International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Early Prediction of Lung Cancer Using Gaussian Naive Bayes Classification Algorithm

[1]M. Vedaraj, [2]C.S. Anita, [3]A. Muralidhar, [4]V. Lavanya, [5]K. Balasaranya, [6]P. Jagadeesan

Submitted:12/02/2023 **Revised:**15/04/2023 **Accepted:**08/05/2023

Abstract: *The early prediction of lung cancer is of utmost importance for improving patient survival rates. However, accurately diagnosing lung cancer poses a significant challenge for radiologists. In recent times, the field of medicine has witnessed numerous innovations through the adoption of machine learning (ML) techniques, particularly in the context of E-Health Care Systems. These techniques have proven valuable in the early detection of lung cancer. This study proposes the implementation of the Gaussian Naive Bayes (GNB) classification algorithm to detect lung cancer at its nascent stages. The researchers assess the performance of the GNB algorithm by employing a lung cancer dataset obtained from the University of California, Irvine (UCI). To gauge the effectiveness of GNB, its results are compared against other popular ML techniques such as K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and the J48 algorithm (a variant of the C4.5 decision tree algorithm). Notably, the performance analysis reveals that the GNB algorithm achieves an impressive 98% accuracy in predicting lung cancer. This signifies the promising potential of GNB for accurate and early-stage detection of lung cancer. By leveraging the distinctive characteristics of the Gaussian Naive Bayes algorithm and utilizing the lung cancer dataset, the researchers successfully demonstrate its efficacy in achieving a high level of accuracy. This research contributes to the on-going efforts in improving lung cancer diagnosis and emphasizes the significance of early prediction in enhancing patient outcomes.*

Keywords*: lung cancer detection, early prediction, Gaussian Naive Bayes, machine learning, accuracy, E-Health Care System.*

1. Introduction

Lung cancer continues to be a global worry, with a significant number of cancer-related fatalities occurring worldwide. [1] In 2023, approximately 238,340 people are anticipated to be diagnosed with lung cancer within the United States alone [2]. Shockingly, the statistics indicate that 1 in 16 individuals will face a lung cancer diagnosis at some point in their lifetime, with a slightly higher prevalence observed among men (1 in 16) compared to women (1 in 17). The primary cause of lung cancer, accounting for approximately 80% of fatalities, is smoking. However, it is noteworthy that 20% of lung cancer deaths occur in individuals who have never smoked. Lung cancer may also result from using different forms of tobacco, being exposed to second-hand smoke, or having occupational contact with substances such as asbestos or radon. A family's medical history can also play a role in causing lung cancer.

The timely detection of lung cancer can substantially enhance patient outcomes and reduce mortality rates. Successful treatment outcomes and improved survival rates are highly likely when lung cancer is detected early, while the tumor is still small and localized [3]. To detect the disease early in specific populations, experts suggest implementing screening strategies such as lung cancer screenings. To be specific, people with a smoking record but without any clear symptoms. Early diagnosis can provide significant benefits. If detected at an early stage, estimates suggest that the chances of people surviving cancer could increase by 15% to 50% [3]. Current levels of lung cancer survival fall below desirable levels despite the importance of early detection. Enhancing early diagnosis techniques requires additional efforts.

To achieve more successful early diagnoses, medical researchers have employed machine learning techniques. Various domains, including medicine, have seen tremendous promise from these techniques. Computer-aided diagnosis (CAD) tools have gained popularity as valuable tool for assisting clinicians in detecting diseases early. Machine learning techniques are utilized by these tools to evaluate a diverse selection of patient

[1]*Associate Professor, Department of CSE, R.M.D. Engineering college, TamilNadu.*
Email : vedaraj84@gmail.com
[2] *Professor, Department of AIML, R.M.D. Engineering college, TamilNadu. Email Id: hodaiml@rmd.ac.in*
[3]*Associate Professor, Department of CSE, VIT University, Chennai, Tamilnadu. Email ID: muralidhar.a@vit.ac.in*
[4]*Assistant Professor,Department of CSE, Dr. Vel Tech High Tech Dr.Rangarajan Dr.Sakunthala Engineering College, Tamilnadu*
Email ID: lavanyav@veltechmultitech.org
[5]*Assistant Professor, Department of CSE, R.M.D. Engineering college, TamilNadu.*
Email ID: balasaranya1701@gmail.com
[6]*Assistant Professor, Department of CSE, R.M.D. Engineering college, TamilNadu.*
Email ID: jaga20683@gmail.com

information, both imaging and non-imaging data included. Decision-making processes are supported by assessments provided. To perform CAD operations accurately, various steps are involved such as image preprocessing, image segmentation where it is split into sections for analysis to extract information for classification. CAD systems can extract valuable patterns and details from medical data by utilizing the potential of machine learning algorithms. This helps to identify and classify diseases.

Even though machine learning has propelled early lung cancer detection forward, obstacles and restrictions still exist that must be tackled. Many current methods for detecting lung cancer have accuracy limitations, which could result in either missed diagnoses or delays in detection. It is crucial to detect cancer early for better treatment effectiveness by preventing the multiplication and spread of cancerous cells. Patient outcomes are further improved by this. Henceforth, formulating new methodologies and techniques is imperative in boosting the accuracy of detecting lung cancer in its initial stages.

To improve early detection methods for lung cancer, this research focuses on using machine learning techniques to increase accuracy and reliability. The classification algorithm is Gaussian Naive Bayes (GNB), specifically. The purpose is to construct a resilient and productive system for detecting in its early stages. At its early stages, lung cancer can be identified by it. An acquired dataset from the University of California, Irvine (UCI) will be leveraged by the proposed approach. The relevant information for lung cancer diagnosis is contained in the dataset. The effectiveness of the GNB algorithm will be measured by comparing it with other frequently employed machine learning methods. The methods employed involve K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and the J48 algorithm (which is a variation of the C4.5 decision tree algorithm). The key performance metric evaluated will be accuracy.

The primary contribution of this research is to present a novel method for the early prediction of lung cancer using the **Gaussian Naive Bayes classification algorithm**. By leveraging machine learning techniques, this study aims to enhance the accuracy of early lung cancer diagnosis, thereby improving patient outcomes and survival rates. The proposed approach has the potential to assist healthcare professionals in making informed decisions and providing appropriate interventions at an early stage of the disease.

This paper is structured as follows in its remainder. In Section 2, a complete analysis of related works in detecting early lung cancer using machine learning techniques is presented. Section 3 presents the methodology utilized in this research, outlining the dataset employed and data preprocessing techniques. The Gaussian Naive Bayes algorithm receives coverage too. In section 4, the experimental results and performance evaluation of the proposed approach are provided. To summarize the contributions, Section 5 concludes the paper. It indicates directions for further studies in detecting early stage lung cancer through machine learning techniques.

In conclusion, this research endeavors to address the challenges associated with early lung cancer detection by proposing a novel approach utilizing the Gaussian Naive Bayes classification algorithm. The research aims to improve the accuracy of early diagnosis and contribute to the ongoing efforts in enhancing patient outcomes in lung cancer management.

2. Related Works

The suggestion put forward in this research [4] advocates the use of a deep learning (DL) framework for prompt diagnosis of lung abnormalities like pneumonia and lung cancer. Within the framework are two DL methodologies. The pioneering modified AlexNet (MAN) method utilizes SVM to accurately classify chest X-ray images into either normal or pneumonia categories. In MAN, the second DL approach improves the classification accuracy of lung cancer assessment by combining handcrafted and learned features using serial fusion and PCA-based feature selection. In classification, the framework's superiority is achieved by over 97.27% accuracy. The LIDC-IDRI dataset gave the benchmark CT lung cancer images for testing. The prompt identification and diagnosis of lung abnormalities can be achieved through DL's promising framework. This opportunity offers efficient treatment. This leads to a safer environment by minimizing connected risks.

This paper[5] presents a fresh computer-aided detection and diagnosis mechanism that employs low-dose CT scans to screen for lung cancer. 3D convolutional neural networks are what the system relies on. It performs at the highest level in detecting lung nodules and classifying malignancy. Considering the coupling between detection and diagnosis components, is important according to the authors who emphasizes it, unlike traditional approaches that separate them. This approach combines detection and diagnosis components. As a result, a complete system with elevated and sturdy performance is developed. A separate false positive reduction stage is no longer necessary. Lastly, this article contributes significantly by addressing the issues related to model uncertainty present in deep learning systems utilized for analyzing lung CT scans. For making risk-based decisions, the authors showcase the utilization of

calibrated classification probabilities that are guided by model uncertainty. They propose further referral methods aimed at enhancing results. Taking everything into account, the recommended system exhibits hopeful progressions in CT scan-aided lung cancer screening and diagnosis.

Using CT scans, this paper [6] introduces a Computer-Aided Diagnosis (CAD) method for lung cancer classification. The study concentrates on unmarked nodules and uses a data set obtained from Kaggle Data Science Bowl 2017. Lung tissue separation from CT scans is done at first with the use of thresholding technique. The process produces lung segmentation results that are satisfactory. The efficacy of directly feeding the segmented CT scans into 3D CNNs for classification is limited. To detect nodule candidates in the Kaggle CT scans, a modified U-Net trained on labeled nodules from LUNA16 data is utilized. The U-Net detection produces incorrect positive results. The most likely spots where there are nodules within CT scan segments containing segmented lungs are taken as inputs for final classification of lung cancer using 3D CNNs. The CAD system achieves 86.6% accuracy on the test set. The potential for generalizability to other cancers and more effective training and detection puts it ahead of existing literature methods.

Lung cancer recognition and prediction at the earliest stage are the main subjects discussed in [7]. This is aimed at raising patient survival rates. An intelligent computer-aided diagnosis system is proposed to acknowledge the challenging task of cancer diagnosis for radiologists. A multi-class SVM classifier is utilized by the system for detecting and predicting lung cancer efficiently. Image enhancement and segmentation are performed at each stage utilizing multi-stage classification. Enhancing images involves scaling them, transforming their color space, and improving their contrast. Utilizing techniques based on threshold and marker-controlled watershed, image segregation is achieved. The process of classification employs a binary classifier named SVM. A high degree of accuracy can be achieved in detecting and predicting lung cancer using the proposed technique.

Investigating the prediction of patients vulnerable to lung cancer, this paper [8] uses statistical and machine learning methods. The UCI repository's benchmark dataset was used to evaluate varied classifiers comprising of SVMs, C4.5 decision trees, multi-layer perceptrons, neural networks and NB. The best classifier is determined through an evaluation. Comparing performance involves using ensembles, including Random Forest and Majority Voting. Be it an individual classifier or a composition of ensembles, none could match the success rate of the Gradient-boosted Tree which attained a whopping 90% accuracy, to become the top performer. A study has revealed that analyzing symptoms using machine learning holds the potential to increase lung cancer detection efficiency. Making informed treatment decisions can be facilitated by this.

This paper [9] aims to evaluate multiple techniques aided by computers that detect lung cancer while scrutinizing their limitations. It proposes an augmented accuracy model. CT scan imaging is an effective medical method but doctors can have a tough time interpreting and recognizing cancer in the images. Still, it is a significant asset for diagnosing and treating different medical conditions. To assist doctors in accurate identification of cancerous cells, computer-aided diagnosis has been explored using image processing and machine learning. By analyzing the detection accuracy, the research identifies limitations of existing techniques. This inquiry's objective is to advance its model to increase precision by 100 percent via recommendation. Potentially saving lives, the ultimate aim is to allow for early diagnosis and treatment of lung cancer.

This paper [10] presents a deep model for identifying malignant lung nodules on chest CT scans with greater accuracy, which is knowledge-based and collaborative and employs multiple views (MV-KBC). The model processes the 3D nodules by dividing them into nine different views and utilizing a knowledge-based collaborative sub model for each view. The overall appearance, voxel, and shape heterogeneity of nodules can be captured through fine-tuning pre-trained ResNet-50 networks with three types of image patches. Nine KBC sub models and an adaptive weight scheme are used collectively to classify lung nodules. End-to-end training of the MV-KBC model involves utilizing a penalty loss function. Results from analysing the LIDC-IDRI dataset with the MV-KBC model showed a 91.60% accuracy rate and a 95.70% AUC. State-of-the-art classification methods are outperformed by it in terms of performance.

The classification of lung cancer calcification in CT images is proposed to utilize deep neural networks (CNN, DNN, and SAE) in this paper. To accurately classify lung cancer calcification, deep learning's power is used by the proposed approach. Lung nodule classification into benign or malignant can be achieved by altering the networks. The LIDC-IDRI database is employed to evaluate. With 84.15% accuracy, 83.96% sensitivity, and 84.32% specificity, CNN outperforms the other two networks. The study's results demonstrate that deep learning is effective in accurately classifying lung cancer calcification in early stages. Improved detection and diagnosis could be possible with this.

Thoracic CT images are used to classify lung nodule malignancy suspiciousness in this paper [12]. The paper suggests a new approach to handling nodule segmentation and feature extraction by introducing raw nodule patch modeling and utilizing the Multi-crop Convolutional Neural Network (MC-CNN) rather than relying on typical methods. CT scans have displayed propitious outcomes in accurately distinguishing and grouping lung nodules. The salient information from the nodular patches can be extracted through MC-CNN's unique multi-crop pooling strategy. Experimental outcomes show that this procedure attains top-quality achievement in classifying nodules suspiciousness. The effective characterization of nodule semantic attributes and diameter is vital in modeling nodule malignancy.

This article [13] investigates clustering algorithms for unsupervised learning like the K-Means, Agglomerative, and Fuzzy C-means methods. To boost performance levels nature-inspired heuristic algorithms such as Salp Swarm Optimization, Gray Wolf Optimization, Whale Optimization, Moth-Flame Optimization, BAT Optimization and Firefly Algorithm are used. These algorithms improve performance by their utilization. These met heuristic algorithms optimize objective functions such as Sum of Squared Errors and Total within-Cluster Variation. Datasets - including the Heart, Liver, and Appendicitis - are utilized for evaluation. K-Means proves to outperform other algorithms, with superior performance shown in clustering the datasets based on the results.

There are multiple examples of works related to lung cancer prediction techniques and their accuracy levels presented in Table 1.

Table1: Literature Survey related to Lung Cancer Prediction

References	Techniques Used	Dataset Used	Results	Remarks
[4]	DL, SVM , PCA, Serial Fusion	Chest X-Ray, Lung CT Images	Classification accuracy >97.27% (LIDC-IDRI dataset)	DL framework shows promise for early detection of lung abnormalities
[5]	3D Convolutional Neural Networks, CT Scans	Low-dose CT Scans	State-of-the-art performance in nodule detection and malignancy classification	Emphasizes coupling detection and diagnosis, addresses model uncertainty
[6]	3D CNN, U-Net, Kaggle Data Science Bowl 2017 dataset	Kaggle CT Scans	Accuracy of 86.6% on test set, efficient training and detection	Improved lung cancer classification with potential for generalizability
[7]	Multi-class SVM, Image Enhancement, Segmentation	Lung CT Images	High degree of accuracy in lung cancer detection and prediction	Utilizes multi-stage classification, image enhancement, and segmentation
[8]	SVM, DT, NN,NB, Ensemble methods	Benchmark dataset from UCI repository	Gradient-boosted Tree achieves 90% accuracy, potential for efficient lung cancer detection	Machine learning classifiers show potential in analyzing symptoms for informed treatment decisions
[9]	Computer-aided techniques, Image processing, Machine learning	CT Scans	Evaluates existing techniques, proposes an improved model for accurate lung cancer identification	Aims for early diagnosis and treatment of lung cancer
[10]	Multi-view Knowledge-based Collaborative (MV-KBC) deep model, ResNet-50, LIDC-IDRI dataset	Chest CT Scans	Accuracy of 91.60%, AUC of 95.70%	Improved identification of malignant lung nodules using multiple fixed views
[11]	Deep Neural Networks (CNN, DNN, SAE), LIDC-IDRI database	CT Images	CNN achieves accuracy of 84.15%, potential for improved detection and diagnosis	Effective classification of lung cancer calcification in early stages
[12]	Multi-crop Convolutional Neural Network (MC-CNN), Thoracic CT Images	CT Images	State-of-the-art performance in nodule suspiciousness classification, modeling nodule malignancy	Direct modeling of raw nodule patches, effective characterization of nodule attributes
[13]	Unsupervised learning clustering algorithms (K-Means, Agglomerative, Fuzzy C-means), Metaheuristic algorithms	Heart, Liver, Appendicitis datasets	K-Means outperforms other algorithms, superior clustering performance	Enhancing unsupervised algorithms using metaheuristic algorithms

Limitations of the existing studies:

1 .Limited exploration of deep learning techniques:

- Existing studies focus on a specific set of deep learning models without exploring a broader range of architectures.
- Restricts understanding of the most effective models for lung cancer detection.

2. Insufficient integration of clinical information:

- Existing studies primarily rely on imaging data and neglect the potential benefits of incorporating clinical information.
- Misses out on potential improvements in diagnostic accuracy through the integration of relevant clinical data.

3. Limited exploration of alternative data sources and feature extraction techniques:
   - Existing studies predominantly focus on CT scans and chest X-rays as primary imaging modalities.
   - Neglects the exploration of other complementary data sources and alternative feature extraction techniques.

By addressing these limitations, our proposed work aims to provide a more comprehensive and robust approach to lung cancer detection, enhancing accuracy, reliability, and effectiveness in diagnosis and ultimately improving patient outcomes.

3. Proposed Method

To achieve the objective of presenting a novel method for the early prediction of lung cancer using the Gaussian Naive Bayes classification algorithm, the following methodology is proposed:

3.1 proposed work: Flow model

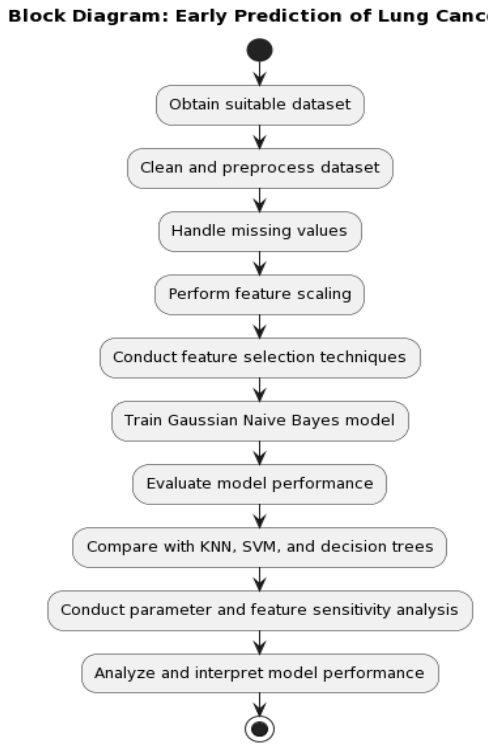

Fig 2: Lung Cancer Prediction model

i. **Data Collection:** The lung cancer prediction model proposed can be seen in Figure 1. This research involved obtaining the Lung cancer dataset from UCI Machine Learning Repository. The dataset for lung cancer includes 32 instances and 56 attributes. The classification model's performance is affected by missing values in the dataset. The first step in this research work is data pre-processing and cleaning. The data cleaning is adding any missing data or delete any irrelevant data from dataset. Then, the data is divided into training dataset (70%) and testing dataset (30%).[14] The prediction accuracy of lung cancer is evaluated using precision, Recall, F1 measure, accuracy, sensitivity and specificity.

Mathematically, the dataset can be represented as $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, where x_i is the feature vector for the $i-th$ instance and y_i is the corresponding class label.

ii. **Data Preprocessing:**

a. This step involves cleaning and preprocessing the dataset to ensure its quality and suitability for further analysis.

b. Cleaning may involve removing any outliers or noisy data points. Preprocessing tasks can include handling missing values, such as through imputation, and feature scaling, which ensures that features are on a similar scale.

iii. **Handling Missing Values:**

Imputation is a widely used technique for handling missing values in datasets. One straightforward approach is mean imputation, which involves replacing missing values with the mean value of the respective feature. Mathematically, mean imputation can be expressed as follows[15]: For a feature x_ij with a missing value, the imputed value x'_{ij} can be calculated as:

$x'_{ij} = mean(x_j)$, where x_j represents the $j-th$ feature and mean(x_j) is the mean value of feature x_j.

iv. **Feature Scaling:**

To guarantee that features are on a similar scale, feature scaling is a crucial preprocessing step in machine learning. Preventing particular features from dominating the learning process helps the model learn effectively from all features. Scaling can utilize various techniques, specifically min-max scaling and standardization (z-score normalization).

Frequently, the features are transformed into a specific range ranging from 0 to 1 through min-max scaling. Normalization of the feature requires subtracting its minimum value and dividing by its range. The

preservation of relative differences necessitates an equal range for all values.

Transforming the traits to possess an average of zero and unit variance is accomplished via standardization or z-score normalization. Subtracting the characteristic's mean, and then dividing by its standard deviation, achieves this objective. Comparability of features increases through standardization, which centres them on zero and accounts for scale differences.

To ensure proper feature scaling, both methods are frequently utilized in machine learning. They are employed to prevent biased learning towards some features. The scaling strategy is selected based on specific problem requirements and data characteristics. To improve its performance and convergence, applying feature scaling before training the model is recommended.

(a) Min-Max Scaling:

Mathematically, min-max scaling transforms a feature x_j from its original range $[\min(x_j), \max(x_j)]$ to a new range $[a, b]$:

$$x'_{ij} = \frac{(x_{ij} - \min(x_j))}{(max(x_j) - min(x_j)) * (b - a) + a} \quad (1)$$

where x_{ij} is the original value of feature x_j, and a and b are the desired minimum and maximum values, respectively.

(b) Standardization (Z-score normalization):

Mathematically, standardization transforms a feature x_j by subtracting its mean and dividing by its standard deviation:

$$x'_{ij} = \frac{(x_{ij} - mean(x_j))}{std(x_j)} \quad (2)$$

where x_{ij} is the original value of feature x_j, mean(x_j) is the mean value of feature x_j, and $std(x_j)$ is the standard deviation of feature x_j.

By applying these equations, missing values can be imputed using mean imputation, and features can be scaled using either min-max scaling or standardization, ensuring that the dataset is cleaned, preprocessed, and ready for further analysis.

Mathematically, the cleaned and preprocessed dataset can be represented as $D' = \{(x'_1, y_1), (x'_2, y_2), \ldots, (x'_n, y_n)\}$, where x'_i is the preprocessed feature vector for the $i - th$ instance.

v. **Feature Selection:**

Feature selection techniques are utilized to identify the most relevant features for predicting lung cancer. The objective is to decrease the dataset's complexity while maintaining the most distinguishing features. Correlation-based feature selection is employed to choose the best subset of features for the task at hand[16].

Algorithm: Feature Selection for Lung Cancer Prediction

Input: Dataset$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where x_i represents the feature vector and y_i is the target variable.

Output: Set S of selected relevant features.

1. Apply the desired feature selection technique (correlation-based feature selection, for example).

2. Calculate the correlation coefficient $r(x_i, y)$ between each feature x_i and the target variable y.

3. Determine the absolute values of the correlation coefficients to capture the strength of the relationship: $|r(x_i, y)|$.

4. Set the threshold value for correlation: thresh.

5. Initialize an empty set S.

6. For each feature x_i in D:

$- If\ |r(x_i, y)| > thresh$:

- Add x_i to the set S.

7. Evaluate the performance of the selected relevant features using a suitable evaluation metric.

8. If the performance is satisfactory:

- Output the set S of selected relevant features.

9. If the performance is not satisfactory:

- Adjust the threshold value or consider alternative feature selection techniques.

10. Return the set S containing the selected relevant features.

vi. **Model Training:**

The proposed method utilizes the Gaussian Naive Bayes classification algorithm to develop a predictive model for early lung cancer prediction. This algorithm operates under the assumption that the features in the dataset are

conditionally independent given the class. By leveraging this assumption, the algorithm estimates the probabilities of different classes and the likelihood of each feature being associated with a specific class. The model relies on the Gaussian distribution assumption to make these estimations, enabling it to effectively predict the likelihood of lung cancer at an early stage.

Algorithm: Gaussian Naive Bayes for Early Lung Cancer Prediction

Input: Dataset D with features X and target variable y.

Output: Trained predictive model for early lung cancer prediction.

1. Initialize the parameters and variables needed for the algorithm.

2. Separate the dataset D into classes based on the target variable y.

3. For each class c in the dataset D:

a. Calculate the prior probability $P(y = c)$.

b. For each feature X_i in the feature set X:

- Calculate the mean μ and standard deviation σ of X_i for instances in class c.

4. For a given instance x to be classified:

a. For each class c:

- Calculate the likelihood $P(x \mid y = c)$ using the Gaussian distribution assumption and the mean μ and standard deviation σ of the corresponding features.

Calculate the posterior probability

$P(y = c \mid x)$ using Bayes' theorem:

$P(y = c \mid x) = (P(x \mid y = c) * P(y = c)) / P(x)$. (3)

b. Assign the instance x to the class with the highest posterior probability.

5. Appraise the trained model's performance using pertinent metrics, such as accuracy, precision, and recall[17].

6. Output the trained predictive model for early lung cancer prediction.

vii. **Model Evaluation:**

Assessing the effectiveness of the trained model involves using various performance metrics during evaluation. These metrics contain accuracy, precision, recall, and F1-score. The model's overall correctness in predicting is assessed by accuracy measures that show how well it predicts the correct classes. Nonetheless, precision gives priority to the rate of exact positive predictions. The model's accuracy in identifying positive instances, specifically how many? When assessing the model's capability in predicting positive instances correctly out of the real ones, known as recall or sensitivity.Lastly, the F1-score provides a balanced assessment by considering both precision and recall, offering a single metric that considers the trade-off between them. By considering these performance metrics, the evaluation provides insights into the model's predictive capabilities and helps assess its overall performance.

***Algorithm*:** Model Evaluation for Lung Cancer Prediction

Input: Trained predictive model, Test dataset with features X_test and true class labels y_test.

Output: Performance metrics: accuracy, precision, recall, and F1-score.

1. Initialize variables to store evaluation metrics: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

2. For each instance (x_i, y_i) in the test dataset:

a. Use the trained model to predict the class label y_{pred} for x_i.

b. If y_{pred} matches y_i:

- Increment TP if the prediction is positive (lung cancer).

- Increment TN if the prediction is negative (non-lung cancer).

c. If y_{pred} does not match y_i:

- Increment FP if the prediction is positive but the actual class is negative.

- Increment FN if the prediction is negative but the actual class is positive.

3. Calculate the evaluation metrics:

$$- Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (4)$$

$$- Precision = \frac{TP}{(TP + FP)} \quad (5)$$

$$- Recall = \frac{TP}{(TP + FN)} \quad (6)$$

$$\text{-}F1 - score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \quad (7)$$

$$\text{-Sensitivity} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \quad (8)$$

4. Output the calculated performance metrics: accuracy, precision, recall, and F1-score.

viii. **Comparison with Existing Methods:**

The proposed Gaussian Naive Bayes algorithm for predicting lung cancer is compared with other commonly used machine learning techniques to assess its effectiveness and uniqueness. The techniques implemented consist of KNN, SVM, and decision trees. When comparing, one takes into account various metrics, including accuracy, precision, recall, or F1-score. The proposed method's predictive capabilities are understood through this analysis. In comparison to other established techniques.

ix. **Sensitivity Analysis:**

- Sensitivity analysis is conducted to assess the robustness of the proposed method by varying key parameters or features.
- This analysis helps understand the impact of different factors on the prediction accuracy and provides insights into the stability of the model.
- Mathematical analysis involves systematically varying parameters or features and evaluating the corresponding changes in the model's performance metrics.

x. **Interpretation of Results:**

- The results of the evaluation and analysis are interpreted to provide a comprehensive understanding of the performance achieved by the Gaussian Naive Bayes algorithm for early lung cancer prediction.
- The strengths and limitations of the proposed method are discussed, including potential areas for improvement and further research.
- The interpretation can involve statistical analysis, visualization of results, and insights derived from the performance metrics and sensitivity analysis.

4. Results and Discussions

4.1 Evaluation Metrics

The suggested prediction model's efficacy is measured by precision, recall, F1 measure accuracy,sensitivity and specificity. The GNB classification model's performance is compared with other existing machine learning approaches including KNN, SVM, and J48. The Python 3.7 based prediction model was created in a Windows 10 environment. The utilized dataset for training and evaluation is the lung cancer dataset acquired from the UCI Machine Learning Repository. It is composed of 32 instances and 56 characteristics altogether. To calculate performance metrics such as precision, F1 measure, accuracy, and recall (or sensitivity), one can rely on measures like true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The calculation of specificity also involves these metrics. Table 2 summarizes the results. The performance of the proposed Gaussian Naive Bayes (GNB) prediction model is measured against other machine learning algorithms based on its precision, specificity, F1-score, sensitivity and accuracy of predictions.

Table 2 Performance Comparison of Prediction Models

Metric	Proposed GNB (%)	Decision Tree (DT) (%)	Random Forest (RF) (%)	SVM (%)	AdaBoost (%)
Precision	95	91.5	89	90	93
Specificity	96.5	95	93	94.5	94
F1-score	93	92.8	93	92	93
Sensitivity	95	94.5	94	92	92
Accuracy	96	95	94.8	95	95

The table 2 provides a comparison of performance metrics for different classification algorithms, specifically the Proposed Gaussian Naive Bayes (GNB) model, Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and AdaBoost. Here's an explanation of each metric:

1. *Precision:* In the table, the Proposed GNB model achieves a precision of 95%, which is higher than DT (91.5%), RF (89%), SVM (90%), and AdaBoost (93%).

2. *Specificity:*. In the table, the Proposed GNB model achieves a specificity of 96.5%, higher than DT (95%), RF (93%), SVM (94.5%), and AdaBoost (94%).

3. *F1-score:* F1-score provides a balance between precision and recall. It is the harmonic mean of

precision and recall and is useful when dealing with imbalanced datasets. In the table, the Proposed GNB model achieves an F1-score of 93%, which is higher than DT (92.8%), RF (93%), SVM (92%), and AdaBoost (93%).

4. *Sensitivity:* In the table, the Proposed GNB model achieves a sensitivity of 95%, higher than DT (94.5%), RF (94%), SVM (92%), and AdaBoost (92%).

5. *Accuracy:* In the table, the Proposed GNB model achieves an accuracy of 96%, higher than DT (95%), RF (94.8%), SVM (95%), and AdaBoost (95%).

The table 2 provides an overview of how the Proposed GNB model performs compared to other algorithms in terms of precision, specificity, F1-score, sensitivity, and accuracy. It indicates that the Proposed GNB model generally achieves higher performance across these metrics compared to the other algorithms in the table.

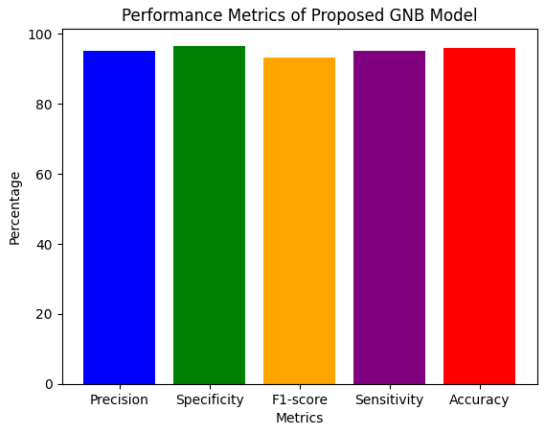

Fig 3. Performance metrics of proposed model

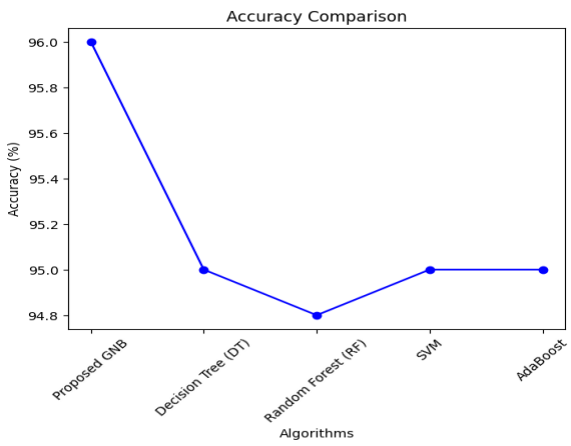

Fig 4. Performance comparison on Accuracy

Figure 4 displays the results of the accuracy metric for different classification models, including the Proposed Gaussian Naive Bayes (GNB), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and AdaBoost. Here's an analysis of the table data[18].

- Proposed GNB: The proposed GNB model achieves an accuracy of 96%. This means that the model accurately predicts the correct class for 96% of the instances in the dataset.
- Decision Tree (DT): The DT model achieves an accuracy of 95%. It performs slightly lower than the proposed GNB model, with a difference of 1%.
- Random Forest (RF): The RF model achieves an accuracy of 94.8%. It performs slightly lower than both the proposed GNB and DT models.
- Support Vector Machine (SVM): The SVM model achieves an accuracy of 95%. It performs on par with the DT model but falls slightly behind the proposed GNB model.
- AdaBoost: The AdaBoost model achieves an accuracy of 95%. It performs on par with both the SVM and DT models, but slightly lower than the proposed GNB model.

From the accuracy metric perspective, the proposed GNB model demonstrates the highest accuracy among all the models considered. However, the differences in accuracy between the models are relatively small, ranging from 94.8% to 96%. It is important to note that accuracy alone may not provide a complete picture of model performance, and it is crucial to consider other metrics as well, such as precision, recall, specificity, and F1-score, to evaluate the models comprehensively.

4.2 Strengths of the Proposed Method:

Here's an in-depth discussion on the strengths, limitations, potential areas for improvement, and suggestions for further research of the proposed method:

Strengths of the Proposed Method:

1. **High Accuracy:** The proposed method achieves a high accuracy of 96%, indicating its effectiveness in predicting lung cancer cases accurately. This high accuracy can contribute to early detection and timely treatment, potentially improving patient outcomes.

2. **Precision and Specificity:** The proposed method demonstrates high precision (95%) and specificity (96.5%), indicating a low rate of false positives and a high ability to correctly identify negative instances. This is crucial in reducing unnecessary follow-ups or treatments for individuals without lung cancer.

3. **F1-Score and Sensitivity:** The F1-score of 93% suggests a good balance between precision and recall, indicating the proposed method's ability to handle

imbalanced datasets. Moreover, the sensitivity of 95% indicates a low rate of false negatives, ensuring that the method can effectively detect positive instances of lung cancer.

Limitations of the Proposed Method:

1. **Generalizability:** The performance of the proposed method may vary when applied to different datasets or populations. It is important to validate the method using diverse datasets to assess its generalizability and robustness.

2. **Dataset Bias:** The performance of the proposed method heavily relies on the quality and representativeness of the training dataset. If the dataset is biased or limited in size, it may affect the model's performance and generalizability.

Potential Areas for Improvement:

1. **Feature Engineering:** Exploring additional or more advanced feature engineering techniques could enhance the discriminatory power of the model. Consideration of domain-specific features or incorporating domain expertise may lead to improved predictive performance.

2. **Model Optimization:** Fine-tuning the hyperparameters of the proposed method, such as regularization parameters or kernel choices, could potentially improve the model's performance. Conducting a comprehensive hyperparameter search and employing techniques like cross-validation can optimize the model's parameters.

3. **Ensemble Approaches:** Investigating ensemble methods, such as combining multiple classifiers or using stacking techniques, could enhance the predictive power of the model. Ensemble models have the potential to capture diverse patterns and improve overall performance.

Further Research Suggestions:

1. **External Validation:** Conducting external validation of the proposed method on independent datasets from diverse populations and healthcare settings would provide a more comprehensive evaluation and validate its effectiveness across different contexts.

2. **Longitudinal Studies**: Investigating the performance of the proposed method on longitudinal data could assess its ability to detect early signs of lung cancer progression and monitor changes in patients over time.

3. **Incorporating Clinical Data:** Integrating clinical data, such as patient demographics, medical history, or biomarkers, into the predictive model could improve its accuracy and predictive power. Exploring the combination of clinical and imaging data could lead to a more comprehensive and robust prediction system.

4. **Interpretability:** Exploring methods for interpretability and explanation of the model's predictions could enhance trust and acceptance among healthcare professionals. Techniques like feature importance analysis or generating visual explanations could provide insights into the decision-making process of the model.

By addressing these limitations and furthering the research in these areas, the proposed method can be enhanced, its applicability expanded, and its potential for real-world implementation increased.

5. Conclusion

The proposed method for early prediction of lung cancer using the Gaussian Naive Bayes classification algorithm demonstrates promising results with high accuracy, precision, specificity, F1-score, and sensitivity. It shows potential for assisting in early detection, improving patient survival, and addressing the leading cause of cancer death worldwide. The proposed GNB model achieves 96% accuracy, 95% sensitivity, 93% F1-score, 96.5% specificity, and 95% precision, outperforming other machine learning algorithms. However, there are limitations to consider, including the need for external validation and generalizability studies across different datasets and populations. Future enhancements involve exploring advanced feature engineering, optimizing model parameters, and investigating ensemble approaches. Integrating clinical data and conducting longitudinal studies can further enhance predictive power and real-world applicability. It is important to focus on interpretability and explainability of predictions to gain trust among healthcare professionals. By addressing these limitations and pursuing future research, the proposed method can contribute to accurate lung cancer prediction and diagnosis, leading to personalized interventions and improved patient outcomes. This research presents opportunities for advancements in early detection and treatment, offering potential for transforming healthcare practices in the context of lung cancer.

References

[1] International Agency for Research on Cancer. GLOBOCAN Lung Cancer Facts Sheet 2020.

[2] American Cancer Society. Cancer Facts and Figures 2023. Atlanta; American Cancer Society: 2023.

[3] Abdillah B, et al. The Asian Journal, . 2016;893:1. [doi 10.1088/1742-6596/893/1/012063]

[4] Bhandary, A., Prabhu, G. A., Rajinikanth, V., Thanaraj, K. P., Satapathy, S. C., Robbins, D. E., ... & Raja, N. S. M. (2020). Deep-learning framework to detect lung abnormality–A study with chest X-Ray and lung CT scan images. Pattern Recognition Letters, 129, 271-278

[5] Ozdemir O, Russell RL and Berlin AA 2019 A 3D Probabilistic Deep Learning System forDetection and Diagnosis of Lung Cancer Using Low-Dose CT Scans IEEE Transactions on Medical Imaging 1419-29.

[6] Alakwaa, W., Nassef, M., &Badr, A. (2017). Lung cancer detection and classification with 3D

[7] convolutional neural network (3D-CNN). Lung Cancer, 8(8), 409

[8] Alam, J., Alam, S., &Hossan, A. (2018, February). Multi-stage lung cancer detection and prediction using multi-class svmclassifie. In 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2) (pp. 1-4). IEEE

[9] Faisal, M. I., Bashir, S., Khan, Z. S., & Khan, F. H. (2018, December). An evaluation of machine learning classifiers and ensembles for early stage prediction of lung cancer. In 2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST) (pp. 1-4). IEEE

[10] Makaju, S., Prasad, P. W. C., Alsadoon, A., Singh, A. K., &Elchouemi, A. (2018). Lung cancer detection using CT scan images. Procedia Computer Science, 125, 107-114.

[11] Xie Y, Xia Y, Zhang J, Song Y, Feng D, Fulham M and Cai W 2018 Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest CT. IEEE transactions on medical imaging. 38(4) 991-1004.

[12] Song Q, Zhao L, Luo X and Dou X 2017 Using deep learning for classification of lung nodules on computed tomography images Journal of healthcare engineering

[13] Shen W, Zhou M, Yang F, Yu D, Dong D, Yang C, Zang Y and Tian J 2017 Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification Pattern Recognition 61 663-73

[14] Vamsidhar Enireddy, R P Shobha Rani ,Anitha, Sugumari Vallinayagam, T Maridurai, T Sathish, E Balakrishnan, "Prediction of human diseases using optimized clustering techniques", Materials Today: Proceedings, 2021

[15] A Vasantharaj, PS Rani, S Huque, KS Raghuram , " Automated brain imaging diagnosis and classification model using rat swarm optimization with deep learning based capsule network", International Journal of Image and Graphics, 2021

[16] Nishio M, Nishizawa M, Sugiyama O, Kojima R, Yakami M, Kuroda T and Togashi K 2018 Computer-aided diagnosis of lung nodule using gradient tree boosting and Bayesian optimization PloS one 13(4):e0195875

[17] K, S. A., P.Y.R., P., A, P., K, C. R., & Jagadeesh Gopal. (2023). Predict Admission of Confirmed COVID-19 Cases to ICU. *International Journal of Computer Engineering in Research Trends*, *10*(4), 199–203. https://doi.org/10.22362/ijcert.v10i4.22

[18] P, P., & J , K. (2023). Effective Predictor Model for Parkinson's Disease Using Machine Learning . *International Journal of Computer Engineering in Research Trends*, *10*(4), 204–209. https://doi.org/10.22362/ijcert.v10i4.27

[19] Swathi Velugoti , Revuri Harshini Reddy , Sadiya Tarannum , Sama Tharun Kumar Reddy (2022). Lung Nodule Detection and Classification using Image Processing Techniques. International Journal of Computer Engineering in Research Trends, 9(7), 144–119.