, High) using the retrieved features and the trained Naive Bayes model. The preprocessed dataset was used to train the model, and the test data was used to assess the classification accuracy. To evaluate the model's performance, metrics like accuracy, precision, recall, and F1-score were calculated. The model's ability to distinguish between the three severity levels was shown in detail by the confusion matrix.

F. Results

The system's output is a classification of each patient's cancer severity within the dataset. Following training, the system is able to estimate the severity level (Low, Medium, High) from fresh patient data. Healthcare practitioners can utilize the system's predictions to evaluate the course of cancer and choose the best course of therapy. The accuracy of the system is used to gauge its performance, and the results indicate that it has potential categorization capabilities for practical use.

3.5 Model Evaluation

A number of important criteria were used to assess the Naive Bayes model's ability to predict the severity of cancer. Assessing the model's capacity to generalize to new data and generate precise predictions across the three severity levels (Low, Medium, and High) was the aim of this study. The model's performance was assessed using the following metrics:

A. Accuracy of Training and Testing

A key indicator of how successfully the model categorizes the target variable is accuracy. To determine how well the model fit the training data and how well it generalized to new data, both training and testing accuracy were computed.

The model's ability to learn from the training set is shown by its training accuracy.

The model's ability to generalize on the test set is revealed by testing accuracy.

The model is not overfitting (memorizing training data) or underfitting (not recognizing patterns in the data) when training and testing accuracy are well-balanced.

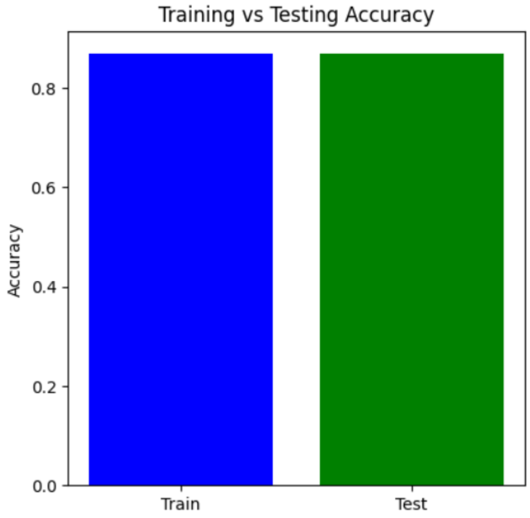


Figure 3. Training Vs Testing Accuracy

B. Confusion Matrix

The model's classification performance was assessed using the confusion matrix, which offers a thorough analysis of true positives, false positives, true negatives, and false negatives for each of the three classes (Low, Medium, and High). The matrix assisted in figuring out:

How often the model successfully classified each severity level.

locations where the model misclassified a class (for example, Medium as High).

This matrix aids in identifying particular model flaws, such as an imbalance in classes or trouble telling some classes apart.

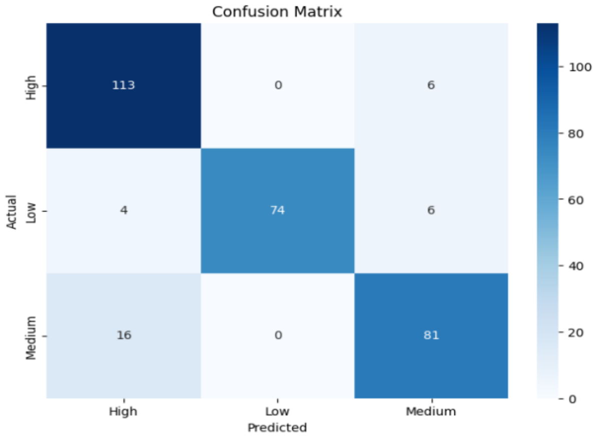


Figure 4. Confusion Matrix

C. Accuracy

Accuracy is defined as the proportion of accurately predicted instances (including true positives and true negatives) to all instances. Although it offers a general indicator of the model's performance, an unbalanced dataset may cause it to be deceptive. Here, accuracy is used as a starting point.

D. Precession

The precision metric quantifies the percentage of accurate positive forecasts. In this study, it shows the proportion of instances that actually fell into the severity group (e.g., High) that was predicted. Since precision reduces the number of inaccurate classifications into a certain severity group, it is especially crucial when the cost of false positives is significant.

E. Recall

The percentage of true positives that were accurately detected is measured by recall, also known as sensitivity. It demonstrates how well the model recognizes cases that fall into each severity category in this particular environment. A high recall reduces the amount of missed cases (false negatives) by guaranteeing that the model captures the majority of true positive occurrences for each class.

F. F1-Score

The harmonic mean of recall and precision is the F1-score. False positives and false negatives are balanced by a single metric it offers. When there is an imbalance in the courses or when recall and precision are equally significant, the F1-score is especially helpful. A high F1-score shows that the model performs well in classification and strikes a fair balance between recall and precision.

G. Outcomes of Performance

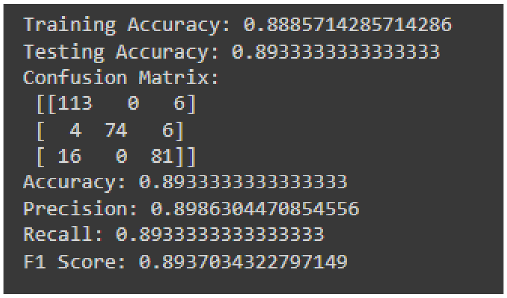
The following conclusions were drawn from the model's performance on various metrics:

Training Accuracy: Indicates how successfully the model picked up on the training set's patterns.

Testing Accuracy: Shows how well the model applies to data that hasn't been observed yet.

Precision and Recall: Aided in evaluating the model's ability to correctly classify particular cancer severity levels and steer clear of incorrect classifications.

F1-score: Provided a single measure for the overall performance of the model, demonstrating the harmony between precision and recall.



```
Training Accuracy: 0.8885714285714286
Testing Accuracy: 0.8933333333333333
Confusion Matrix:
 [[113   0   6]
 [  4  74   6]
 [ 16   0  81]]
Accuracy: 0.8933333333333333
Precision: 0.8986304470854556
Recall: 0.8933333333333333
F1 Score: 0.8937034322797149
```

Figure 5. Performance Outcomes

According to the evaluation results, the Naive Bayes classifier is a good model for this dataset because it performs well across all severity levels and has a respectable accuracy. Nevertheless, more optimization (such as feature selection and tuning) might improve the model's capacity to distinguish across severity levels.

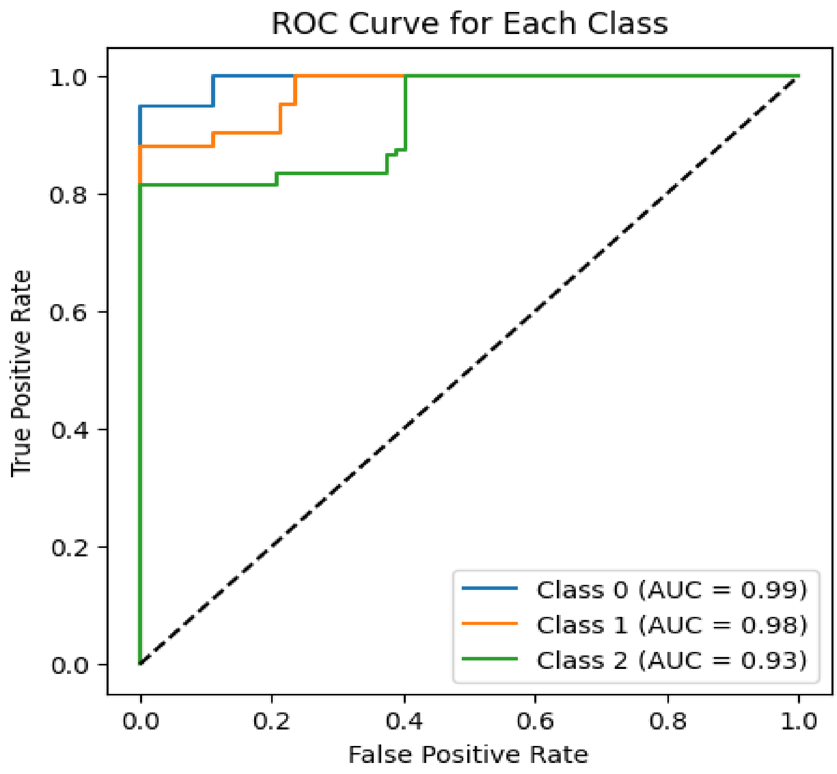


Figure 6. ROC Curve for Each Class

To see each classifier's performance, confusion matrices were plotted. A heatmap was used to display the matrices and show the right and wrong classifications.

Logistic Regression

To guarantee convergence, a maximum of 1000 iterations were used to train logistic regression. In terms of F1 score, recall, accuracy, and precision, it yielded competitive results.

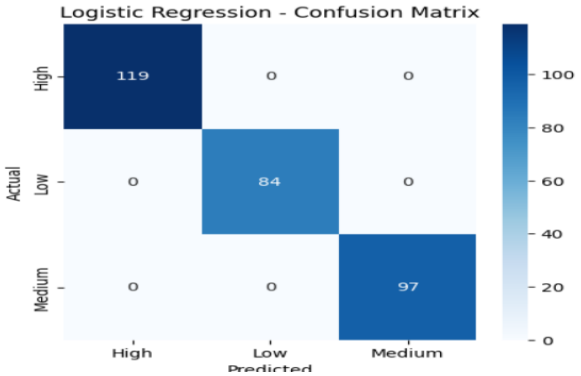


Figure 7. Logistic Regression – Confusion Matrix

Naive Bayes

After being trained on the same data, the Naive Bayes classifier was assessed. Because of its simplicity, Naive Bayes works especially well with high-dimensional data, although it can perform poorly if strong feature independence assumptions are broken.

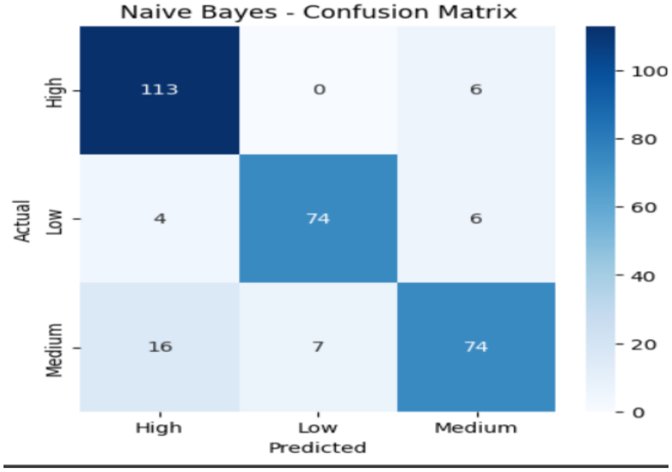


Figure 8. Naïve Bayes – Confusion Matrix

Support Vector Machine (SVM)

Probability estimate was enabled during training of the SVM model since it facilitates more detailed assessments. Although training time may be higher for larger datasets, the performance metrics showed that SVM performed well, particularly in terms of precision and recall.

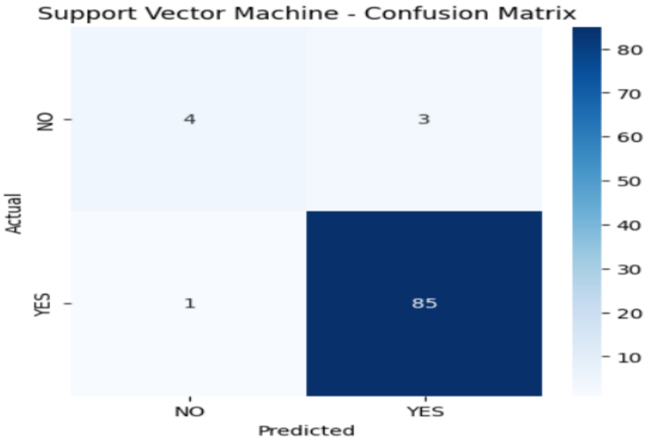


Figure 9. Support Vector Machine (SVM) -– Confusion Matrix

Random Forest

Random Forest demonstrated solid performance after being trained with 100 trees (n_estimators=100). Because Random Forest is an ensemble approach, it is resistant to overfitting and typically produces good accuracy.

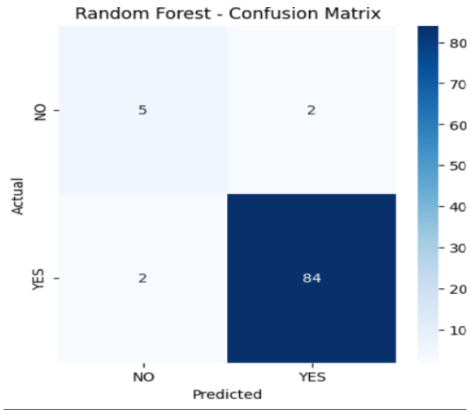


Figure 10. Random Forest – Confusion Matrix

XGBoost

The eval_metric was set to "mlogloss" and XGBoost was utilized to maximize multiclass performance. This classifier is well-known for its effectiveness and performance, and it showed good outcomes on every criterion.

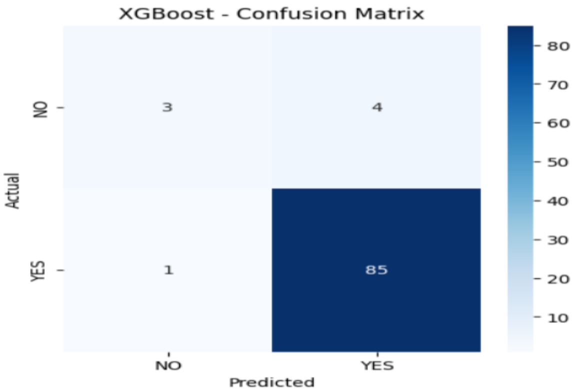


Figure 11. XGBoost – Confusion Matrix

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.95	0.95	0.95	0.95
Naive Bayes	0.87	0.871	0.87	0.86
Support Vector Machine	0.956	0.953	0.956	0.953
Random Forest	0.956	0.956	0.956	0.956
XGBoost	0.98	0.981	0.98	0.98

Table 1. Recorded Results for each Classifier

Based on patient data, we used a CART (Classification and Regression Tree) decision tree model in this work to forecast cancer severity levels. In order to preprocess the dataset, non-essential columns like the target variable Level, index, and patient ID were removed. To make it easier to employ in machine learning methods, the target variable—which reflects various cancer severity levels—was converted into numerical form using LabelEncoder. To guarantee reproducibility, the dataset was subsequently divided into training (70%) and testing (30%) sets using a random state. To assess the quality of splits inside the tree, we used the Gini impurity criteria in the decision tree classifier. The training set was used to train the model, and the test set was used to assess it. Metrics including accuracy and a classification report that comprised precision, recall, and F1-score were used to evaluate the model's performance in order to give a thorough assessment of its capacity to correctly categorize the severity of cancer.

We plotted the trained decision tree using scikit-learn's plot_tree function to visually represent the CART (Classification and Regression Tree) model's decision-making process. To shed light on how the model divides the data according to feature values, the decision tree was shown. To guarantee readability and clarity, the figure was sized at 12 by 8. To ensure accurate depiction of the anticipated cancer severity levels, the target class names were taken from the LabelEncoder, and the feature names used for the splits were derived from the dataset's column names. Plotting the tree with color-coded nodes allowed for a better comprehension of the model's decision-making processes.

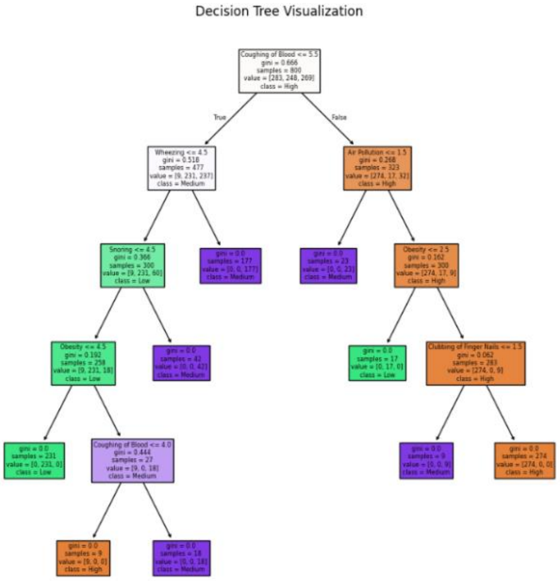


Figure 12. Decision Tree Visualization

i. **Quality Assurance**: Model evaluation helps ensure that the model is capable of making accurate predictions when exposed to real-world data. It acts as a quality control mechanism to validate the model's generalization ability.

ii. **Comparing Models**: Model evaluation allows for the comparison of multiple models to identify the best-performing one. It helps data scientists and stakeholders make informed decisions about which model to deploy.

iii. **Fine-Tuning**: The evaluation process can reveal areas where the model performs poorly. This information is valuable for refining the model, making it more robust, and addressing its limitations.

iv. **Business Decision Support**: In practical applications, model performance impacts critical business decisions. A well-evaluated model provides confidence to stakeholders, leading to better decision-making.

v. **Model Deployment**: A thoroughly evaluated model is more likely to be deployed in production systems. It instils trust in the model's predictions, which is essential in real-world applications.

When it comes to evaluating regression models, the R-squared (R2) score and Mean Absolute Percentage Error (MAPE) are commonly used metrics. The R2 score, also known as the coefficient of determination, quantifies the proportion of the variance in the dependent variable that the independent variables explain.

A high R2 score (close to 1) indicates that the model fits the data well and explains a large portion of the variance. Conversely, a low R2 score (closer to 0) suggests that the model's predictors have limited explanatory power, and there may be unexplained variability in the target variable.

Assume a dataset has n values marked $y_1,...,y_n$ (collectively known as y_i or as a vector $\boldsymbol{y} = [y_1,...,y_n]^T$), each associated with a fitted (or modelled, or predicted) value $f_1,...,f_n$ (known as f_i, or sometimes $\hat{y}_i$, as a vector $\boldsymbol{f}$).

Define the residuals as $e_i = y_i - f_i$ (forming a vector $\boldsymbol{e}$).

If $\bar{y}$ is the mean of the observed data: $\bar{y} = (\frac{1}{n}) * \sum_{i=1}^{n} y_i$

then the variability of the data set can be measured with two sums of squares formulas:

- The sum of squares of residuals, also called the residual sum of squares:

$$SS_{res} = \sum_{i=1}^{n} e_i^2$$

- The total sum of squares (proportional to the variance of the data):

$$SS_{tot} = \sum_{i=1}^{n}(y_i - \overline{y})^2$$

The most general definition of the coefficient of determination is

$$R^2 = 1 - (\frac{SS_{res}}{SS_{tot}})$$

Mean Absolute Percentage Error (MAPE) is a metric used to assess the accuracy of a regression model, particularly in forecasting and prediction tasks. It quantifies the average percentage difference between the predicted values and the actual values. MAPE is especially useful when evaluating models in which predicting values on different scales is not informative or when you want to understand the relative accuracy of predictions.

$$MAPE = (\frac{1}{n}) \sum_{t=1}^{n} |\frac{A_t - F_t}{A_t}|$$

where At is the actual value and Ft is the forecast value. Their difference is divided by the actual value At. The absolute value of this ratio is summed for every forecasted point in time and divided by the number of fitted points n.

3.6 Constraints

We work within a set of particular limitations in our lung cancer detection project, which influence how we approach the solution's design and development. These limitations guarantee that our model complies with crucial factors and restrictions pertaining to healthcare and medical data:

i. **Authenticity**: We accept the possibility of incomplete or erroneous data. Our dataset may contain errors due to patient-reported symptoms or environmental factors that don't always match real situations. This danger emphasizes how crucial it is to put data verification procedures in place to guarantee the validity and dependability of the data used to train and test our model, lessening the effect of any potential errors on the final predictions.

ii. **Privacy:** When handling medical data, security and privacy are crucial. To safeguard private patient data, we follow stringent data access and privacy guidelines. Our initiative ensures that no personally identifiable information is utilized or disclosed by adhering to all applicable legal and ethical requirements, including HIPAA compliance. These limitations are necessary to protect the privacy of data and guarantee that the use of medical data complies with the law.

iii. **Cost:** Although our dataset was obtained from a publicly accessible website such as Kaggle, we acknowledge that producing or obtaining high-quality patient data for the detection of lung cancer sometimes entails monetary expenses. This covers costs for operations, maintenance, and data collecting (such as imaging, clinical research, or medical testing). To ensure cost-effectiveness without sacrificing accuracy or data quality, it is imperative that we strike a balance between these expenses and our project goals.

iv. **Data Quality:** The effectiveness of our lung cancer detection model depends on ensuring excellent data quality and integrity. We are constrained by the need to uphold strict data quality standards, which entails procedures like data cleansing, validation, and verification to eliminate errors or noise. To increase our model's accuracy and dependability, we need high-quality data, especially in the healthcare industry where accuracy is crucial.

v. **Resource Availability:** The main limits of our project are computer power, access to medical datasets, and human knowledge. Our goal is to maximize the utilization of the resources available by designing and implementing our model as efficiently as possible. This entails choosing suitable algorithms and methods (such the Naive Bayes classifier) that strike a compromise between computational effectiveness and precise forecasts, guaranteeing that the project stays viable and scalable in light of our resource limitations.

3.7 Cost and sustainability Impact

Our approach to the creation and execution of our lung cancer detection project is heavily influenced by sustainability consequences as well as cost concerns. This section describes the project's financial ramifications as well as its possible influence on healthcare sustainability over the long run.

A. Cost Consequences

Infrastructure and Equipment:

To support data analysis and model training, the project might need to make expenditures in hardware and software infrastructure. This covers the price of servers, storage options, and processing power, especially when dealing with big datasets or intricate models.

Costs of Operations:

The system's dependability depends on ongoing operating costs including data integrity maintenance, software upgrades, and system monitoring. Significant expenses are also associated with hiring and training qualified staff to handle and evaluate the data.

Costs of Data Acquisition:

Although our original dataset came from Kaggle, obtaining more datasets—especially proprietary or clinical data—may be expensive in order to guarantee thorough and high-quality data for lung cancer diagnosis. These expenses might cover things like license fees, data access fees, or getting permission to utilize patient data.

Benefit-Cost Analysis

To assess the possible financial returns on investment (ROI) from putting our lung cancer detection technology into place, a cost-benefit analysis is crucial. Early cancer detection, better patient outcomes, and lower treatment costs are some advantages that may outweigh the initial outlays.

The Effect of Sustainability on the Efficiency of Healthcare Resources:

The project can help make better use of healthcare resources by offering a useful tool for detecting lung cancer. Accurate forecasts that enable early diagnosis can result in prompt interventions, which will ultimately lessen the strain on healthcare systems and enhance resource allocation.

Sustainability of the Environment:

By eliminating the need for substantial physical resources like paper-based records and manual reporting, the use of digital tools for lung cancer diagnosis can minimize waste. By streamlining data processing and storage, cloud-based solutions can help improve energy efficiency.

Long-Term Health Outcomes: By increasing lung cancer early detection rates, the study seeks to improve public health. Long-term savings in healthcare expenses, decreased death rates, and enhanced patient quality of life can all result from better results.

Community Involvement and Awareness: Raising community involvement in health screenings and preventative measures can result from raising awareness of lung cancer detection through our system. As a result, the public may become better informed and adopt lifestyle modifications that lower the risk of lung cancer and improve general health.

Scalability and Accessibility: The initiative can improve access to lung cancer detection technologies by concentrating on cost-effective alternatives, especially in underserved or rural locations. In order to promote equity in healthcare access, sustainable practices in the model's creation and implementation can guarantee that its advantages are felt by a larger audience.

3.7 Use of Standards

i. Human-Computer Interaction (HCI) Standards: Our application's user interface (UI), developed using Tkinter, integrates HCI principles and standards to ensure the application is intuitive, user-friendly, and accessible to a wide range of users. HCI standards guide the design of the user interface to enhance usability and user experience.

ii. Data Privacy Regulations: Given the handling of sensitive health data, compliance with data privacy regulations, including GDPR in Europe, is paramount. Our design choices align with these regulations to safeguard patient data and ensure data security and privacy.

iii. Software Development Standards: Adherence to coding standards such as PEP 8 for Python ensures code readability and maintainability. These standards have a positive impact on the organization and structure of our code, enhancing its quality and sustainability.

iv. Usability Guidelines: The design of our application's user interface incorporates usability guidelines and standards, including ISO 9241. These guidelines influence the layout, labeling, and interactivity of the graphical user interface, creating an intuitive and efficient user experience.

v. Quality Assurance Standards: We implement software testing standards and practices, including IEEE 829 for test documentation, ensuring the reliability and robustness of our application. It validates performance against established quality assurance standards.

vi. Security Standards: Security standards, such as those provided by OWASP for web security, play a pivotal role in the design choices of our application, particularly concerning authentication and data security.

vii. Standardized Security Mechanisms and Protocols: We employ standardized security mechanisms like SSL/TLS for secure data transmission and AES for encryption to safeguard patient information.

viii. Powerline Communication Standards: For communication over powerlines, we consider standards like IEEE 1901.2 to ensure reliable and compliant communication.

ix. Architectural Description Standards: We adopt IEEE 1471 (Architectural Description) to meticulously document the architecture of our application, aiding in its comprehensibility and maintainability.

x. Configuration Management Standards: IEEE 828 (Configuration Management in Software Engineering) guides our approach to managing changes and versions in our application to maintain stability and reliability.

xi. Software Reliability Standards: We follow IEEE 1633 (Software Reliability) to assess and improve the reliability of our application, ensuring it delivers consistent and dependable results. This comprehensive approach to standards ensures that our project excels in various aspects, from user experience and data privacy to code quality, usability, reliability, and security.

3.8. Experiment / Product Results (IEEE 1012 & IEEE 1633)

Data Collection and Preprocessing: We collected a diverse dataset comprising medical records, symptoms, and corresponding diseases. Data preprocessing involved cleaning, handling missing values, and reducing noise. The dataset was then split into training and testing sets.

CHAPTER-4

IMPLEMENTATION

4.Implementation

4.1 Environment Setup

To guarantee the smooth operation of our lung cancer classification models, we used a strong environment designed for data analysis and machine learning tasks in this project. Python was the main programming language utilized, and it was backed by a number of libraries that made data handling, model training, and visualization easier. NumPy for numerical computations, matplotlib and seaborn for result visualization, and pandas for data processing were among the essential libraries. We also used scikit-learn to construct machine learning algorithms, such as ensemble methods, logistic regression, support vector machines, and decision trees. Because of the XGBoost library's effectiveness in improving performance with structured data, it was particularly used.

Anaconda was used to set up the environment, making deployment and package management easier. Pandas was used to preprocess the dataset after it was loaded into the environment from local storage. To get the dataset ready for modeling, data preprocessing involved encoding categorical variables, addressing missing values, and feature scaling. A normal desktop computer with at least 8GB of RAM and an Intel i5 processor were among the hardware parameters used for this project, enabling effective model and data processing operations.

4.2 Sample Code for Preprocessing and MLP Operations

To guarantee the caliber and dependability of the input data for our machine learning models, the preprocessing stage was crucial. Several preprocessing procedures were performed on the dataset, which included a variety of variables pertaining to clinical data and patient demographics for lung cancer. Those included encoding the target variable, 'Level,' using scikit-learn's LabelEncoder and eliminating superfluous columns, such 'index' and 'Patient Id,' which don't aid in predictive modeling. Because it transforms categorical labels into a numerical format appropriate for model training, this transformation is essential.

```
from sklearn.neural_network import MLPClassifier
from sklearn.metrics import accuracy_score
# Initialize and train the MLP model
mlp_model = MLPClassifier(hidden_layer_sizes=(100, ), max_iter=500, random_state=42)
mlp_model.fit(X_train, y_train)

# Predictions and evaluation
y_pred = mlp_model.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy of MLP model:", accuracy)

from sklearn.metrics import confusion_matrix
import seaborn as sns
import matplotlib.pyplot as plt

# Confusion matrix visualization
conf_matrix = confusion_matrix(y_test, y_pred)
sns.heatmap(conf_matrix, annot=True, fmt='d', cmap='Blues',
            xticklabels=label_encoder.classes_, yticklabels=label_encoder.classes_)
plt.title('Confusion Matrix for MLP Model')
plt.ylabel('Actual')
plt.xlabel('Predicted')
plt.show()
```

CHAPTER-5

Experimentation and Result Analysis

5. Experimentation and Result Analysis

Using the lung cancer dataset, several machine learning models were trained during the experimentation phase, and their performance was assessed using a range of metrics. To determine how well each model predicted the severity of lung cancer, we methodically evaluated its accuracy, precision, recall, and F1 score.

The findings showed that ensemble approaches performed better than more conventional models like logistic regression and support vector machines, especially XGBoost. The model performed better because it was resilient against overfitting and could accommodate missing values. Additionally, the MLP model demonstrated encouraging outcomes, particularly after being adjusted using hyperparameter optimization methods.

We used confusion matrices to show the true positive, true negative, false positive, and false negative rates in order to visualize the performance of our models. This study shed light on the models' advantages and disadvantages by identifying instances of incorrect classification, especially in early-stage cancer diagnosis.

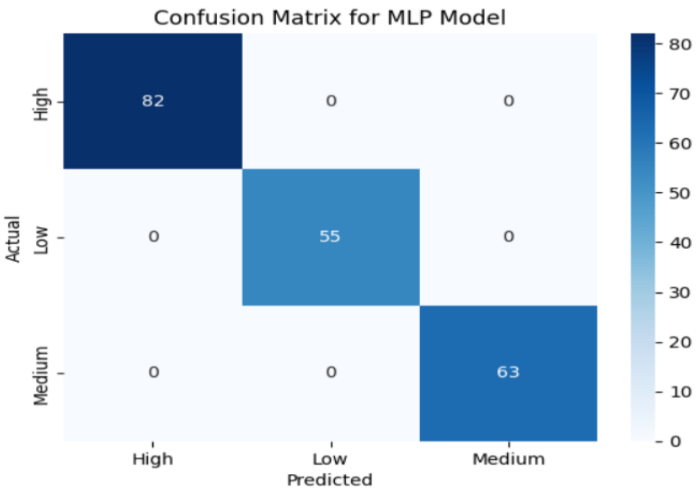


Figure 13. Confusion Matrix for MLP Model

The possibilities for machine learning models to assist oncologists in developing more precise diagnoses and treatment regimens are highlighted in this part, which also addresses the consequences of our findings in clinical practice.

CHAPTER-6

CONCLUSION

6.Conclusion

In Conclusion, this experiment highlights how machine learning approaches can improve lung cancer detection and therapy. We showed that algorithms like XGBoost and Multi-Layer Perceptron (MLP) can efficiently evaluate complicated clinical datasets and provide insightful predictions about patient outcomes by methodically putting different machine learning models into practice and assessing them. The findings show that in addition to achieving high accuracy, these models offer insights into the underlying patterns linked to the severity of lung cancer, which can help medical practitioners make well-informed judgments.

Even with our study's promising results, there are still a number of obstacles to overcome. The correctness and completeness of the data are essential for machine learning models to function well. Data in healthcare settings may have missing values or discrepancies and can originate from a variety of sources. Strong data management techniques and cooperation between researchers, data scientists, and healthcare professionals are needed to address these problems. Another major obstacle in clinical applications is the interpretability of machine learning models. Even though sophisticated algorithms are capable of producing precise forecasts, practitioners frequently find it challenging to comprehend the reasoning behind particular choices due to their complexity. Future research should concentrate on creating strategies to improve these models' interpretability and transparency so that medical practitioners can have confidence in and comprehend the insights they produce.

Combining genomic, transcriptomic, and proteomic data—also referred to as multi-omics data—represents a viable strategy for further research. These techniques could lead to more accurate predictions and a better understanding of the molecular mechanisms behind lung cancer by expanding the dataset. Furthermore, by testing model performance across a variety of populations, real-world data—such as patient registries and electronic health records—may enhance generalizability and therapeutic utility.

In summary, the results of this study show how machine learning has great promise for the study and management of lung cancer. These technologies have the potential to completely transform patient treatment as they develop further, improving survival rates and the quality of life for those who have lung cancer. In order to fully utilize machine learning and develop creative solutions that tackle the urgent problems associated with lung cancer diagnosis and treatment, data scientists and medical professionals must continue to collaborate.

REFERENCES

[1] S. Agarwal, S. Thakur, and A. Chaudhary, “Prediction of Lung Cancer Using Machine Learning Techniques and their Comparative Analysis,” *2022 10th Int. Conf. Reliab. Infocom Technol. Optim. (Trends Futur. Dir. ICRITO 2022*, pp. 1–5, 2022, doi: 10.1109/ICRITO56286.2022.9965052.

[2] P. Divya, R. Anuradha, and D. Palanivel Rajan, “Early Identification on Lung Cancer disease by using different ML Approaches,” *8th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2022*, vol. 1, pp. 1586–1591, 2022, doi: 10.1109/ICACCS54159.2022.9784969.

[3] D. D. Arka, S. M. Tafhim, R. M. Anan, N. Rahat, S. M. Ishan, and S. Tanvir, “Lung Cancer Detection Using Machine Learning Methods,” *Proc. 2023 IEEE Asia-Pacific Conf. Comput. Sci. Data Eng. CSDE 2023*, pp. 1–5, 2023, doi: 10.1109/CSDE59766.2023.10487685.

[4] S. Bharathy, R. Pavithra, and B. Akshaya, “Lung Cancer Detection using Machine Learning,” *Proc. - Int. Conf. Appl. Artif. Intell. Comput. ICAAIC 2022*, no. Icaaic, pp. 539–543, 2022, doi: 10.1109/ICAAIC53929.2022.9793061.

[5] R. H. Khan, J. Miah, S. A. A. Nipun, M. Islam, M. S. Amin, and M. S. Taluckder, “Enhancing Lung Cancer Diagnosis with Machine Learning Methods and Systematic Review Synthesis,” *ICEEIE 2023 - Int. Conf. Electr. Electron. Inf. Eng.*, pp. 1–5, 2023, doi: 10.1109/ICEEIE59078.2023.10334739.

[6] G. Sruthi, C. L. Ram, M. K. Sai, B. P. Singh, N. Majhotra, and N. Sharma, “Cancer Prediction using Machine Learning,” *Proc. 2nd Int. Conf. Innov. Pract. Technol. Manag. ICIPTM 2022*, vol. 2, pp. 217–221, 2022, doi: 10.1109/ICIPTM54933.2022.9754059.

[7] K. Shreya, J. Gopalakrishnan, R. Sivayogitha, S. Vijayabaskar, S. Hariharan, and V. Kukreja, “Lung Cancer Analysis using Machine Learning Approach,” *2nd Int. Conf. Autom. Comput. Renew. Syst. ICACRS 2023 - Proc.*, no. Ml, pp. 736–740, 2023, doi: 10.1109/ICACRS58579.2023.10404078.

[8] M. Singh, C. Shah, and P. Patel, “Lung Cancer Prediction Using Machine Learning Models,” *Lect. Notes Networks Syst.*, vol. 765 LNNS, no. Icimia, pp. 613–618, 2023, doi:

10.1007/978-981-99-5652-4_54.

[9] A. Indumathi, M. Sathanapriya, N. Vinodh, M. Ashok, and N. Aishwarya, “Machine Learning based Lung Cancer Detection & Analysis,” *Int. Conf. Sustain. Comput. Smart Syst. ICSCSS 2023 - Proc.*, no. Icscss, pp. 361–365, 2023, doi: 10.1109/ICSCSS57650.2023.10169329.

[10] A. Alomar, M. Alazzam, H. Mustafa, and A. Mustafa, “Lung Cancer Detection Using Deep Learning and Explainable Methods,” *2023 14th Int. Conf. Inf. Commun. Syst. ICICS 2023*, pp. 1–4, 2023, doi: 10.1109/ICICS60529.2023.10330443.

[11] A. A. Hasan, A. T. A. Salih, and A. Ghandour, “Lung Cancer Detection using Evolutionary Machine learning and Deep learning: A survey,” *5th Int. Conf. Inf. Technol. Appl. Math. Stat. ICITAMS 2023*, pp. 129–133, 2023, doi: 10.1109/ICITAMS57610.2023.10525500.

[12] S. Kukreja, M. Sabharwal, and D. S. Gill, “A Survey of Machine learning algorithms for Lung cancer detection,” *Proc. - 2022 4th Int. Conf. Adv. Comput. Commun. Control Networking, ICAC3N 2022*, pp. 338–342, 2022, doi: 10.1109/ICAC3N56670.2022.10074272.

[13] M. Jaeyalakshmi, P. K. Janani, P. J. Priya, M. Bhavani, and K. E. Narayanan, “Detection of Lung Cancer Using Deep Learning Model and Radiomics Method,” *2024 Int. Conf. Commun. Comput. Internet Things, IC3IoT 2024 - Proc.*, 2024, doi: 10.1109/IC3IoT60841.2024.10550376.

[14] S. Murthy Nimmagadda, K. Likhitha, G. Srilatha, and S. M. Sree, “Lung Cancer Prediction and Classification Using Machine Learning Algorithms,” *Proc. - 2024 Int. Conf. Expert Clouds Appl. ICOECA 2024*, pp. 1012–1015, 2024, doi: 10.1109/ICOECA62351.2024.00176.

[15] H. Kasaudhan, K. K. Shukla, R. Kushwaha, K. Sharma, U. Gupta, and A. Sharma, “Early Detection and Analysis of Lung Cancer Using Artificial Intelligence,” *Proc. - Int. Conf. Comput. Power, Commun. Technol. IC2PCT 2024*, vol. 5, pp. 1470–1474, 2024, doi: 10.1109/IC2PCT60090.2024.10486335.

[16] T. Singh, B. Regmi, S. B. Jadhav, and S. Singh, “Early Stage Lung Cancer Detection Using Deep Learning,” *2024 MIT Art, Des. Technol. Sch. Comput. Int. Conf. MITADTSoCiCon 2024*, pp. 1–6, 2024, doi: 10.1109/MITADTSoCiCon60330.2024.10575345.

[17] V. A. Binson and M. Subramoniam, “Advances in Early Lung Cancer Detection: A Systematic Review,” *2018 Int. Conf. Circuits Syst. Digit. Enterp. Technol. ICCSDET 2018*, pp. 1–5, 2018, doi: 10.1109/ICCSDET.2018.8821188.

[18] N. Nawreen, U. Hany, and T. Islam, “Lung cancer detection and classification using CT scan image processing,” *2021 Int. Conf. Autom. Control Mechatronics Ind. 4.0, ACMI 2021*, vol. 0, no. July, pp. 1–6, 2021, doi: 10.1109/ACMI53878.2021.9528297.

[19] B. Meylia *et al.*, “Determining the Main Symptoms of Lung Cancer with Machine Learning Methods,” *10th Int. Conf. ICT Smart Soc. ICISS 2023 - Proceeding*, pp. 1–6, 2023, doi: 10.1109/ICISS59129.2023.10291539.

[20] O. Khouadja and M. S. Naceur, “Lung Cancer Detection with Machine Learning and Deep Learning: A Narrative Review,” *Proc. 2023 IEEE Int. Conf. Adv. Syst. Emergent Technol. IC_ASET 2023*, pp. 1–8, 2023, doi: 10.1109/IC_ASET58101.2023.10150913.