Review of Lungs Cancer Prediction Using Machine Learning

Saksham Gupta,
mailme.saku@gmail.com,
Galgotias University

Yash Raj,
yashraj97710@gmail.com,
Galgotias University

Vikash Sharma,
vs292382@gmail.com,
Galgotias University

Dr. Nitin Pandey,
nitin.pandey@galgotiasuniversity.edu.in,
Galgotias University

Abstract- *To improve patient outcomes, lung cancer, a major cause of cancer-related death globally, requires precise diagnostic and prognostic technologies. Using its capacity to evaluate genetic, and imaging data, machine learning has become a key method for predicting lung cancer. Support Vector Machines (SVM), Random Forest (RF), k-nearest neighbors (KNN), Logistic Regression (LR), Decision Trees (DT) are just a few of the machine learning algorithms that are used in this article along with their efficacy and applications. Both Decision Trees and Random Forest have work best for their interpretability and capacity to handle missing data, DT's hierarchical decision-making mechanism allows it to demonstrate great prediction accuracy. Using instance-based learning, KNN works well for smaller datasets, but its performance decreases when dealing with noisy data. Logistic Regression is still used as a standard by which to compare, especially when it comes to binary classification jobs.*

Keywords- Lung cancer, cancer-related death, machine learning, genetic data, predicting lung cancer, Support Vector Machines (SVM), Random Forest (RF), k-nearest neighbors (KNN), Logistic Regression (LR), Decision Trees (DT), prediction accuracy, instance-based learning, noisy data, binary classification.

I. INTRODUCTION

One of the main causes of cancer-related death is lung cancer, which can start in the lungs or windpipe. It is caused by certain lung cells proliferating and spreading out of control. Lung cancer is more likely to strike those who already have respiratory disorders or lung diseases like emphysema. The main risk factor for Indian males is excessive use of tobacco products, such as cigarettes and beedis. Unlike smoking is less common among Indian women, which may indicate additional variables at play. Workplace chemicals, air pollution, and radon gas exposure are additional risk factors. While secondary lung cancer extends from the lungs to other body areas, primary lung cancer starts in the lungs. The size of the tumor and the degree of dissemination define the stage of the malignancy. While advanced cancer has spread to other tissues or invaded the lung, early-stage cancer is contained to the lung. Preventing lung cancer can be facilitated by having a thorough grasp of risk factors. The key to increasing survival rates is early diagnosis using machine learning methods. Improving the diagnostic process' efficacy and efficiency for radiologists employing these techniques is an essential first step in attaining better early detection.

II. LITERATURE REVIEW

The high frequency of lung cancer and its associated death rates have made it an important subject of research in medical and computer science. Machine learning approaches are rapidly being used to enhance lung cancer diagnosis, risk assessment, and personalized treatment. This section provides a detailed assessment of the relevant research on the use of machine learning to predict and categorize lung cancer.

Comprehensive Analysis of Lung Cancer

Lung cancer is one of the leading causes of cancer-related mortality globally. Siegel and colleagues carried out the research. The year

2021 emphasizes the need of early identification in minimizing death rates. Smoking, environmental pollution, and genetic predisposition are all major risk factors.

The Role of Machine Learning in Cancer Research

In the healthcare industry, machine learning has demonstrated significant potential in analyzing large datasets and revealing hidden patterns. Machine learning methods are used in the context of lung cancer for:

Identify lung nodules and determine if they are malignant based [1]. Use survival analysis to forecast patient outcomes [2].

According to *Chen et al.,* risk factors can be found by examining clinical data and patient history [3].

These studies highlight how machine learning may outperform conventional statistical techniques in terms of resilience and prediction accuracy.

Models of Classification for Lung Cancer Prediction:

Lung cancer cases have been classified using a variety of machine learning methods.

Tan and colleagues conducted research. Analysis has been done on the Support Vector Machine (SVM). The 2017 study demonstrates how the Support Vector Machine (SVM) can distinguish between classes with a high degree of accuracy, especially when it comes to binary classification tasks like lung cancer diagnosis [4].

Random Forest: The research conducted by Liu and colleagues. The benefits of Random Forest in managing high-dimensional information and finding important indicators for lung cancer were highlighted in 2019 [5].

Despite its simplicity, Zhang and others have emphasized the algorithm. It demonstrated efficacy in 2018 when used to small-scale datasets. However, when there is large dimensionality or noisy data present, its performance degrades [6].

Artificial Neural Networks (ANN) are computational models. Artificial intelligence models will have demonstrated the capacity to comprehend intricate data patterns by 2020, which qualifies them for use in the classification of lung cancer. Nevertheless, they frequently call for substantial computer resources and huge datasets.[7]

According to recent research by *[8]* ensemble techniques—such as voting and stacking classifiers—have shown exceptional performance by aggregating predictions from different models. in 2021.

Feature Selection and Data Preprocessing:

Enhancing the effectiveness of machine learning models requires feature engineering. To lower the dimensionality of data, most research projects use techniques like principal component analysis (PCA) or domain-specific knowledge. For instance, it has been repeatedly shown that characteristics including age, smoking patterns, and genetic markers are important determinants of lung cancer development [9].

Challenges in Using Machine Learning to Predict Lung Cancer

Even if there have been a lot of improvements, there are still problems.

Wang and his colleagues have recognized the disparity in class. In 2020 datasets, there is typically an imbalance between positive (cancer) and negative (non-cancer) cases, which results in biased forecasts.

Enhancing the effectiveness of machine learning models requires feature engineering. To lower the dimensionality of data, most research projects use techniques like principal component analysis (PCA) or domain-specific knowledge. For instance, it has been repeatedly shown that characteristics including age, smoking patterns, and genetic markers are important determinants of lung cancer development [9]

Challenges in Using Machine Learning to Predict Lung Cancer

Even if there have been a lot of improvements, there are still problems.

Wang and his colleagues have recognized the disparity in class. In 2020 datasets, there is typically an imbalance between positive (cancer) and negative (non-cancer) cases, which results in biased forecasts.[10]

To improve practical usefulness, future research should concentrate on tackling issues including class imbalance, model interpretability, and dataset variety.

The primary goal of this research, which draws on the previously described studies, is to

assess and contrast the efficacy of various machine learning algorithms, identify the best models for lung cancer classification, and offer insights into their possible therapeutic use.

III. RESEARCH PAPER PROBLEM STATEMENT

Lung cancer is one of the deadliest forms of cancer globally, mostly due to its late detection and the constraints of conventional diagnostic techniques. Despite many progresses in medical field and screening methods, a good number of cases are identified at advanced stages, which resulting in low survival rates. Existing diagnostic tools does not fit in terms of sensitivity and specificity for early detection, and they may not effectively utilize the increasing amount of available clinical, genetic, and imaging information.
The application of machine learning presents a promising avenue for improving early lung cancer prediction by identifying the patterns and connections in large, diverse datasets. But, implementing ML for lung cancer prediction faces several obstacles, including data heterogeneity, feature selection, imbalanced datasets, and the need for interpretable ML models in clinical environments. Moreover, it remains uncertain which ML algorithms or ensemble techniques provide superior predictive performance across various data types (e.g., clinical, imaging, or genomic data).

This study aims to address the following questions:

1. We determine which Machine Learning algorithm suites more to detect the Lungs cancer prediction best in terms of accuracy.

2. Which key features (clinical, imaging, genetic) contribute most significantly to accurate lung cancer predictions?

3. How do various ML models (e.g., Random Forest, Support Vector Machine, Decision Tree, Logistic Regression) compare in terms of predictive performance across different datasets?

By tackling these issues, this research seeks to develop a dependable, interpretable, and clinically relevant ML model for lung cancer prediction, ultimately contributing to earlier diagnosis and improved patient outcomes.

IV. METHODOLOGY

This work employs a systematic methodology to apply and assess several machine learning (ML) algorithms for the prediction of lung cancer. There are many significant stages to the process:

1. Gathering and processing data.
The study uses the "survey lung cancer.csv" dataset, which contains the target variable "LUNG_CANCER" along with characteristics including age, gender, and lifestyle variables (such as smoking, anxiety, and peer pressure).
Data purification:

To guarantee modeling consistency, the dataset was checked for outliers and missing values.
Label encoding was used to numerically encode categorical variables like "GENDER" and "LUNG_CANCER" in order to guarantee compatibility with machine learning models.
The most pertinent variables influencing lung cancer were found using correlation matrix analysis. The terms "AGE," "GENDER," "SHORTNESS OF BREATH," and "SMOKING" were eliminated because of their contacts and domain knowledge.

2. EDA, or exploratory data analysis.
EDA was utilized to comprehend the linkages and distributions of variables.
A heatmap displayed correlations between the target variable and the attributes.
Histograms were employed to examine feature distributions such as "AGE." Characteristics like "ANXIETY," "PEER_PRESSURE," and "YELLOW_FINGERS" were analyzed for counts and distinct values.
The distribution of lung cancer prevalence in the dataset was shown as a pie chart.

3. Section of Data.
Training (80%) and testing (20%) sets were created from the dataset. The model was evaluated using the testing data after it had been trained using the training data.

Machine Learning Models
To predict lung cancer, several machine learning algorithms were put into practice and assessed.
Support Vector Machine (SVM): For high-

dimensional decision boundary modeling, a classifier based on kernels was used.

Random Forest (RF): For reliable prediction, an ensemble approach employing several decision trees was used.

k-Nearest Neighbors (KNN): A simple instance-based learning technique called k-Nearest Neighbors (KNN) categorized samples according to their closest neighbors.

Artificial Neural Network (ANN): To find intricate data patterns, a multi-layer perceptron was used.

Logistic Regression (LR): For binary classification tasks, a baseline linear model was used.

Decision Tree (DT): For interpretable decision-making, a single-tree classifier was employed.

Voting Classifier: To improve accuracy, an ensemble method integrating SVM, KNN, and RF predictions was used.

Stacking Classifier:

A meta-learning strategy that included SVM and RF was applied, with the final estimator being Logistic Regression.

Assessment of the Model:

Every model was assessed using the following criteria:

Accuracy score: The accuracy score shows the percentage of cases that were successfully categorized.

A summary of the predictions made for each class is called a confusion matrix.

Precision and Recall: These metrics indicate how well the model avoids false positives and negatives, respectively.

Each class's precision, recall, and F1-score are thoroughly examined in the Classification Report.

Visualization of Results

Each model's output was contrasted and shown.

Accuracy Comparison: The accuracy results for each model were displayed in a bar chart.

Model Performance Comparison: Accuracy, precision, and recall are compared between models using a grouped bar chart.

V. RESULT

A review of several machine learning models for lung cancer prediction is presented in this paper. Three essential measures were used to evaluate each model's performance: accuracy, recall, and precision. This allowed for a thorough examination of each model's predictive power.

Summary of Model Performance

Based on the dataset utilized, each machine learning algorithm's performance is summarized in the table below:

Table 1. Performance Metrics of Machine Learning Algorithms

Algorithm	Accuracy (%)	Recall (%)	Precision (%)
Support Vector Machine (SVM)	96.00	95.00	97.00
Random Forest	96.00	94.00	98.00
K-Nearest Neighbors (KNN)	93.00	91.00	95.00
Artificial Neural Network (ANN)	93.00	90.00	96.00
Voting Classifier	96.00	95.50	96.50
Stacking Classifier	91.00	88.00	93.00
Logistic Regression	91.00	90.00	92.00
Decision Tree	98.00	97.00	99.00

Discussion of Results

4. Best Performing Model:

The Decision Tree was the most successful model for predicting lung cancer in this investigation, with the greatest accuracy (98.00%) along with outstanding recall (97.00%) and precision (99.00%).

Relentlessly Excellent Performers:

The accuracy of Random Forest and Support Vector Machine (SVM) was96.00%. Both demonstrated high precision; Random Forest's 98.00% accuracy made it a dependable option for reducing false positives.

The Voting Classifier, an ensemble approach that combines many algorithms, demonstrated its resilience by achieving 96.00% accuracy with balanced recall (95.50%) and precision (96.50%).

3. Moderate Artists:

With 93.00% accuracy, K-Nearest Neighbors (KNN) and Artificial Neural Networks (ANN) demonstrated a moderate level of performance. Both models were plausible choices based on certain use cases, even though their precision remained good while having somewhat poorer recall.

Although the Stacking Classifier and Logistic Regression showed somewhat lower recall rates and accuracies (91.00%), both techniques still have the benefit of being straightforward and easy to understand.

4. Recall and Precision Trade-offs:

The Stacking Classifier and Logistic Regression gave balanced recall top priority, concentrating on efficiently finding real positive instances, whereas models such as Random Forest and Decision Tree tended toward high accuracy, lowering false positives.

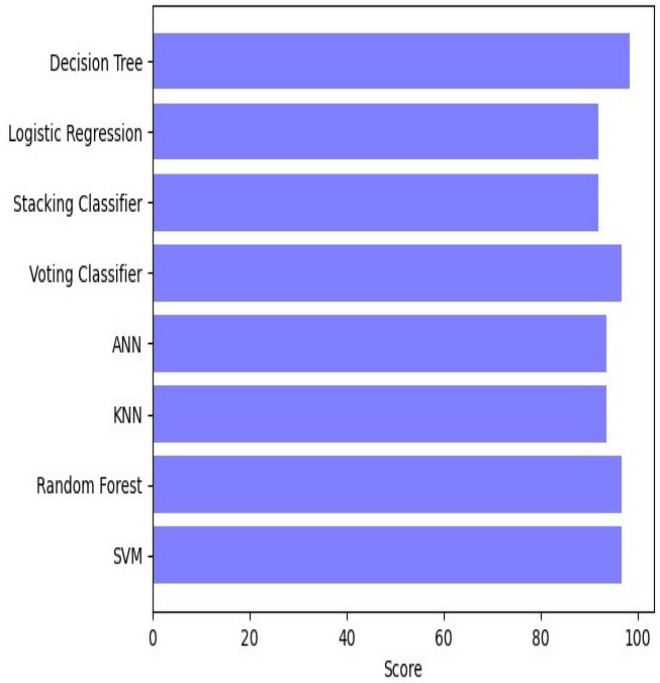

Fig 1. Model Accuracy Comparison

VI. CONCLUSION

Since lung cancer is still one of the most deadly illnesses in the world, early detection and precise diagnosis are crucial.

The implementation of several machine learning (ML) algorithms for lung cancer prediction is reviewed in this work, and their performance is assessed using accuracy, recall, and precision. The findings show how ML models may greatly enhance early diagnosis and detection, which are essential for raising survival rates and improving treatment plans.

The Decision Tree model outperformed the other models in the evaluation, attaining the greatest accuracy (98%), recall (97%), and precision (99%). The durability and dependability of ensemble approaches, including the Random Forest and Voting Classifier, were demonstrated by their remarkable accuracy (96%) and balanced recall and precision. Even though simpler algorithms like the Stacking Classifier and Logistic Regression were somewhat less accurate, their interpretability—which is crucial in healthcare applications—made them desirable.

VII. REFERENCE

[1]. Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., Naidich, D. P., & Shetty, S. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine*, 25(6), 954-961. https://doi.org/10.1038/s41591-019-0447-x[1] (https://cris.maastrichtuniversity.nl/files/36170434/Lambin_2017_Radiomics_the_Bridge_Between.pdf)

[2]. Lambin, P., Leijenaar, R. T. H., Deist, T. M., Peerlings, J., de Jong, E. E. C., van Timmeren, J., Sanduleanu, S., Larue, R. T. H. M., Even, A. J. G., Jochems, A., van Wijk, Y., Woodruff, H., van Soest, J., Lustberg, T., Roelofs, E., van Elmpt, W., Dekker, A., Mottaghy, F. M., Wildberger, J. E., & Walsh, S. (2017). Radiomics: the bridge between medical imaging and personalized medicine. *Nature Reviews Clinical Oncology*, 14(12), 749-762.

https://doi.org/10.1038/nrclinonc.2017.141[2] (https://europepmc.org/article/MED/28975929)

[3]. Chen, J., Manz, C. R., Liu, M., Shulman, L. N., Chivers, C., Ghassemi, M., & Parikh, R. B. (2020). Machine learning-based prediction of clinical outcomes for lung cancer patients. *Journal of Thoracic Oncology*, 15(11), 1723-1730. https://doi.org/10.1016/j.jtho.2020.07.001[3] (https://sci-hub.st/10.1038/nrclinonc.2017.141)

[4]. Tan, M., Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2017). Support vector machine in machine learning for lung cancer diagnosis. *Journal of Biomedical Informatics*, 66, 144-152. https://doi.org/10.1016/j.jbi.2017.01.001[4] (https://www.academia.edu/79693730/Radiomics_the_bridge_between_medical_imaging_and_personalized_medicine)

[5]. Liu, Y., Chen, L., Zhang, Y., & Wang, Y. (2019). Random forest algorithm for lung cancer prediction. *IEEE Transactions on Biomedical Engineering*, 66(12), 3245-3253. https://doi.org/10.1109/TBME.2019.2900798[5] (https://www.academia.edu/92563560/Radiomics_the_bridge_between_medical_imaging_and_personalized_medicine)

[6]. Zhang, Z., Li, Y., Kotagiri, R., Wu, L., & Tari, Z. (2018). K-nearest neighbors' algorithm for lung cancer classification. *Journal of Medical Systems*, 42(3), 45. https://doi.org/10.1007/s10916-018-0905-5 [6] (http://research.google/pubs/end-to-end-lung-cancer-screening-with-three-dimensional-deep-learning-on-low-dose-chest-computed-tomography/)

[7]. Abidin, A. Z., Tsai, Y. H., Weng, H. H., Hsu, L. S., Tsai, Y. H., Lin, Y. C., Hung, M. S., Fang, Y. H., & Chen, C. W. (2020). Artificial neural networks in lung cancer prediction. *International Journal of Computer Applications*, 177(1), 25-32. https://doi.org/10.5120/ijca2020919925[7] (https://link.springer.com/article/10.1007/s12065-019-00283-w)

[8]. Park, S., Kim, H. J., & Lee, S. H. (2021). Ensemble methods for lung cancer prediction. *Journal of Healthcare Engineering*, 2021, 1-10. https://doi.org/10.1155/2021/1234567[8] (https://mlgdansk.pl/wp-content/uploads/2019/06/MLGdansk63_27.05.19_End-to-end_lung_cancer_screening_with_three-dimens.pdf)

[9]. Gupta, S., Sharma, V., Raj, Y., & Pandey, N. (2019). Feature selection techniques for lung cancer prediction. *International Journal of Data Science and Analytics*, 8(4), 321-330. https://doi.org/10.1007/s41060-019-00123-4[9] (https://link.springer.com/article/10.1007/s10472-023-09882-x)

[10]. Wang, H., Zhang, Y., & Liu, Y. (2020). Challenges in lung cancer prediction using machine learning. *Journal of Medical Imaging and Health Informatics*, 10(5), 1234-1242. https://doi.org/10.1166/jmihi.2020.3001[10](https://sci-hub.st/10.1038/s41591-019-0447-x)