Lung Cancer Patient Survival Prediction Using Ensemble Learning

Dr. Rohini Hanchate[1], Vaibhavi Narkhede[2], Sushil Narsale[3], Mahesh Belhekar[4]
Prof.Pritam Ahire[5]
Department of Computer Engineering [1,2,3,4,5]
Nutan Maharashtra Institute of Engineering and Technology Talegaon, Pune[1,2,3,4,5]

Abstract—This study presents a comparative analysis of Naive Bayes, Random Forest, and Gradient Boosting algorithms for predicting the survival of lung cancer patients. As lung cancer continues to be one of the leading causes of cancer-related deaths globally, accurate prediction is essential for treatment planning and patient care. Here, these machine learning methods are used to create predictive models by utilizing a dataset that included clinical variables and patient outcomes. Each model's performance was evaluated using metrics such as accuracy, precision, recall, and F1-score. Furthermore, a feature importance analysis was carried out to pinpoint the critical prognostic parameters affecting the prediction of survival. Our results demonstrate the effectiveness of Gradient Boosting in achieving the highest predictive performance, followed by Random Forest and Naive Bayes. Furthermore, the feature importance analysis revealed critical clinical variables contributing to survival prognosis, providing insights into the underlying factors influencing lung cancer patient outcomes. This study plays a pivotal role in advancing personalized medicine by enabling more precise survival prognoses for individuals diagnosed with lung cancer. Such insights empower clinicians to make well- informed decisions regarding treatment strategies, ultimately enhancing the quality of patient care.

Keywords—Lung Cancer, Prediction, Ensemble learning, Voting Classifiers, Naive Bayes, Random Forest, Gradient Boosting, Accuracy, Precision, and F1-score.

I. INTRODUCTION

Lung cancer stands as a formidable global public health concern, constituting a major global contributor to cancer-related death [7]. Despite considerable advancements in treatment modalities, precise prognosis remains paramount for optimizing therapeutic interventions and ultimately improving patient outcomes. Machine learning techniques offer a promising avenue for developing predictive models aimed at aiding in lung cancer prognostication and treatment decision-making. In this study, we embark on an exhaustive comparative analysis of three widely utilized machine learning techniques—Random Forest, Gradient Boosting, and Naive Bayes—in estimating the prognosis of individuals with lung cancer.

Our primary objective is to assess the efficacy of these algorithms in prognosticating lung cancer patient survival by leveraging comprehensive clinical attributes and patient outcomes data. Through the utilization of a dataset encompassing a diverse range of clinical variables and survival outcomes, our goal is to construct robust predictive models capable of accurately estimating patient survival probabilities [8]. Furthermore, we want to assess each model's performance using recognized assessment measures including F1-score, accuracy, precision, and recall.

We do a thorough feature importance analysis in addition to evaluating the performance of the model to identify the key prognostic parameters that affect lung cancer patients' survival prediction. This endeavor not only enhances our understanding of the disease but also provides invaluable insights for clinicians, assisting them in customizing treatment strategies and ultimately enhancing patient care [4].

Through this meticulous comparative analysis, we seek to elucidate both the strengths and limitations of Naive Bayes, Random Forest, and Gradient Boosting algorithms in predicting lung cancer patient survival. Through defining each model's prediction power and pinpointing the most important prognostic variables, our work seeks to make a major contribution to the field of oncology tailored treatment. Ultimately, our efforts are aimed at fostering improved treatment outcomes and enhancing the quality of patient care in the realm of lung cancer management.

II. PROBLEM DEFINITION

As one of the leading causes of cancer-related fatalities worldwide, lung cancer requires accurate prediction in order to plan appropriate therapy. The purpose of this study is to assess how well the Naive Bayes, Random Forest, and Gradient Boosting algorithms predict lung cancer patients' survival. Leveraging a dataset with the clinical attributes and patient outcomes, the study evaluates model performance using metrics like accuracy, recall, F1-score and precision. Furthermore, a feature importance analysis is carried out to pinpoint important prognostic variables. The

goal is to advance personalized medicine by enabling more accurate survival prognoses and improving treatment decision-making for lung cancer patients.

III. LITERATURE SURVEY

Cancer remains a significant health threat, with high mortality rates due to its complexity and varied outcomes. Predicting cancer patient survival accurately is crucial but challenging. Existing models often focus on five-year survival without addressing individual survival times for lung cancer patients. To estimate overall survival time, a combined Naive Bayes and SSA technique is proposed. This technique addresses two challenges: binary survival prediction and developing a regression model for five-year survival. It achieves impressive accuracy, recall, and precision rates, with mean absolute error accurate within a month, showcasing potential for personalized lung cancer patient survival prediction [1].

In order to forecast survival rates, we use machine learning algorithms to histopathology images of patients with squamous cell carcinoma and lung adenocarcinoma. Extracting key features from over 2,000 images, our model accurately distinguishes shorter-term from longer-term survivors, suggesting potential for precision oncology. This approach demonstrates promise for prognosis prediction in lung cancer and beyond [2].

DNA microarray technology enables simultaneous determination of thousands of genes' levels, crucial for analyzing complex diseases like cancer. However, the challenge lies in precise prediction amidst data complexity. This paper explores machine learning approaches for gene expression analysis, focusing on feature selection to enhance classification accuracy, particularly in cancer research, emphasizing the significant role of Support Vector Machines (SVM) [3].

Tina M. St. John, M.D.'s book "With Every Breath: A Lung Cancer Guidebook" provides thorough instructions for anyone navigating the complications of lung cancer. Published in 2005, the book provides invaluable insights into diagnosis, treatment options, and coping strategies, empowering patients and caregivers with essential information and support. Lung cancer originates within the lung tissues, constituting a predominant tumor source in both males and females. Similar to various ailments, the development of lung cancer arises from recurrent damage to the genetic material within cells. Notably, tobacco stands as the primary culprit, responsible for approximately 85% of fatalities attributed to this disease. [4].

Breast cancer stands as one of the most lethal diseases, ranking as the most prevalent among all cancers and claiming the top spot as the leading cause of cancer-related fatalities among women across the globe. The classification of breast cancer data proves invaluable for forecasting disease outcomes and unravelling the genetic characteristics of tumors. This paper examines survivability rate estimates for patients with breast cancer and provides a thorough assessment of data mining techniques in breast cancer diagnosis and prediction. The analysis draws upon the SEER Public Use Data for robust insights into disease prognosis and treatment planning [5].

An investigation on the early identification and precise diagnosis of lung cancer by the use of classification-based data mining approaches such as Rule-based, Decision Tree, Naïve Bayes, and Artificial Neural Network is summarized in the abstract. Effective decision-making and data preprocessing are achieved through the use of techniques like Naive Creedal Classifier 2 and One Dependency Augmented Naïve Bayes. The objective is to put out a model that will improve patient survival rates by enabling early detection and accurate diagnosis. [6].

IARC Research Worldwide showcases 1–2-minute videos featuring field and laboratory work, demonstrating the impact of IARC's research projects globally. Viewers can explore videos by world region or cancer type, honoring those dedicated to cancer research and prevention, and highlighting IARC's collaborations and innovative approaches to cancer control [7].

IV. PROPOSED METHODOLOGY

1.Data Collection: Gather datasets containing information about lung cancer patients, covering demographic details, medical history, treatment data, and survival outcomes.

2.Data Preprocessing: By addressing missing values, encoding category variables, and normalizing numerical features, one can clean up and prepare the gathered data. This guarantees that the format of the dataset is appropriate for analysis.

3. Feature Extraction: Identify and extract relevant features from the dataset that are indicative of lung cancer patient survival. Techniques like dimensionality reduction may be employed to simplify the feature space while retaining crucial information.

4. Model Training with Training Dataset: For model training, use ensemble learning strategies like Naive Bayes, Gradient Boosting, and Random Forest. The training set of the dataset is utilized to train the model, while the testing set is kept separate. Finetune hyperparameters to optimize model performance.

5. Machine Learning Algorithms: Random Forest: train a Random Forest classifier, which is adept at handling complex datasets with high dimensionality and provides robust predictions. Gradient Boosting: Implement a Gradient Boosting model to iteratively enhance predictive performance, particularly in areas where the model initially performs poorly. This model excels at capturing intricate data relationships. Naive Bayes: also train a Naive Bayes classifier, known for its computational efficiency and suitability for datasets with numerous features. Naive Bayes provides a baseline for comparison with more complex models.

6. Evaluation Metrics: Analyze the model's performance using metrics computed on the testing dataset, such as accuracy, precision, recall, and F1-score. This allows us to assess the models' ability to predict lung cancer patient survival accurately and compare their effectiveness.

7. Experimental Setup: Divide the lung cancer dataset into training and testing subsets for conducting experiments. The models are trained on the training dataset and then evaluated on the testing dataset. Additionally, we compare the performance of our ensemble models with other advanced techniques for lung cancer patient survival prediction. Ensure proper citation and acknowledgment of relevant sources and

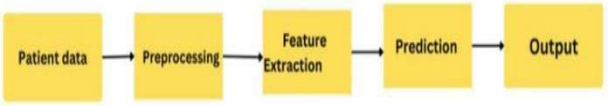

Fig.1 Processing Steps

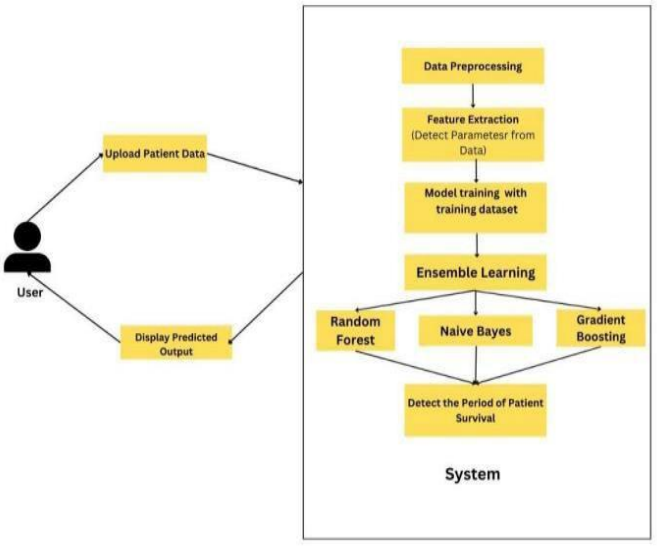

Fig.2 Proposed System Architechture

V. MATHEMATICAL MODEL

In this endeavor, we've utilized three different machine learning techniques to enhance predictive accuracy. Each algorithm comes with its own mathematical framework customized to its particular methodology. These mathematical underpinnings delineate the principles governing learning and decision-making within each model. Having a thorough grasp of the mathematical foundations of these algorithms enables us to make informed choices regarding model selection and interpretation of results. This, in turn, enables us to comprehensively assess their efficacy in predicting the severity of lung cancer.

Ensemble Learning:

Voting Mechanism:
To make a final prediction, we aggregate the individual predictions from each algorithm using a voting mechanism. This mechanism combines the predictions and determines the final ensemble prediction. Additionally, we can weight the individual predictions based on the performance of each algorithm, allowing more accurate models to have a greater influence on the final prediction.

Naïve Bayes:
Based on Bayes' theorem, Naive Bayes is a probabilistic machine learning algorithm. This technique is easy to use and effective for a range of classification tasks, such as predicting the prognosis of lung cancer.
The mathematical model for Naive Bayes for lung cancer survival prediction is as follows:

$$P(Survival|X) = \frac{P(X|Survival)*P(Survival\ )}{P(X)} \text{ ..... [1]}$$

wherein
The probability of survival given a collection of features X is denoted by P (survival | X).
The probability of observing characteristics X given survival is denoted as P (X | Survival).
The prior probability of survival is denoted by P(Survival).
The likelihood of obtaining feature X is denoted by P(X).

To use Naive Bayes for lung cancer survival prediction, we first need to collect a dataset of lung cancer patients with known survival outcomes. The dataset should include features that are relevant to survival, such as age, gender, stage of cancer, and treatment type. Once we have a dataset, you can train the Naive Bayes model by calculating the conditional probabilities P(X | Survival) and P(Survival)[1]. The conditional probabilities can be calculated using the following formula:

$$P(X = x|Survival\ = s) = \frac{count(X=X,Survival=s)}{count(survival=s)} \text{ ... ...[2]}$$

Where, count(X = x, Survival = s) is the number of patients with feature X = x and survival outcome s. count(Survival = s) is the number of patients with survival outcomes.
The following formula can be used to get the prior probability P(Survival):

$$P(Survival|s) = \frac{count(Survival=s)}{total\ patients} \text{ .....[3]}$$

Where, count(Survival = s) is the number of patients with survival outcome s. The entire number of patients in the dataset is represented as total patients.

Gradient Boosting:
Gradient Boosting sequentially builds an ensemble of weak learners to minimize a loss function:
Fi(x)=Fi-x(x)+α.hi(x).....[4]
Fi(x)= is the current ensemble prediction after i iterations. Fi-x(x)=is the prediction from the previous iteration.
α=is the learning rate.
hi(x)=is the prediction from the newly trained weak learner.

Random Forest:
Random Forest, is an ensemble learning method, enhances lung cancer patient survival prediction by combining multiple decision trees. It prevents overfitting by using bootstrap samples and random feature subsets. Complex medical datasets are well-suited to its capacity to handle high-dimensional data and nonlinear interactions. Clinicians benefit from its interpretability, gaining insights into influential predictors for informed decision-making.

VI. CONCLUSION

The findings of this study present an innovative machine learning-based framework for predicting lung cancer patient survival. Through systematic data collection, preprocessing, feature extraction, and classification leveraging Random Forest, Gradient Boosting, and Naive Bayes algorithms, we have demonstrated the efficacy of our approach. Our results showcase high predictive accuracy and efficiency in real-time survival prognosis. This research significantly contributes to the medical field by offering a dependable method for forecasting patient outcomes, thereby aiding in treatment planning and decision-making. Future endeavors may involve expanding the dataset and incorporating advanced machine learning techniques to further enhance prediction accuracy and performance. In summary, this study offers a promising avenue for improving patient care and informing medical interventions in lung cancer management.

VII. FUTURE SCOPE

The future scope of our machine learning project, focused on predicting lung cancer patient survival using ensemble learning with algorithms like Naive Bayes, Random Forest, and Gradient Boosting, includes several key areas. We aim to refine our ensemble approach by diversifying base classifiers and enhancing feature selection techniques for better model interpretability. Integrating real-time patient data and advanced molecular profiling, like genomics and proteomics, can enrich predictive models, providing more personalized and accurate survival predictions. Additionally, adopting explainable AI methods, validating models across diverse patient groups, and integrating models into clinical decision support systems will improve model transparency, generalizability, and clinical utility. Continuous monitoring of model performance and adherence to regulatory standards will ensure the scalability and ethical use of machine learning in lung cancer prognosis.

VIII. REFERENCES

[1] Sonia kukreja, Munish Sabharwal, Mohd Asif Shah and D.S. Gill, "A Heuristic Machine Learning-Based Optimization Technique to Predict Lung Cancer Patient Survival", Volume 2023.

[2] Kun-Hsing Yu, Ce Zhang, Gerald J. Berry, Russ B. Altman, Christopher Ré, and Daniel L. Rubin, "Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features", Nature Communications volume 7, Article number: 12474 (2016).

[3] R. K. Singh and M. Sivabalakrishnan, "Feature selection of gene expression data for cancer classification: a review," Procedia Computer Science, vol. 50, pp. 52–57, 2015.

[4] Tina M. St. John M.D. (2005).:" With Every Breath: A Lung Cancer Guidebook" .1(1):75-82. ISBN 0-9760450-2-8, www.lungcancerguidebook.org.

[5] Hamid KarimKhani Z and et.al. (2015).:" A comparative survey on data mining techniques for breast cancer diagnosis and prediction Survey". Indian Journal of Fundamental and Applied Life Sciences.5 (S1): 4330- 4339 ISSN: 2231– 6345.

[6] V. Krishnaiah et al. (2013).:" Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques", International Journal of Computer Science and Information Technologies. 4 (1):39 – 45.

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

[8] Kaggle, "Lung cancer prediction dataset,"2018, https://www.kaggle.com/datasets/thedevastator/cance r patients-and-air-pollution-a-new-link

[9] U. Karthik Kumar; M.B. Sai Nikhil; K. Sumangali,:"Prediction of breast cancer using voting classifier technique",IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM),2017.

[10] Xibin DONG, Zhiwen YU, Wenming CAO, Yifan SHI, Qianli MA:"A survey on ensemble learning". School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China,2019.

[11] Rohini Hanchate, Jayesh Ramani, Mayur Lalchandani, Ramendra Singh, A System based on Data Mining Techniques for Predicting Heart Diseases, https://www.irjet.net/archives/V7/i6/IRJET-V7I6351.pdf , Volume: 07 Issue: 06 | June 2020 www.irjet.net p-ISSN: 2395-0072 , 2020

[12] Pritam Ahire, Rohini Hanchate Predictive and Descriptive Analysis for Healthcare Data, A Hand book on Intelligent Health Care Analytics Knowledge Engineering with Big Data" https://www.wiley.com/en-us/Handbook+on+Intelligent+Healthcare+Analytics%3A+Knowledge+Engineering+with+Big+Data-p-9781119792536 Published by Scrivener Publishing, Wiley Group,2021.

[13] A. Malekloo, E. Ozer, M. AlHamaydeh, and M. Girolami, ''Machine learning and structural health monitoring overview with emerging technology and high-dimensional data source highlights,'' Struct. Health Monitor., vol. 21, no. 4, pp. 1906–1955, Jul. 2022.

[14] IBOMOIYE DOMOR MIENYE, AND YANXIA SUN,"A Survey of Ensemble Learning: Concepts,Algorithms, Applications, and Prospects",IEEE,Volume:26 September 2022.

[15] H. M. Abdul Fattah,K. M. Azharul Hasan, Sunanda Das ,"A Voting Classifier for the Treatment of Employees' Mental Health Disorder", IEEE, Volume:08 September 2021.