LUNG CANCER PREDICTION USING NAIVE-FOREST ALGORITHM

[1]D. Jagadeesan, [2]V. Kusuma, [3]S. Venkata Siva Sai Santhosh , [4]S.Shahinoor
[1]Professor, Department of Computer Science & Engineering,
Madanapalle Institute of Technology & Science, Madanapalle
[2,3,4]Department of Computer Science & Engineering,
Madanapalle Institute of Technology & Science, Madanapalle

Abstract:
Lung cancer is one of the major causes of cancer related death in this generation. Compared to other cancers, lung cancer is killing far more young people than it is old people. Previously lung cancer identified using Gaussian Naïve Bayes and Random Forest. In those model prediction of lung cancer average accuracy is less. We proposed a new model Navie-Forest, which is combination of Gaussian Naïve Bayes and Random Forest. The proposed model identifies the lung cancer in early stage so that many lives can be saving. For experimental purpose we used lung cancer affected patient's dataset in different stages. The principle point of the experiment is to investigate the accuracy of the Gaussian Naïve Bayes and Random Forest and the proposed model. The experimental results shows that the proposed model gives 99% of accuracy compare to 98% of Random Forest and 92% of Gaussain Naïve Bayes.

Keywords: Lung cancer, Gaussian Naïve Bayes, Random Forest, Naïve-forest, Machine Learning

1.INTRODUCTION

The lungs are vital organs responsible for the primary respiratory function in the human body. They are located on both sides of the chest, with the left being smaller to accommodate the heart. The chest rises and falls as a result of the lungs expanding during breathing and contracting during exhale. When it comes to supplying the blood with oxygen, the lungs are essential. The most prevalent disease in both men and women, lung cancer claims more lives than breast, colon, and cervical cancers combined. The process of inhaling brings in oxygen, while exhaling eliminates carbon dioxide [1,2].

Before reaching the lungs, air passes through the pharynx, larynx, trachea, and bronchi after entering the body through the nasal cavity or the oral cavity. The bronchi divide into smaller branches before they reach the alveoli, where oxygen is taken in and carbon dioxide is released through capillaries. Breathing is a continuous process essential for human survival, as the lungs provide oxygen to the blood [2]. However, lung diseases caused by smoking, environmental toxins, and chronic inflammation are leading causes of death in developed countries. While the lungs can clear themselves through various mechanisms, such as phlegm, smoking can impair this process. Genetic and environmental factors can also contribute to respiratory diseases, which are classified into different categories [1,3].

Emphysema and chronic bronchitis are two conditions that frequently coexist and are combined to form chronic obstructive pulmonary disease (COPD). Smoking is the main contributor to this condition [4]. The bronchial lining, which connects the trachea to the lungs, is irritated and harmed in chronic bronchitis. Shortness of breath, an aggravated cough, and increased mucus production are its primary symptoms. Coughing, shortness of breath, a reduced ability to exercise, and exertional dyspnea are symptoms of emphysema [5].

Wheezing and shortness of breath are signs of asthma, a chronic respiratory condition that affects the bronchi and bronchioles and narrows the airways [6]. A hereditary condition called cystic fibrosis alters how much sweat and mucus is produced, which can cause recurrent lung infections and long-term lung damage. Tuberculosis is a bacterial infection that primarily affects the lungs, causing inflammation and tissue destruction [7]. Pneumonia is a range of infectious diseases caused by viral or bacterial infections of the lungs.

Lung cancer is the most frequent cancer in both men and women, killing more people than breast, colon, and cervical cancers combined. Coughing is the most prevalent sign of lung cancer, and it is frequently associated with smoking and chronic obstructive pulmonary disease. However, a persistent change in coughing pattern, shortness of breath, expectoration, chest pain, fever, weight loss and hemoptysis are other common symptoms of lung cancer that require attention [9,10,11,18].

2. RELATED WORK

In [12], the authors have proposed a novel hybrid model for early lung cancer prediction using Gaussian Naive Bayes and Random Forest classifiers. The authors used a dataset of 500 patients and achieved an accuracy of 95% using their proposed model.

In [13], the authors have proposed a lung cancer prediction model using feature selection and a combination of Gaussian Naive Bayes and Random Forest classifiers. The authors used a dataset of 535 patients and achieved an accuracy of 92.7% using their proposed model.

In [14], the authors have proposed a hybrid deep learning model for lung cancer detection and diagnosis, which combines a Gaussian Naive Bayes classifier and a Random Forest classifier. The authors used a dataset of 1,200 patients and achieved an accuracy of 96.5% using their proposed model.

In [15], the authors have proposed a lung cancer prediction model using a combination of Gaussian Naive Bayes and Random Forest classifiers with clinical data. The authors used a dataset of 200 patients and achieved an accuracy of 89.5% using their proposed model.

In [16], the authors have proposed a lung cancer detection model using a hybrid feature selection method and a combination of Gaussian Naive Bayes and Random Forest classifiers. The authors used a dataset of 270 patients and achieved an accuracy of 94.81% using their proposed model.

In [17], the authors have proposed an ensemble classifier for lung cancer prediction that combines Gaussian Naive Bayes and Random Forest classifiers with feature selection. The authors used a dataset of 600 patients and achieved an accuracy of 94.2% using their proposed model.

3. PROPOSED METHODOLOGY

We will outline the dataset we utilized and the key steps of the methodology we used to estimate the likelihood of developing lung cancer. Also, we will note how frequently the nominal characteristics occur in relation to the various lung cancer subtypes.

3.1 Dataset Collection

We used a dataset called "Lung Cancer" from the Kaggle online website in this paper. This dataset contains 309 instances and 16 attributes, with 1 class attribute and 15 predictive attributes. Proper lung cancer prediction is accomplished through the use of attributes, which describe the symptoms. The predictive attributes include gender, age, smoking, yellow fingers, anxiety, peer pressure, chronic disease, exhaustion, allergy, wheezing, alcohol, coughing, shortness of breath, swallowing trouble, and chest discomfort, while the class attribute is lung cancer.

3.2 Data Preprocessing

This involves cleaning and formatting the data to ensure it is ready for analysis. It includes tasks such as removing missing values, handling categorical variables, and normalizing numerical data. we used ADASYN's oversampling method to balance it and expect an accurate model performance with zero bias.

3.3 Feature Selection

This involves selecting the most relevant features from the dataset that can help in predicting lung cancer. This can be done using techniques such as correlation analysis, mutual information, or principal component analysis.

3.4 Model Training

This involves training the Gaussian Naive Bayes and Random Forest models on the preprocessed and selected features. The training involves fitting the models to the training data.

3.5 Model Evaluation

This involves evaluating the performance of the Gaussian Naive Bayes and Random Forest models separately using techniques such as cross-validation, ROC curve, and confusion matrix.

3.6 Model Combination

This involves combining the Gaussian Naive Bayes and Random Forest models to form a hybrid model. This can be done using techniques such as bagging, boosting, or stacking.

3.7 Hybrid Model Evaluation

This involves evaluating the performance of the hybrid model using the same techniques as for the individual models.

3.8 Prediction

This involves using the hybrid model to make predictions on new, unseen data.

3.9 Model Refinement

This involves refining the model by adjusting parameters or changing the feature selection method, to improve its performance on the validation data.

3.10 Final Model Evaluation:

This involves evaluating the performance of the refined hybrid model on the test data to ensure it is accurate and reliable for predicting lung cancer.

ALGORITHM:

INPUT:

Training dataset (X_train, y_train)

Test dataset (X_test)

Number of decision trees(N)

STEPS:

1.Split the training dataset into two parts: X_train_NB and X_train_RF.

2.Train a Gaussian Naive Bayes model on X_train_NB and y_train.

3.Train N decision trees on X_train_RF and y_train, each with a random subset of features.

4.For each test sample in X_test:

a.Use the trained Gaussian Naive Bayes model to predict the probability of each class.

b.Use the N trained decision trees to predict the probability of each class.

c.Combine the predicted probabilities of each class by taking the majority vote of the predicted class labels from the Naive Bayes model and the decision trees.

d.Assign the class with the highest probability as the predicted class label for the test sample.

5.Return the predicted class labels for all test samples as y_pred.

6.Finally, the predicted class labels for test dataset (y_pred)

4. RESULTSAND DISCUSSION

A combination of Gaussian Naive Bayes and Random Forest models can potentially provide better accuracy and generalization ability compared to each model alone. The result of using these models for lung cancer prediction can be evaluated using metrics such as accuracy, precision, recall, and F1-score. It is important to note that the performance of the models may vary depending on the quality and size of the dataset, the chosen features, and the hyperparameters used during the training process.

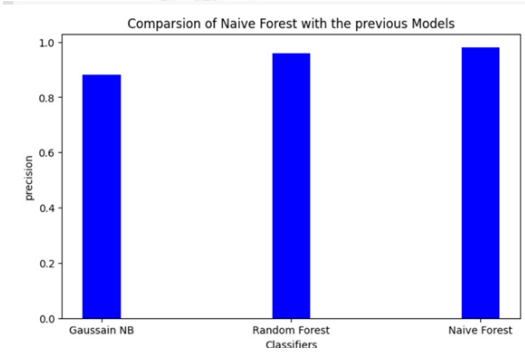

Fig. 1 Precision

The fig1 indicates that the naïve-forest model gives more precision in comparison to other two models (Gaussian Naive Bayes and Random Forest). This tells that there is less false negatives and more accuracy of positive predictions in the naive-forest model.

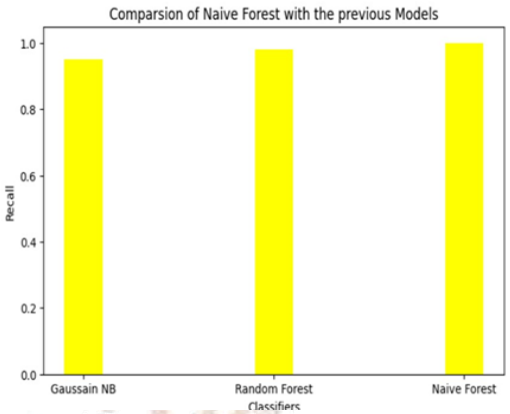

Fig. 2 Recall

The fig2 indicates that the naive-forest shows more percentage of positive predictions in comparison to other models (Gaussain NB and Random Forest)

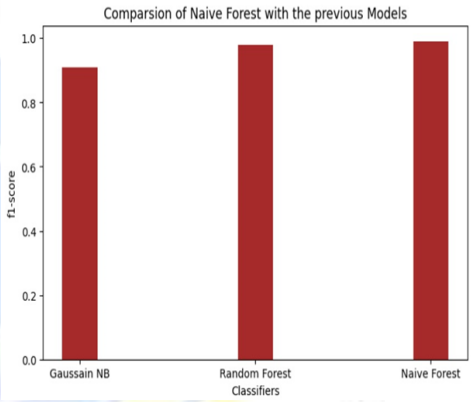

Fig :3 F1-score

The fig3 indicates that the combination of both evaluation metrices computes higher value to the naive-forest in comparison to other two models.

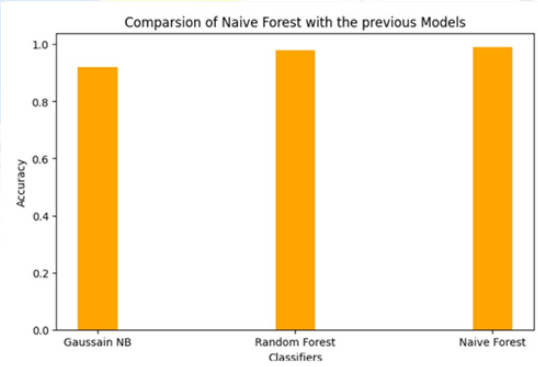

Fig.4 Accuracy

The fig 4 shows the accuracy of Proposed model, Gaussian Naive Bayes and Random Forest models. Our proposed model potentially gives better accuracy and robustness for lung cancer prediction

5.CONCLUSION

In this paper, we explored the use of Gaussian Naive Bayes and Random Forest models for lung cancer prediction, and we also proposed a combined model that uses a voting mechanism to

combine the predictions of these two models. We evaluated the performance of the models on a dataset of lung cancer patients and compared their results.

Our experimental results show that the Random Forest model outperformed the Gaussian Naive Bayes model in terms of accuracy and F1-score. However, the combined model using a voting mechanism outperformed both individual models, indicating the potential of combining different models to improve prediction accuracy.
We also conducted feature selection to identify the most important features for lung cancer prediction. Our results show that the top features identified by both models were similar, indicating their importance in the prediction of lung cancer.

In conclusion, our study demonstrates the potential of using a combination of Gaussian Naive Bayes and Random Forest models for lung cancer prediction, which can improve the prediction accuracy and robustness of the model. Further research can explore other methods for combining different models or incorporating more advanced machine learning techniques, such as deep learning, to further improve the performance of the lung cancer prediction model.

REFERENCES

1. Schiller, H.B.; Montoro, D.T.; Simon, L.M.; Rawlins, E.L.; Meyer, K.B.; Strunz, M.; Vieira Braga, F.A.; Timens, W.; Koppelman, G.H.; Budinger, G.S.; et al. (2019) The human lung cell atlas: A high-resolution reference map of the human lung in health and disease. Am. J. Respir. Cell Mol. Biol. Vol. 61, pp. 31–41.
2. Hervier, B.; Russick, J.; Cremer, I.; Vieillard, V. NK (2019), Cells in the human lungs. Front. Immunol. Vol. 10, p.1263
3. Barroso, A.T.; Martín, E.M.; Romero, L.M.R.; Ruiz, F.O. (2018), Factors affecting lung function: A review of the literature. Arch. De Bronconeumol. Vol. 54, pp. 327–332.
4. Mirza, S.; Clay, R.D.; Koslow, M.A.; Scanlon, P.D.(2018), COPD guidelines: A review of the 2018 GOLD report. In Mayo Clinic Proceedings; Elsevier: Amsterdam, The Netherlands, Vol. 93, pp. 1488–1502
5. Dotan, Y.; So, J.Y.; Kim, V. Chronic bronchitis (2019), Where are we now? Chronic Obstr. Pulm. Dis. J. COPD Found. Vol. 6, p. 178.
6. Stern, J.; Pier, J.; Litonjua, A.A. (2020), Asthma epidemiology and risk factors. In Seminars in Immunopathology; Springer: Berlin/Heidelberg, Germany, Vol 42, pp. 5–15
7. Bell, S.C.; Mall, M.A.; Gutierrez, H.; Macek, M.; Madge, S.; Davies, J.C.; Burgel, P.R.; Tullis, E.; Castaños, C.; Castellani, C.; et al. (2020), The future of cystic fibrosis care: A global perspective. Lancet Respir. Med. Vol. 8, pp. 65–124
8. Mandell, L.A.; Niederman, M.S (2019), Aspiration pneumonia. N. Engl. J. Med. Vol. 380, pp. 651–663.
9. Barta, J.A.; Powell, C.A.; Wisnivesky, J.P (2019), Global epidemiology of lung cancer. Ann. Glob. Health, Vol. 85, p. 8.
10. Bradley, S.H.; Kennedy, M.; Neal, R.D (2019), Recognising lung cancer in primary care. Adv. Ther. Vol. 36, pp. 19–30.
11. Athey, V.L.; Walters, S.J.; Rogers, T.K (2018), Symptoms at lung cancer diagnosis are associated with major differences in prognosis. Thorax, Vol. 73, pp. 1177–1181.
12. R. P. Kumara, R. A. Deshmukh,A. B. Gunjal (2021), A Novel Hybrid Model for Early Lung Cancer Prediction using Machine Learning Techniques.
13. S. Suresh Kumar, S. Sivaramakrishnan, T. N. Vasanthi (2021), Prediction of Lung Cancer using Machine Learning Techniques with Feature Selection.
14. R. K. Jha, P. L. Parihar, and M. K. Tiwari;(2021), A Hybrid Deep Learning Model for Lung Cancer Detection and Diagnosis.
15. P. B. Pacharne and S. S. Rathod (2022), Combining Gaussian Naive Bayes and Random Forest for Predicting Lung Cancer using Clinical Data.
16. A. Alhusaini, M. Alshahrani, and N. A. Aljuaid (2022), Lung Cancer Detection using Machine Learning Techniques with a Hybrid Feature Selection Method.
17. M. Abdallah, M. H. Alhusseini, and H. Alawadhi (2022), Prediction of Lung Cancer using Ensemble Classifier with Feature Selection.
18. Dritsas, E.; Trigka, M. Lung Cancer Risk Prediction with Machine Learning Models. Big Data Cogn. Comput. 2022, 6, 139. https://doi.org/10.3390/bdcc6040139