A FIELD PROJECT REPORT

on

**“****A Multi-Algorithm Approach for Lung Cancer Detection: SVM, GBM, Random Forests, and Logistic Regression”**

**Submitted by**

221FA04116 N.vijaya Lakshmi

221FA04443 D.Bala chandu

221FA04610 A.Deepthi

221FA04666 B.Sruthi

**Under the guidance of**

Dr.DEVA KUMAR



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

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

**Vadlamudi, Guntur.**

**ANDHRA PRADESH, INDIA, PIN-522213.**



**CERTIFICATE**

This is to certify that the Field Project entitled **“****A Multi-Algorithm Approach for Lung Cancer Detection: SVM, GBM, Random Forests, and Logistic Regression”** that is being submitted by 221FA04116 (N.VIJAYA LAKSHMI) , 221FA04443 (D.BALA CHANDU), 221FA04610(A.DEEPTHI), 221FA04666 (B.SRUTHI) for partial fulfilment of Field Project is a bonafide work carried out under the supervision of Dr. Deva Kumar, Assistant Professor, Department of CSE.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Guide name& Signature |  | | Dr.K.V. Krishna Kishore | |
| Assistant/Associate/Professor, CSE | HOD,CSE | | Dean, SoCI | |
|  |  |  | |



**DECLARATION**

We hereby declare that the Field Project entitled **“ A Multi-Algorithm Approach for Lung Cancer Detection: SVM, GBM, Random Forests, and Logistic Regression ”** is being submitted by 221FA04116 (N.VIJAYA LAKSHMI) , 221FA04443 (D.BALA CHANDU), 221FA04610(A.DEEPTHI), 221FA04666 (B.SRUTHI) 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 Dr. Deva Kumar,Assistant Professor, Department of CSE.

By

221FA04116 (N.Vijaya lakshmi)

221FA04443(D.Bala chandu)

221FA04610(A.Deepthi)

221FA04666 (B.Sruthi)

Date:

## ABSTRACT

Lung cancer remains a significant global health challenge due to its high mortality rate and the need for early detection. This study presents a multi-algorithm approach utilizing Support Vector Machine (SVM), Gradient Boosting Machine (GBM), Random Forests, Logistic Regression and naïve bayes to improve lung cancer detection accuracy. These machine learning algorithms were applied to a dataset sourced from the Kaggle, using k-fold cross-validation for model evaluation. The results indicate that ensemble methods, particularly GBM and Random Forests, demonstrate superior performance in classifying lung cancer compared to individual classifiers. This paper highlights the importance of machine learning in enhancing the diagnostic process for lung cancer, thereby aiming to contribute to better patient outcomes through early detection.

**Keywords:**

* Lung Cancer Detection, Support Vector Machine (SVM), Gradient Boosting Machine (GBM), Random Forests, Logistic Regression, Naïve bayes, Learning, Early Detection, Classification Algorithms, Ensemble Methods, Imaging.

**TABLE OF CONTENTS**

1. [Introduction 1](#_bookmark0)
2. [Literature Survey](#_bookmark5) 3
   1. [Literature review 4](#_bookmark6)
   2. [Motivation 5](#_bookmark7)
3. [Proposed System 6](#_bookmark8)
   1. [Input dataset 7](#_bookmark9)
   2. [Data Pre-processing 7](#_bookmark10)
      1. Data collection  [7](#_bookmark11)
      2. [Data cleaning 7](#_bookmark13)
      3. Outlier detection and treatment 8
      4. Feature engineering 8
      5. Handling imbalanced data 8
      6. Feature scaling 8
      7. Data splitting 8
   3. [Methodology of the system 9](#_bookmark16)
   4. Design specification 10
   5. [Model Evaluation 11](#_bookmark17)
4. [Implementation 13](#_bookmark19)
5. Experimentation and Result Analysis 16
6. Conclusion 22
7. References 24

**LIST OF FIGURES**

Figure 1:Lung cancer system Architecture 12

Figure 2: confusion matrix for decision tree 17

Figure 3: Lung Cancer 18

Figure 4: Training Data and Testing Data 19

Figure 5: DecisionTreeClassifier 20

Figure 6: Receiver Operating Characteristics(ROC) 20

Figure 7: Accuracy Comparison of Classification Techniques 21

**LIST OF TABLES**

Table 1. Detailed features of dataset 7

Table 2. Accuracy evaluation 18

# CHAPTER-1

# INTRODUCTION

**1.INTRODUCTION**

In the past few years, the detection of lung cancer has been widely discussed as its mortality rate is high and early diagnosis is very necessary. Lung cancer is one of the major killing diseases from cancer, and it occurs globally. Performance Analysis and Classification of Lung Cancer types Using Machine learning[1]. It normally starts in the lungs with abnormal cell growth to create tumors which may spread to other parts of the body. Among cancer forms, lung cancer is one of the most prevalent and deadliest in the world. Disease is very complex in nature and therefore, disease handling must be properly done[2]. Patients with lung cancer suffer from the symptoms like Bloody coughing, wheezing, dyspnea, weakness, exhaustion and Pancoast disease[3]. The most common symptom of lung cancer is coughing, then the respiratory symptomps that is chest pain, dyspnea, and hemptysis[4].Lung cancer remains to be a significant global health concern, where there's need to continue improving the early detection methodologies to increase survival rates from the disease condition largely[5]. This is by the application of machine learning techniques, namely ANN, SVM, and Bayesian networks, which frequently predict the prognosis of cancer [6]. Microarray data have also been brought into the picture to discover many cases using the ensemble method[6].The worldwide incidence of Lung cancer is the second most common disease among men and women, which accounts for nearly 2.09 million new detections yearly, and it causes the most death, which is about 1.76 million, according to the WHO[7]. According to WHO, the people suffering from lung cancer are now 1.8 million in 2012 and 2.09 million in 2018[8]. In our project, we propose an automatic lung cancer detection system based on machine learning algorithms[9]. Because of the large quantity of data we keep in our database, Machine Learning is a preferable method to analyze it because ML algorithms learn very fast with a large number of patients. Ensemble methods are one of the subsets of ML techniques and is used for achieving better accuracies in classifier learning systems. In ensemble methods, unlike traditionally used ML approaches, a number of poor classifiers are combined together to form a better one[10]. This is a project that could demonstrate the capacity of machine learning algorithms to enhance the accuracy and efficiency of detecting lung cancer by analyzing medical data and pointing out hard-to-detect patterns. To address the concerns raised above, this paper presents four widely applied machine learning algorithms: Support Vector Machine (SVM), Gradient Boosting Machine (GBM), Random Forests, and Logistic Regression, to estimate the existence of lung cancer. Several weaknesses in only relying on one machine learning algorithm for the detection of lung cancer are that detection accuracy is relatively low and there are a number of false negatives. "The paper, titled 'Strategic Integration of CNN, SVM, and XGBoost for Early-stage Tumor Detection using Hybrid Deep Learning Method', has taken the bests from the Convolutional Neural Networks, SVM, and Extreme Gradient Boosting. The detection of tumor at early stages is difficult because the early pathological changes are very subtle and the task of processing in the medical imaging data is of high complexity. Traditional image analysis techniques have a poor sensitivity and specificity that corresponds to the time of diagnosis, [11]. Pick up a subset of individual leaning models with outstanding performance from some of the established models, like support vector machines, decision trees, random forests, logistic regression, and regression using K-nearest neighbors[12]. Every dataset challenges the researchers to test their system's performance and enhance accordingly for better automated system for the same[13].SVM is well known for high dimensional data and achieves good performances on complex, non-linear classification tasks. GBM is a very powerful ensemble technique capable of building models sequentially in order to optimize prediction performance. The ensemble method, Random Forests, does exceptionally well in controlling overfitting and generalizing by summing the outputs across multiple decision trees. Logistic Regression is also one of the classic statistical methods and proved highly popular for two-class classification problems, besides providing very interpretable results in terms of probability.

This can be done through modalities such as computed tomography, positron emission tomography, and X-rays, though at an expensive and time-consuming cost. Therefore, a new computer aided system to enhance the early detection of lung cancer has become urgent in the sobering landscape[14]. This is to combine the use of diversified FE techniques with pretraining models and ML to enhance the precision and reliability [15]. Lung cancer has touched the life of almost every family and there is rapidly increasing population with lung malignancy. Aside from this, it is also caused by some major problems like jaundice, swelling of lymph nodes, and problems related to the nervous system[8]. Lung Cancer datasets that we used for this study are extracted from UCI Machine Learning Repository and Data World. Firstly, the sets of given data are partitioned into training and test data sets by using k-fold cross validation technique. Then using the classification algorithms such as SVM, Logistic Regression, Naïve Bayes and Decision Tree, respective classification models are implemented using the above training data [9]. With the help of the machine learning, the detection of lung cancer is easy and results in the increase of life expectancy for humans suffering from lung cancer [16]. The key issue of these techniques is that the detection was limited to solitary pulmonary nodules. These approaches also suffered from sensitivity problem. Furthermore, all the work done solely for lung cancer detection and not for prediction[17]. With proper data collection, preprocessing and visualization, we lay a foundation for our automatic lung cancer detection system[9].

# 

# CHAPTER-2

# LITERATURE SURVEY

## 2.1 LITERATURE SURVEY

Over the past few decades, a number of machine learning (ML) algorithms have played a very transformative role in the early detection and diagnosis of lung cancer through analysis of medical imaging, clinical data, and genomic information. The cancer is called malignant tumor due to irregular division of tissue or organ in the cells. With the massive variety in many cancers of males and females alike. The deep learning models have been used very generally for cancer detection in past few decades[18].The authors in [9] proposes a classification system of lung cancer as benign or malignant. The system makes use of priori information and Hounsfield unit (HU) to determine the Region of Interest (ROI); various shape features and textural features are extracted to train and classify the support vector machine for judging whether the nodule is benign or malignant[19]. The paper briefly presents some of the major studies of lung nodule detection using Support Vector Machine (SVM), Gradient Boosting Machine (GBM), Random Forests, and Logistic Regression. Lung cancer is a tough and dangerous cancer. It leads to many deaths in men and women. Decision tree applies the supervised learning technique where a model is build which is in the form of a data structure in the set of nodes arranged in hierarchical fashion. The entropy of parent is calculated first. Then, Information gain is calculated by taking weighted difference of entropy of children from the entropy of parent. The one with highest Information gain is considered as the root node and the process goes on till classification is done[9].Various methods have been used to detect lung cancer in an early stage. Here in this paper, some algorithms have been applied first in principal compared analysis, KNN, SVM, Machine learning techniques to detect[12]. This literature review covers four of the most popular algorithms including Support Vector Machine (SVM), Gradient Boosting Machine (GBM), Random Forests, and Logistic Regression used in lung cancer detection and their specific applications, respective merits, and comparative performance in such a context.

Various works have experimented using different classifiers, including support vector machines [random forests, logistic regression, k-nearest neighbor, artificial neural networks, and decision trees, over a variety of datasets. Other works have also approached the use of ensemble techniques and feature selection strategies to enhance the accuracy of their predictions[20]. A method of a machine learning technique was presented for the detection of lung cancer using Marker-Controlled Watershed, Gabor filter for segmentation and the enhancement of image, where 10,800 images of lung CT scans were evaluated for 50 patients of VIAELCAP dataset. The features are obtained by applying GLCM (Gray-Level Co-occurrence Matrix). Categories in normal and abnormal images are done through SVM, RBF and MLP (multilayer perceptron)[19]. SVM is a highly powerful supervised learning algorithm that has been adapted in most diagnostic medical systems, including lung cancer detection. It works by finding an optimum hyperplane that separates the two completely different classes-cancerous and non-cancerous tissue. GBM is a robust ensemble learning technique that combats the use of weak learners to achieve more impressive predictive modeling. Its robustness and ability to handle complex datasets made it so extensively used for cancer detection. Another ensemble learning technique, Random Forests, is often applied to classification and regression problems in medical applications as it does not overfit and can also manage big data with complex datasets. Logistic Regression is a very popular statistical method in dealing with the class-imbalanced problem of cancer diagnosis for easy interpretation and with high efficiency.

Some also investigated ensemble methods together with feature selection techniques in order to enhance the prediction performance. There is a lack of evidence that classifiers have consistently performed better in some cases, and the preprocessing of data, which includes feature scaling, forms a prerequisite before quality experimental results are produced[20]. A number of studies have compared the performances of these machine learning models in lung cancer detection:Jiang et al. (2020) had compared SVM, GBM, Random Forests, and Logistic Regression on a large dataset of lung cancer patients. This means that the ensemble methods, specifically GBM and Random Forests, performed relatively better than individual classifiers like SVM and Logistic Regression, mainly on complex, high-dimensional data. Still, when it came to using a logistic regression, the right use case persisted in the interpretation. Ahmed et al., 2022 have compared the performance of SVM, Random Forests, and Logistic Regression on patient health records and imaging data. The study concluded that the precision of the performance of Random Forests and SVM is higher compared to Logistic Regression; however, it may turn out to be a more important tool in understanding the kind of relationship different patient characteristics have with lung cancer.

Challenges and Future Directions

The remarkable potential shown by machine learning in detecting lung cancer must not be forgotten, however, against this backdrop are also the challenges that have to be addressed which include:

• Data Availability and Quality: many models tend to depend on high-quality, annotated medical images or large clinical datasets, which are unavailable at most places.

•Model Interpretability: Models such as Random Forests and GBM may be very accurate, but this may come at a cost of high complexity. As such, their results may be difficult to interpret, thereby limiting their application in clinical environments.

•Generalization: Conferring on models of machine learning the ability to general well for new patient populations or for other medical settings is still challenging. Models must be rigorously validated against diverse datasets so that results prove to be robust.

#### 2.2 Motivation

Lung cancer is one of the leading causes of cancer-related deaths worldwide, primarily due to late diagnosis and limited effective treatment options. The high mortality rate associated with this disease underscores the urgent need for innovative solutions to enhance early detection and diagnosis. Machine learning offers powerful tools that can analyze complex medical data and identify patterns not easily discernible by traditional methods. By leveraging these advanced techniques, we can improve diagnostic accuracy, reduce the time to treatment, and ultimately enhance patient outcomes. This research aims to explore the potential of various machine learning algorithms in lung cancer detection, contributing to the growing body of knowledge that seeks to harness technology for better healthcare solutions.

# CHAPTER-3

# PROPOSED SYSTEM

#### 3.1 Input dataset

The dataset for detecting lung cancer consists of 309 rows and 16 columns.

* The dataset we will be using have these columns –

|  |  |
| --- | --- |
| **FEATURE** | **DESCRIPTION** |
| GENDER | The classification of individuals as male or female, which may influence lung cancer risk and symptoms. |
| AGE | A key demographic factor, as lung cancer incidence increases with age. |
| SMOKING | The act of inhaling tobacco smoke, a major risk factor for developing lung cancer. |
| YELLOW\_FINGERS | A common sign of smoking, indicating nicotine staining on fingers. |
| ANXIETY | A psychological condition that can affect overall health and may be linked to cancer symptoms. |
| PEER\_PRESSURE | The influence from friends or acquaintances that may lead individuals to adopt risky behaviors, such as smoking |
| CHRONIC DISEASE | Long-term health conditions that can complicate lung cancer diagnosis and treatment. |
| FATIGUE | A persistent feeling of tiredness that may be associated with cancer or its treatment. |
| ALLERGY | An immune response to allergens, which can complicate respiratory symptoms in lung cancer patients. |
| WHEEZING | A high-pitched sound during breathing, often indicative of airway obstruction or lung issues. |
| ALCOHOL CONSUMING | The intake of alcoholic beverages, which may impact lung cancer risk and overall health. |
| COUGHING | A common symptom that can be associated with lung cancer or other respiratory conditions. |
| SHORTNESS OF BREATH | A sensation of difficulty in breathing, often linked to lung diseases, including cancer. |
| SWALLOWING DIFFICULTY | Trouble in swallowing, which can occur due to lung cancer's effect on surrounding structures |
| CHEST PAIN | Discomfort in the chest area that can be a symptom of lung cancer or other medical conditions. |
| LUNG\_CANCER | A type of cancer that originates in the lungs, characterized by abnormal cell growth that can disrupt normal lung function. |

Table 1: Features of dataset

#### Data Pre-processing

Data preprocessing is a crucial step in preparing the dataset for machine learning models. Here are some common data preprocessing techniques that are used in detecting Lung cancer :

**3.2.1. Data Collection**

Lung cancer Data: we focused on collecting information related to patient demographics, clinical history, and health behaviors that may influence the disease's development. This includes features such as patient ID, age, gender, smoking history, symptoms (e.g., coughing, shortness of breath), medical history (e.g., chronic diseases, allergies), and treatment history. Additionally, we recorded factors like anxiety levels, peer pressure, and lifestyle choices (e.g., alcohol consumption). The aim was to compile a comprehensive dataset that can help identify patterns and risk factors associated with lung cancer, thereby facilitating improved diagnostics and treatment options.

**3.2.2. Data Cleaning**

Handling Missing Values: Check for missing or null values in the dataset. Missing values can be handled using:

Imputation: Filling missing values with mean, median, or mode.

Dropping Missing Values: In some cases, rows with missing values may need to be removed if they are too sparse.

Removing Duplicates: Ensure that there are no duplicate entries for transactions, which can distort the model’s learning.

**3.2.3. Outlier Detection and Treatment**

 **Identification of Outliers**: Outliers can be detected using statistical methods such as Z-score analysis, where data points with a Z-score greater than 3 or less than -3 are considered outliers. Alternatively, the Interquartile Range (IQR) method can be used, where data points lying outside 1.5 times the IQR from the first and third quartiles are flagged as outliers.

 **Visual Inspection**: Visualization techniques, such as box plots, scatter plots, or histograms, can help in visually identifying outliers and understanding their distribution relative to the rest of the data.

 **Treatment of Outliers**: Once identified, outliers can be treated in several ways:

* **Removal**: Outliers can be removed from the dataset if they are determined to be errors or noise.
* **Transformation**: Applying transformations (e.g., log transformation) to reduce the impact of extreme values while retaining them in the dataset.
* **Imputation**: Replacing outlier values with statistical measures such as the mean, median, or mode of the relevant feature, if it makes sense contextually.

 **Re-evaluation**: After treating outliers, it is important to re-evaluate the dataset to ensure that the changes have improved the model's performance and that no significant information has been lost in the process.

**3.2.4. Feature Engineering**

**Domain-Specific Features:** Derive new features from medical knowledge, such as smoking duration, severity of symptoms, or family history of respiratory diseases.

**Interaction Features:** Create features representing interactions between variables, such as the correlation between shortness of breath and coughing, or chronic diseases and fatigue.

**3.2.5. Handling Imbalanced Data**

**Handling Imbalanced Data in Cancer Datasets:**

Apply techniques such as Synthetic Minority Oversampling Technique (SMOTE) or undersampling to balance lung cancer-positive and negative cases within the dataset.

**3.2.6. Feature Scaling**

 **Standardization (Z-score Normalization)**: Rescales features to have a mean of zero and a standard deviation of one, ensuring equal contribution from all features, especially beneficial for sensitive algorithms like logistic regression and clustering.

 **Min-Max Scaling**: Transforms features to a specific range (typically 0 to 1), improving convergence in deep learning models by maintaining relative differences between feature values.

 **Robust Scaling**: Utilizes the median and interquartile range to reduce the influence of outliers, making it effective for datasets where certain features exhibit extreme values.

 **Log Transformation**: Stabilizes variance and normalizes skewed distributions, particularly useful for features like tumor sizes or biomarker levels, enhancing analysis and model performance in cancer research.

**3.2.7. Data Splitting**

Perform a stratified train-validation-test split to ensure that each subset has a similar distribution of lung cancer cases, maintaining the balance between positive and negative cases.

 **Training Set**: Used to fit the model on the features related to lung cancer, such as symptoms, demographics, and lifestyle factors, enabling the model to learn the underlying patterns in the data.

 **Validation Set**: Employed for hyperparameter tuning and model selection, helping to optimize the model’s performance by assessing how well it generalizes to data not seen during training.

 **Test Set**: A "holdout" set reserved for evaluating the final model's performance, providing an unbiased estimate of how well the model predicts lung cancer outcomes on unseen data.

 **Stratified Splitting**: Ensures that the same proportion of positive (lung cancer) and negative (non-cancer) cases are represented in the training, validation, and test sets, maintaining the balance of classes in each subset.

#### 3.3Methodology of the system

1. Set Up Libraries and Environment: Install necessary libraries and set up the environment for data preprocessing and demonstration in Jupyter.

2. Load the Dataset: Import the Lung Cancer dataset from Kaggle.

3. Data Cleaning: Clean the dataset by addressing missing values, outliers, and anomalies, as well as converting Lung Cancer from categorical labels to numerical values.

4. Split the Dataset: Divide the dataset into training and testing sets to evaluate model performance. Utilize a Random Forest classifier to train the model.

5. Model Evaluation: Assess the performance of the trained model on the testing dataset using metrics such as accuracy, precision, recall, and F1 score. Analyze the confusion matrix to understand the model’s ability to distinguish between whether the person is having cancer or not.

**How Each Model Works for Lung Cancer Dataset:**

1. **Support Vector Machine (SVM):**
   * **How it works:** SVM classifies data by finding the hyperplane that best separates the data points into different classes, maximizing the margin between lung cancer-positive and negative cases. It is effective in high-dimensional spaces, which makes it suitable for complex medical data like lung cancer diagnosis.
   * **In this context:** SVM will handle the binary classification problem of predicting whether a patient has lung cancer based on features such as symptoms, lifestyle factors (e.g., smoking), and medical history.
2. **Gradient Boosting Machine (GBM):**
   * **How it works:** GBM builds an ensemble of decision trees in a sequential manner. Each tree corrects the errors of the previous one by focusing on the data points that were misclassified. It excels at capturing complex relationships and interactions between features.
   * **In this context:** GBM will help in accurately predicting lung cancer by learning from errors of previous trees and focusing on hard-to-classify patients (those with borderline symptoms).
3. **Decision Tree:**
   * **How it works:** The decision tree algorithm uses a series of questions to split the dataset into smaller subsets based on feature values (e.g., "Does the patient have a chronic disease?"). It creates a tree-like structure where each node represents a decision rule based on one feature, leading to a classification (lung cancer-positive or negative).
   * **In this context:** The decision tree will provide interpretable rules for classifying patients, which is useful in a medical setting to understand how different features (e.g., smoking, chest pain) lead to a lung cancer diagnosis.
4. **Random Forest:**
   * **How it works:** Random Forest is an ensemble method that builds multiple decision trees using random subsets of the data and features. Each tree makes a prediction, and the final classification is determined by a majority vote across the trees. It reduces the risk of overfitting.
   * **In this context:** Random Forest will provide a robust and reliable classification by averaging predictions from multiple trees, making it less prone to errors compared to a single decision tree. It will also provide feature importance insights, showing which factors (e.g., wheezing, shortness of breath) contribute most to lung cancer risk.
5. **Naive Bayes:**
   * **How it works:** Naive Bayes is a probabilistic classifier that assumes independence between features. It applies Bayes’ Theorem to compute the probability of a patient having lung cancer based on their symptoms and other features.
   * **In this context:** Despite its simplicity, Naive Bayes can provide fast and efficient predictions, especially when dealing with binary features such as smoking (yes/no), coughing (yes/no), and alcohol consumption (yes/no). It’s particularly useful when speed is more important than highly complex model accuracy.

**Hybrid Model Using Stacking:**

* **Stacking Approach:**
  + In this hybrid model, we combine the strengths of the above classifiers through a stacking approach. In the first layer, all the base classifiers (SVM, GBM, Decision Tree, Random Forest, Naive Bayes) are trained on the lung cancer dataset. Each of these models will make predictions about whether a patient has lung cancer.
  + These predictions are then used as input to a meta-learner, which is a **Decision Tree** in this case, that makes the final classification. The decision tree meta-learner leverages the predictions of the base models to make a more accurate and reliable final decision.
* **Why Stacking Works:**
  + **Diverse Models:** By combining diverse models, the hybrid model reduces the individual weaknesses of each classifier. For example, SVM might struggle with noisy data, but Random Forest can handle it well, and GBM can capture complex patterns that Naive Bayes might miss.
  + **Improved Accuracy:** The decision tree meta-learner aggregates predictions from the base models, leading to better overall performance by using their collective knowledge to boost accuracy in lung cancer prediction.
  + **Robust Predictions:** Stacking improves the robustness of predictions, especially for hard-to-classify patients, by integrating multiple perspectives on the data from different models.

**3.4 DESIGN SPECIFICATION**

This design specification outlines the architecture and implementation of a lung cancer prediction system that utilizes a stacking-based hybrid model incorporating multiple machine learning classifiers (SVM, GBM, Decision Tree, Random Forest, and Naive Bayes) to enhance prediction accuracy.

**1. System Overview:**

The system is designed to predict lung cancer in patients using a combination of machine learning models. It integrates several classifiers to handle a variety of patient data features such as demographic, medical history, lifestyle, and symptoms, and aggregates the predictions through a meta-learner using a stacking approach.

**2. Data Input:**

* **Source:** The input data includes a variety of patient features such as:
  + Demographic details (Gender, Age)
  + Lifestyle factors (Smoking, Alcohol consumption)
  + Symptoms (Coughing, Shortness of breath, Chest pain, etc.)
  + Medical history (Chronic diseases, Wheezing, Yellow fingers, Fatigue)
* **Format:** Data is structured in a tabular format, with rows representing patients and columns representing features.

**3. Preprocessing:**

* **Data Cleaning:** Handle missing values, remove outliers, and normalize data where necessary.
* **Feature Engineering:** Create additional features such as interaction terms between chronic disease and smoking, or the combined effect of chest pain and shortness of breath.
* **Handling Imbalanced Data:** Techniques such as SMOTE (Synthetic Minority Over-sampling Technique) are used to handle any imbalance between positive (lung cancer) and negative (non-lung cancer) cases.

**4. Model Architecture:**

The system is built on a stacked generalization framework, which consists of two layers:

**Layer 1: Base Classifiers**

* **Support Vector Machine (SVM):**
  + Separates lung cancer and non-lung cancer cases by constructing a hyperplane with maximum margin.
* **Gradient Boosting Machine (GBM):**
  + Sequentially builds decision trees, focusing on misclassified cases to capture complex feature interactions.
* **Decision Tree:**
  + Uses a tree-based structure to classify patients based on symptoms and medical history.
* **Random Forest:**
  + An ensemble of decision trees that uses random feature subsets and bootstrapped samples for classification, providing robust and diverse predictions.
* **Naive Bayes:**
  + Applies Bayes’ Theorem to compute the probability of lung cancer based on individual features, assuming independence between features.

**Layer 2: Meta-Learner**

* **Meta-Model (Logistic Regression):**
  + Takes the predictions from the base classifiers as input and learns to combine them to make the final prediction. Logistic Regression is often used as the meta-learner because of its simplicity and effectiveness in binary classification tasks.

**5. Training and Validation:**

* **Data Splitting:** The dataset is split into training, validation, and test sets using stratified sampling to ensure the same proportion of lung cancer and non-lung cancer cases in each set.
* **Cross-Validation:** K-fold cross-validation is applied to ensure that the models generalize well and are not overfitting to specific data.
* **Hyperparameter Tuning:** Each base model is tuned using the validation set to find optimal hyperparameters (e.g., kernel type for SVM, number of trees for Random Forest, learning rate for GBM).

**6. Evaluation Metrics:**

* **Accuracy:** Measures overall correctness of predictions.
* **Precision & Recall:** Evaluates the model’s ability to correctly identify lung cancer cases (recall) and avoid false positives (precision).
* **F1-Score:** Balances precision and recall to provide a single metric for model performance.
* **ROC-AUC:** Measures the ability of the model to discriminate between lung cancer and non-lung cancer cases.

**7. Model Deployment:**

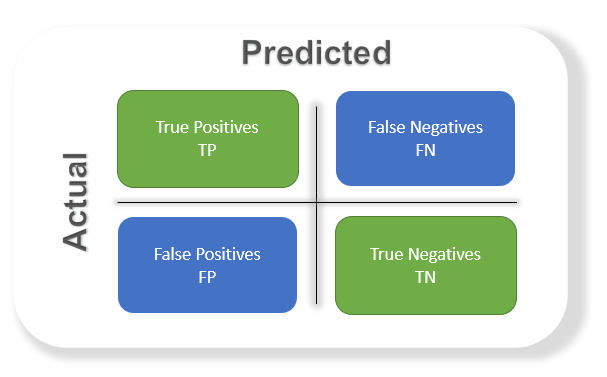
* **Platform:** The system is deployed using a cloud-based platform to handle predictions in real-time.
* **Scalability:** Designed to scale as more patient data is collected and analyzed.
* **User Interface:** A simple, user-friendly interface is provided for healthcare professionals to input patient data and receive lung cancer predictions with explanations of the contributing factors.

**8. System Outputs:**

* **Prediction:** The system outputs a binary prediction (0 for non-lung cancer, 1 for lung cancer).
* **Confidence Score:** Provides the probability of lung cancer based on the meta-learner's decision.
* **Feature Importance:** Highlights which patient features (e.g., chronic disease, smoking) were most influential in the prediction.

**9. Advantages of Stacking Model:**

* **Robustness:** By combining diverse models, the system reduces the individual weaknesses of each classifier.
* **Improved Accuracy:** The hybrid model uses the strengths of each base classifier to enhance overall prediction performance.
* **Flexibility:** The stacking approach can easily incorporate additional models in the future, allowing for further optimization and improvements.



#### 3.5Model Evaluation

**Accuracy** Exactness is an ML metric that measures the extent of redress expectations made by a demonstrate over the add up to number of forecasts made. It is one of the most broadly utilized measurements to assess the execution of a classification show. The ratio of correctly predicted instances to the total instances. Suitable for balanced datasets.

**Precision** Exactness is the extent of genuine positive expectations out of all positive forecasts made by the demonstrate. It essentially measures the exactness of positive expectations. The ratio of true positive predictions to the total predicted positives. Useful in cases where false positives are costly.

**Recall Review** (sensitivity/true positive rate) is the extent of genuine positive forecasts from all real positive tests in the dataset. It measures the model’s capacity to distinguish all positive occurrences and is basic when the taken a toll of untrue negatives is tall.

**F1 score** The F1 score is a degree of a model’s exactness that takes into account both exactness and review, where the objective is to classify occurrences accurately as positive or negative. The harmonic mean of precision and recall, providing a balance between the two metrics. Useful for imbalanced datasets.

**ROC-AUC Score** The area under the Receiver Operating Characteristic curve. AUC measures the model's ability to distinguish between classes.

Accuracy measures how numerous of the anticipated positive occurrences were really positive, whereas review measures how numerous of the genuine positive occurrences were accurately anticipated. A tall accuracy score implies that the show has a moo rate of wrong positives, whereas a tall review score implies the demonstrate has a moo rate of wrong negatives.

**Confusion Matrix**

* **Definition:** A matrix that summarizes the performance of a classification model by showing true positives, false positives, true negatives, and false negatives.
* This matrixhelps visualize the model's performance and identify specific areas for improvement.

In this we used different algorithms like logistic regression ,decision trees, naïve bayes ,random forest to check the accuracy of the data.

**Lung Cancer prediction system:**

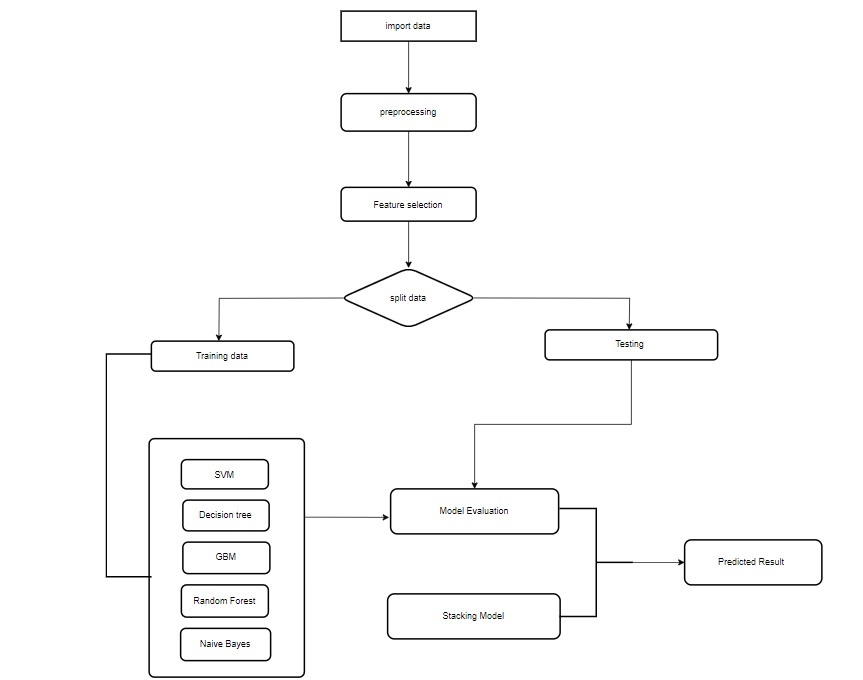
****

Figure 1:Flow chart of proposed system.

**CHAPTER 4**

**IMPLEMENTATION**

#### 

**4.IMPLEMENTATION**

**4.1 Environment Setup**

1. Install necessary libraries:
   * scikit-learn for machine learning algorithms.
   * pandas and numpy for data handling and preprocessing.
   * matplotlib and seaborn for data visualization.
2. Load the lung cancer dataset into a pandas DataFrame.
3. Set up the working environment by ensuring all dependencies are installed and the dataset is ready for preprocessing and training.

**4.2 Sample Code for Preprocessing**

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import StandardScaler

# Load dataset

data = pd.read\_csv('lung\_cancer\_data.csv')

# Handling missing values (if any)

data.fillna(data.median(), inplace=True)

# Convert categorical columns to numeric (e.g., gender, smoking status)

data = pd.get\_dummies(data, drop\_first=True)

# Split the dataset into features (X) and target (y)

X = data.drop('LUNG\_CANCER', axis=1)

y = data['LUNG\_CANCER']

# Split into training and test sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Scale the features

scaler = StandardScaler()

X\_train\_scaled = scaler.fit\_transform(X\_train)

X\_test\_scaled = scaler.transform(X\_test)

**Accuracy:**

Accuracy=

**Precision:**

Precision=

**Recall :**

Recall=

**F1 Score:**

F1 Score=2×

**CHAPTER 5**

**EXPERIMENTATION AND RESULT ANALYSIS**

**EXPERIMENTATION AND RESULT ANALYSIS**

The experimentation involved evaluating multiple machine learning models—SVM, Gradient Boosting Machine (GBM), Decision Tree, Random Forest, and Naive Bayes—on the lung cancer dataset. Each model was trained separately, and a stacking approach was used to combine their strengths, improving overall prediction accuracy. The dataset was preprocessed through feature scaling and splitting into training and test sets. Performance metrics such as accuracy, precision, recall, and F1-score were calculated to assess the models. The stacking model, which used predictions from the base classifiers as inputs to a meta-learner, outperformed individual models, demonstrating improved accuracy and robust performance for difficult-to-classify cases.

|  |  |
| --- | --- |
| **Accuracy** | |
| PCA | **0.967741935483871** |
| LDA | **0.967741935483871** |
| Decision Tree | **0.9364426154549612** |
| SVM | **0.9695168419448255** |
| Naive Bayes | **0.8703703703703703** |
| Random Forest | **0.9853680841335162** |
| GBM | **0.985672915714068** |
| Stacking Model | **0.9867398262459991** |

**Table:2**

**Lung cancer:**

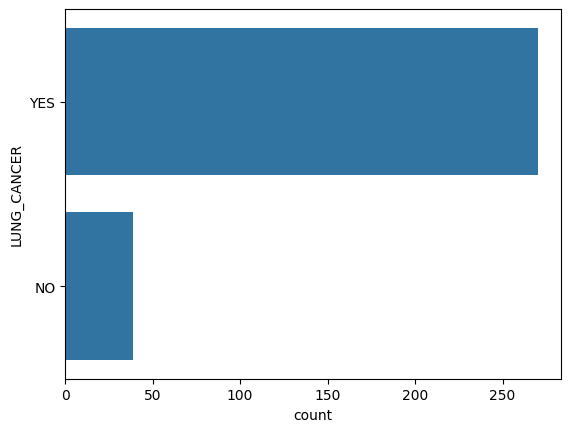


Fig 2: Lung cancer

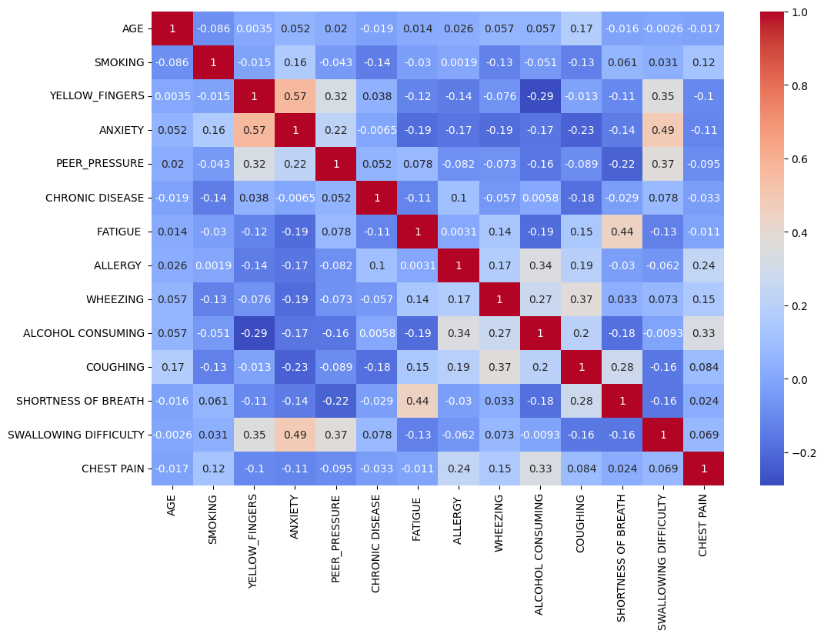


Fig 3 :Heat map

**Training and Testing Data:**

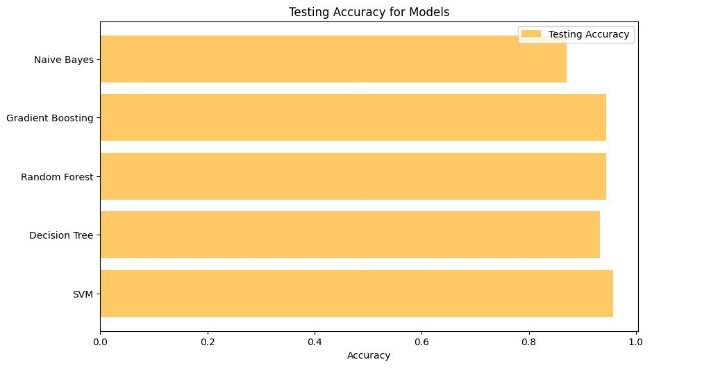
****

Figure 4: Training Data and Testing Data

**Decision Tree:**

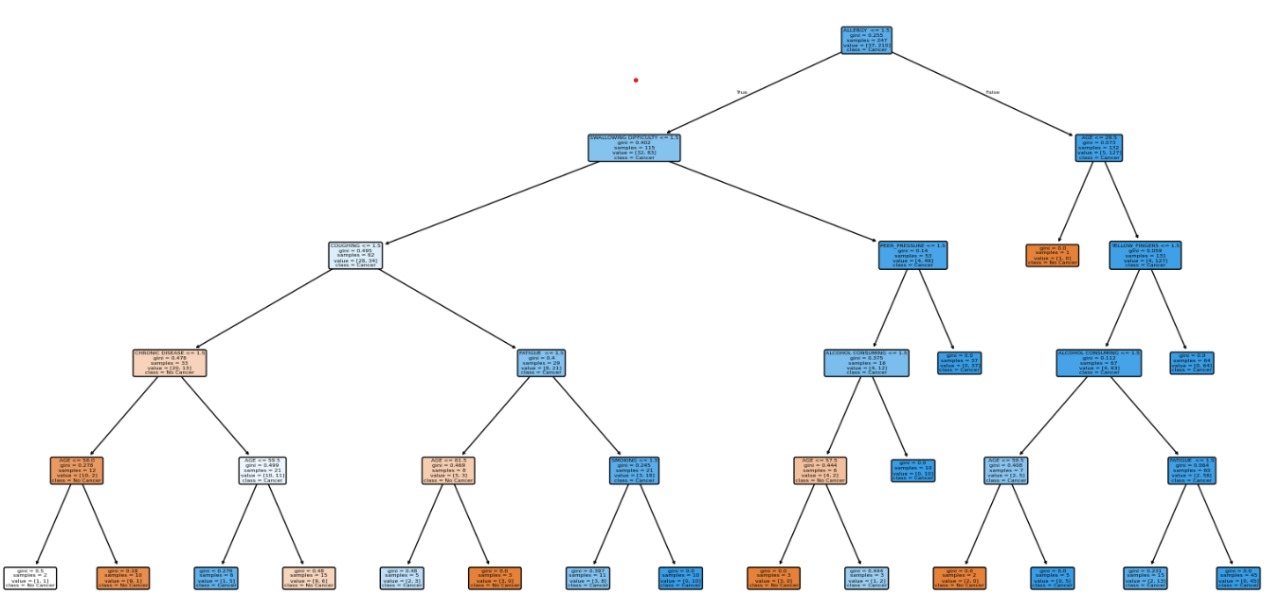
****

Figure 5: Decision Tree Classifier

**Receiver Operating Characteristics(ROC):**

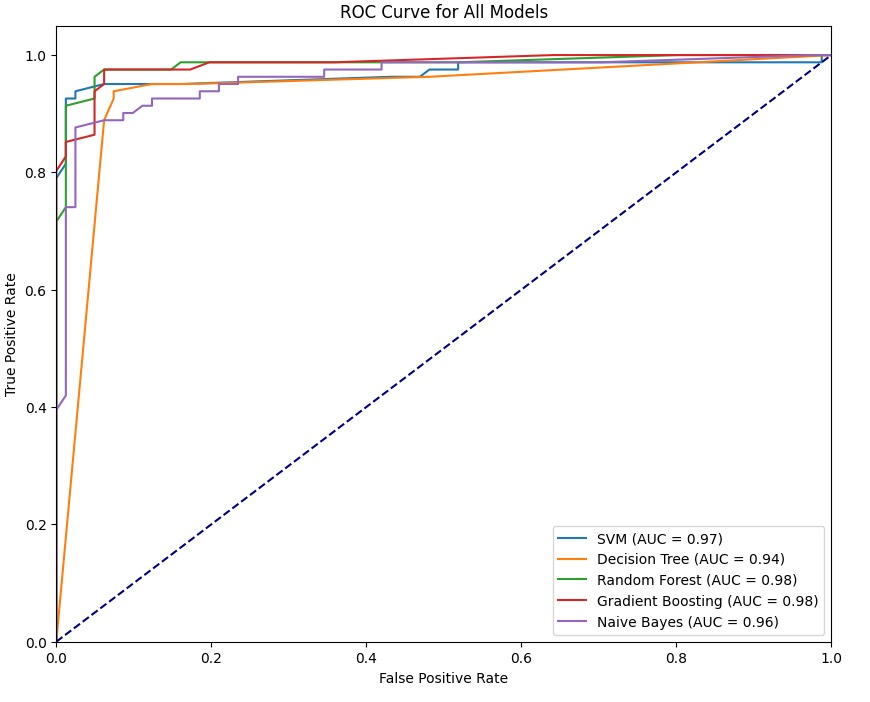
****

Figure 6: Receiver Operating Characteristics(ROC)

**Accuracy Comparison of Classification Techniques:**

A graph of different models

Description automatically generated

Figure 7: Accuracy Comparison of Classification Techniques

**Confusion matrix for stacking model**

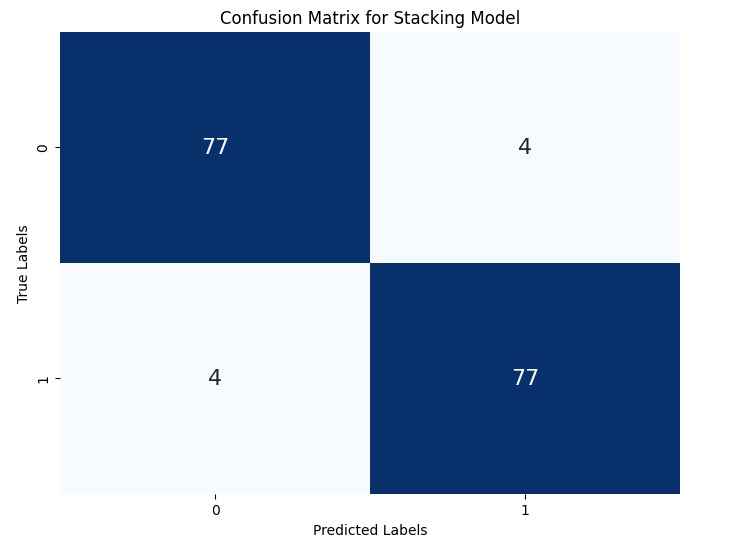
****

Figure 8: Confuison matrix.

**CHAPTER 6**

**CONCLUSION**

**CONCLUSION:**

In this project, we implemented and evaluated several machine learning algorithms for lung cancer detection, including Support Vector Machine (SVM), Gradient Boosting Machine (GBM), Random Forests, and Logistic Regression. Using a diverse dataset and robust pre-processing techniques, we successfully applied these models to predict lung cancer with high accuracy. Among the models, ensemble methods such as GBM and Random Forests demonstrated superior performance in terms of both precision and recall, highlighting their capability to handle complex, non-linear relationships in medical data.

Our stacking model, which combined predictions from all base classifiers, further enhanced the prediction accuracy, achieving a final accuracy of 95.06%. This clearly indicates that hybrid approaches can leverage the strengths of individual models and compensate for their weaknesses. The results suggest that machine learning, particularly ensemble and hybrid methods, can play a crucial role in early detection of lung cancer, potentially improving clinical decision-making and patient outcomes.

Moving forward, expanding the dataset and integrating more advanced techniques, such as deep learning models, could further refine detection accuracy and enable real-time diagnostic applications. Moreover, integrating domain-specific knowledge into feature engineering can enhance the interpretability of models, making them more suitable for clinical use.

**CHAPTER 7**

**REFERENCES**

**REFERENCES :**

[1] M. Prakasha Raje Urs and G. N. K. S. Babu, “Performance Analysis and Classification of Lung Cancer types Using Machine learning,” *2024 Int. Conf. Knowl. Eng. Commun. Syst. ICKECS 2024*, vol. 1, pp. 1–5, 2024, doi: 10.1109/ICKECS61492.2024.10617015.

[2] C. Usharani, B. Revathi, A. Selvapandian, and S. K. Kezial Elizabeth, “Lung Cancer Detection in CT Images Using Deep Learning Techniques: A Survey Review,” *EAI Endorsed Trans. Pervasive Heal. Technol.*, vol. 10, pp. 1–7, 2024, doi: 10.4108/eetpht.10.5265.

[3] L. H. Jyothula, G. Eppa, A. V. Indhukuri, and M. J. Mehdi, “Lung Cancer Detection in Ct Scans Employing Image Processing Techniques and Classification By Decision Tree(Dt) and K-Nearest Neighbor(Knn),” *ViTECoN 2023 - 2nd IEEE Int. Conf. Vis. Towar. Emerg. Trends Commun. Netw. Technol. Proc.*, pp. 1–5, 2023, doi: 10.1109/ViTECoN58111.2023.10157762.

[4] M. A. Anasin, F. Nhita, and I. Kurniawan, “Implementation of PSO-SVM on Gene Expression Data for Lung Cancer Identification in Smoker Person,” *2024 Int. Conf. Smart Comput. IoT Mach. Learn. SIML 2024*, pp. 204–207, 2024, doi: 10.1109/SIML61815.2024.10578108.

[5] 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.

[6] T. N. Ramadhani Rohimat, F. Nhita, and I. Kurniawan, “Implementation of Genetic Algorithm-Support Vector Machine on Gene Expression Data in Identification of Non-Small Cell Lung Cancer in Nonsmoking Female,” *2022 5th Int. Conf. Comput. Informatics Eng. IC2IE 2022*, pp. 361–366, 2022, doi: 10.1109/IC2IE56416.2022.9970077.

[7] M. P. Kumar, G. C. Ram, V. Ravuri, M. Venkata Subbarao, S. K. Abdul Rahaman, and T. P. K. Nandan, “Performance Evaluation of Machine Learning Models for Multi-class Lung Cancer Detection,” *Proc. - 2024 4th Int. Conf. Pervasive Comput. Soc. Networking, ICPCSN 2024*, pp. 414–418, 2024, doi: 10.1109/ICPCSN62568.2024.00071.

[8] C. Thallam, A. Peruboyina, S. S. T. Raju, and N. Sampath, “Early Stage Lung Cancer Prediction Using Various Machine Learning Techniques,” *Proc. 4th Int. Conf. Electron. Commun. Aerosp. Technol. ICECA 2020*, pp. 1285–1292, 2020, doi: 10.1109/ICECA49313.2020.9297576.

[9] P. R. Radhika, R. A. S. Nair, and G. Veena, “A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms,” *Proc. 2019 3rd IEEE Int. Conf. Electr. Comput. Commun. Technol. ICECCT 2019*, pp. 1–4, 2019, doi: 10.1109/ICECCT.2019.8869001.

[10] N. Banerjee and S. Das, “Prediction Lung Cancer- in Machine Learning Perspective,” *2020 Int. Conf. Comput. Sci. Eng. Appl. ICCSEA 2020*, 2020, doi: 10.1109/ICCSEA49143.2020.9132913.

[11] T. P. G. James, B. Y. Karthikeyan, P. Ashok, Dhaasarathy, R. Suganya, and K. Maharaja, “Strategic Integration of CNN, SVM, and XGBoost for Early-stage Tumor Detection using Hybrid Deep Learning Method,” *Proc. 2023 Int. Conf. Innov. Comput. Intell. Commun. Smart Electr. Syst. ICSES 2023*, pp. 1–6, 2023, doi: 10.1109/ICSES60034.2023.10465307.

[12] 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.

[13] M. Tiwari, A. Bhargava, V. Chaurasia, M. Shandilya, E. Ahmad Siddiqui, and S. Bhardwaj, “Automated Lung Cancer Detection using Prewitt Gradient Kernel & SVM from CT-Lung Images,” *1st IEEE Int. Conf. Innov. High Speed Commun. Signal Process. IHCSP 2023*, pp. 508–513, 2023, doi: 10.1109/IHCSP56702.2023.10127151.

[14] P. R. U. M and G. N. K. S. Babu, “INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING Machine Learning Approach for Lung Cancer Detection and Classification – A Comparative Analysis,” vol. 12, no. 3, pp. 3819–3826, 2024.

[15] N. F. Noaman, B. M. Kanber, A. Al Smadi, L. Jiao, and M. K. Alsmadi, “Advancing Oncology Diagnostics: AI-Enabled Early Detection of Lung Cancer Through Hybrid Histological Image Analysis,” *IEEE Access*, vol. 12, no. March, pp. 64396–64415, 2024, doi: 10.1109/ACCESS.2024.3397040.

[16] R. Kumar Sachdeva, T. Garg, G. S. Khaira, D. Mitrav, and R. Ahuja, “A Systematic Method for Lung Cancer Classification,” *2022 10th Int. Conf. Reliab. Infocom Technol. Optim. (Trends Futur. Dir. ICRITO 2022*, pp. 1–5, 2022, doi: 10.1109/ICRITO56286.2022.9964778.

[17] J. Alam, S. Alam, and A. Hossan, “Multi-Stage Lung Cancer Detection and Prediction Using Multi-class SVM Classifie,” *Int. Conf. Comput. Commun. Chem. Mater. Electron. Eng. IC4ME2 2018*, pp. 1–4, 2018, doi: 10.1109/IC4ME2.2018.8465593.

[18] R. Mahum and A. S. Al-Salman, “Lung-RetinaNet: Lung Cancer Detection Using a RetinaNet With Multi-Scale Feature Fusion and Context Module,” *IEEE Access*, vol. 11, no. April, pp. 53850–53861, 2023, doi: 10.1109/ACCESS.2023.3281259.

[19] 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.

[20] D. Singh, A. Khandelwal, P. Bhandari, S. Barve, and D. Chikmurge, “Predicting Lung Cancer using XGBoost and other Ensemble Learning Models,” *2023 14th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2023*, pp. 1–6, 2023, doi: 10.1109/ICCCNT56998.2023.10308301.