Leveraging Machine Learning for Lung Cancer Risk Assessment Based on Survey Insights

Sumit Mhaske[1]; Mandar Dakhorkar[2]; Vanshika Khiyani[3]; Rudrani Patil[4]; Ganesh Shelke[5]

[1;2;3;4]Student; [5]Assistant Professor

[1,2,3,4]Vishwakarma Institute of Information Technology, Pune, India
[5]Vishwakarma Institute of Technology, Pune, India

Publication Date: 2025/04/24

Abstract: Lung cancer is still among the top cancers to cause cancer death in humans around the world. It has a lot to do with lifestyle and smoking- individual factors that contribute to lung cancer development. This research study seeks to analyze the viability of the machines through algorithms for the likely risk prediction of lung cancer through survey data- that is, symptoms, behavioral traits, and demographic data. The dataset consists of information such as smoking habits along with anxiety levels, fatigue, and other symptoms employed. Various machine learning models were trained and evaluated on Logistic Regression, Decision Tree, Random Forest, Naive Bayes, and Support Vector Machines (SVM) algorithms. Among those, Random Forest proved to be the best predictor giving about 96.7% accuracy and strong precision and recall values, indicating its effectiveness in identifying high-risk subjects. This research indicates that machine learning can be applied to healthcare for early diagnosis and screening.

Keywords: *Lung Cancer; Machine Learning; Prediction Model; Survey Data; Random Forest; Health Informatics; Risk Assessment.*

How to Cite: Sumit Mhaske; Mandar Dakhorkar; Vanshika Khiyani; Rudrani Patil; Ganesh Shelke. (2025). Leveraging Machine Learning for Lung Cancer Risk Assessment Based on Survey Insights. *International Journal of Innovative Science and Research Technology*, 10(4), 1102-1109. https://doi.org/10.38124/ijisrt/25apr1335

I. INTRODUCTION

Lung cancer is one of the major causes of cancer death, approximating about 1.8 million deaths globally per year. One of the major issues in lung cancer treatment is diagnosing the disease late, as earlier-stage symptoms are usually not recognized or are misunderstood as trivial conditions. Conventional methods of diagnosis, such as biopsies, CT scans, and chest X-rays, while sure of their effectiveness, are invasive, expensive, and therefore not always available, particularly in developing countries. Thus, it has become increasingly important to find alternative tools for diagnosis that are low-cost and non-invasive to the patient and that may be of assistance in the early identification of individuals at risk. Lifestyle factors such as smoking, alcohol intake, and pollutant exposure, as well as physical symptoms like fatigue, coughing, wheezing, and chest pains, have a strong association with lung cancer risk and can serve as reasonable indicators when studied combined.

In the past few years, the evaluation of healthcare has taken a considerable leap with machine learning. With this capability, machine learning can uncover hidden pivotal patterns within vast and intricate datasets. The ability of machine-learning models to analyze multiple risk factors at the same time and provide predictive insights in the support of clinical decision-making. In this study, we utilize a publicly available survey dataset comprising demographic and behavioral data to design predictive models for lung cancer risk. With the application of various supervised learning algorithms, viz., Random Forest, Decision Tree, Logistic Regression, Naive Bayes, and Support Vector Machines (SVM), we seek to assess their suitability to offer guidance on detecting individuals at risk for developing lung cancer. The research substantiates the use of survey data for predictive modeling, which then cushions the development of AI-powered tools for screening purposes that can facilitate timely intervention and reduce mortality rate incidences.

II. LITERATURE SURVEY

This part of the paper gathers all the researches and information from different sources:

A Survey on Detection of Lung Cancer Using Different Image Processing Techniques uses Computed Tomography (CT) scans over Positron Emission Tomography (PET), and Magnetic Resonance Imaging (MRI) due to higher image

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr1335

clarity and reliability (Reference Paper). The accuracy is around 85-90%. This study emphasis that early detection can save lives upto 60-70% [1].

A study on lung cancer detection was done using MATLAB software, the database was obtained from Cancer imaging Archive (TCIA). Median filter was applied to the lung CT images to prevent the edges. Image enhancement was done using contrast adjustment [2].

Another study presented a model to effectively detect SCLC. In comparison to other models this one stood to be the most effective model with a classification performance of 91.5% [3].

One of the study shows how challenging it is to automate categorization of lung cancer. And this detection has a lot of potential when employing Machine learning. Therefore Johnson Reducer algorithm of gene expression was been reported in this study [4].

A study uses a dataset of 1000 images of chest scans for different lung cancer. In this study multiple machine learning algorithms are used and compared of which CNN turns out to be the best algorithm. VGC-16 was implemented on data set which helped to check severity [5].

Machine learning, deep learning, and Artificial intelligence methods were used in one of the studies. Comparison of various algorithms including convolutional neural networks (CNNs), Random forest algorithm, ensemble extreme boosting (XGBoost) algorithm, Support Vector Machine (SVM), AdaBoost classification model (ADB-C), Long Short-Term Memory Networks (LSTM) was done [6].

Baker, Q. B in his study used various deep learning methods to predict survival time of Non-Small Cell Lung Cancer (NSCLC) patients. Computerized tomography (CT) images dataset was being used that contained data of 300 patients and concordance index (C-index) was used to evaluate the models. An accuracy of 70% was gained [7].

K, A. proposed a real-time lung cancer detection system that leverages the power of Convolutional Neural Networks (CNN) and Generative Adversarial Networks (GAN). The system classifies lung cancer in LDCT images. The result highlights potential of combining GAN - generated synthetic images with CNN-based classification [8].

S. H. Begum proposed a novel approach based on Clustering and Classification using a CNN based image processing algorithm. The proposed study applied SMOTE algorithm for both balanced and unbalanced dataset [9].

A. Jawaid implemented lung cancer detection using CT scan by employing image processing, segmentation, feature extraction and classification of lung cancer with stage detection. It is built on android smartphones and aims to make lung cancer detection as simple as possible. The accuracy turns out 96.5% [10].

A. A. Sagaya Priscilla in his study used various performance measures the Adam optimizer provided the highest accuracy among the optimizers used [11].

III. METHODOLOGY

The research methodology comprises a series of steps including data collection, preprocessing, feature transformation, model selection, training, and finally performance evaluation. Each step is crucial for accurately developing a machine learning model for predicting lung cancers based on survey responses.

➢ *Data Collection*

The dataset applied into this study is "Survey Lung Cancer", which is publicly archived and contains responses collected from 309 individual subjects. Furthermore, the dataset consists of up to 16 features capturing demographic data (e.g., age, gender), behavioral patterns (e.g., smoking, alcohol consumption), and medical symptoms (e.g., coughing, shortness of breath, chest pain). The target variable is, as already described, LUNG_CANCER, and it is a binary variable that indicates whether the subject is reported to be newly diagnosed (YES) or has never been diagnosed with lung cancer (NO). Because the dataset was based on self-report surveys, it serves as a non-invasive means of data collection for predictive modeling.

➢ *Data Preprocessing*

Several preprocessing steps were implemented to maintain both model performance and consistency:

- Handling Missing Values: The dataset was scanned for null or missing values. All records with missing values in the target column were purged to ensure incomplete labelling did not propagate during training.
- Encoding Categorical Variables: The GENDER column was label encoded (e.g., Male=1, Female=0) and the target column LUNG_CANCER was encoded to binary format (YES=1, NO=0).
- Feature Scaling: Since machine learning models such as support vector machines and logistic regressions are sensitive to the scale of input data, numerical features were standardized using StandardScaler, which transforms them so that they have mean 0 and standard deviation 1.
- Splitting the Dataset: The data were split on the training and testing sets through an 80:20 ratio. The training set served for building the model while the other was for testing.

➢ *Model Selection and Training*

Five different machine learning algorithms were selected for the study, primarily due to their popularity and usage in different classification problems. They include the following:

- Decision Tree Classifier: A model based on rules that divides the dataset on branches according to features for the purpose of prediction.
- Random Forest Classifier: This is an ensemble procedure built from different decision trees and aggregates results in

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr1335

such a way that it improves the accuracy and limits overfitting.

- Logistic Regression: Statistical method of using a logistic function to model the probability of a binary outcome.
- Naive Bayes Classifier: This is also known as a probabilistic model that's based on Bayes' theorems and it's assumed that predictors are independent.
- Support Vector Machine (SVM): A classification technique that seeks to find the best hyperplane for the data points of various classes.

Each model was trained using the scaled training data and optimized with default hyperparameters.

➢ *Performance Evaluation*

To check the performance of each model, the performance measures calculated on the test set are:

- Accuracy: Ratio of all correct predictions by the model.
- Precision: Absolute ratio of true positive predictions to the total predicted positives which reflects how well the model performs at avoiding false positives.
- Recall (Sensitivity): Ratio of true positive predictions to all positive actuals, showing in how much the model detects actual cases having lung cancer.
- F1-Score: harmonic mean of precision and recall, which balances the two.
- Classification Report: A detailed breakdown for both classes in terms of precision, recall, F1-score, and support.

It helps to comprehensively compare all five models. Among them, the Random Forest Classifier has the highest accuracy, precision, and recall, thus the best-performing model for this task.

IV. RESULTS AND DISCUSSION

This is where we present the results of the evaluation, followed by some discussion on the highlighted results.

➢ *Age Distribution of Participants:*

Fig 1 shows the histogram that exhibits the age distribution of the participants in the domain of this study. Most individuals lie in the age group of 55-70 years, which is more representative of the middle-aged to elderly category. The distribution is approximately typically normal, with a peak toward age 60. Since lung cancer occurs mainly in older adults, this age distribution fits nicely with studying risk in a susceptible population.

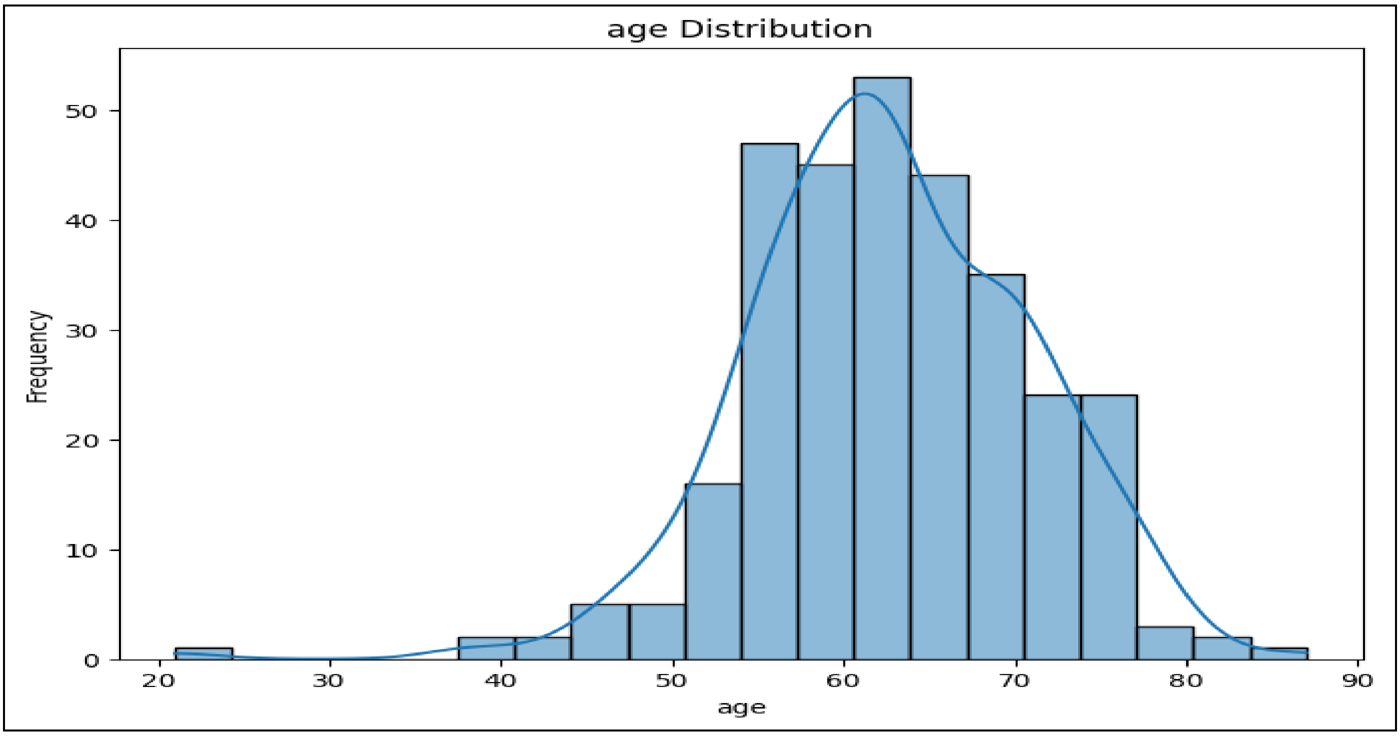

Fig 1 Age Distribution

➢ *Correlation Heatmap:*

Fig 2 illustrates the heatmap represents the Pearson correlation coefficients among all features in the data set, including the target variable lung cancer. A few interesting relationships emerge from this picture.

- Moderate positive correlations with lung_cancer were found for Smoking, Alcohol_consuming, Coughing, Shortness_of_breath, and Chest_pain.
- Smoking is notably associated with Yellow_fingers and Anxiety, which may imply behavioral correlates.
- The generally low to moderate correlation values suggest that no one feature alone can be used for prediction, thus emphasizing the need for a multi-feature-based machine-learning approach.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr1335

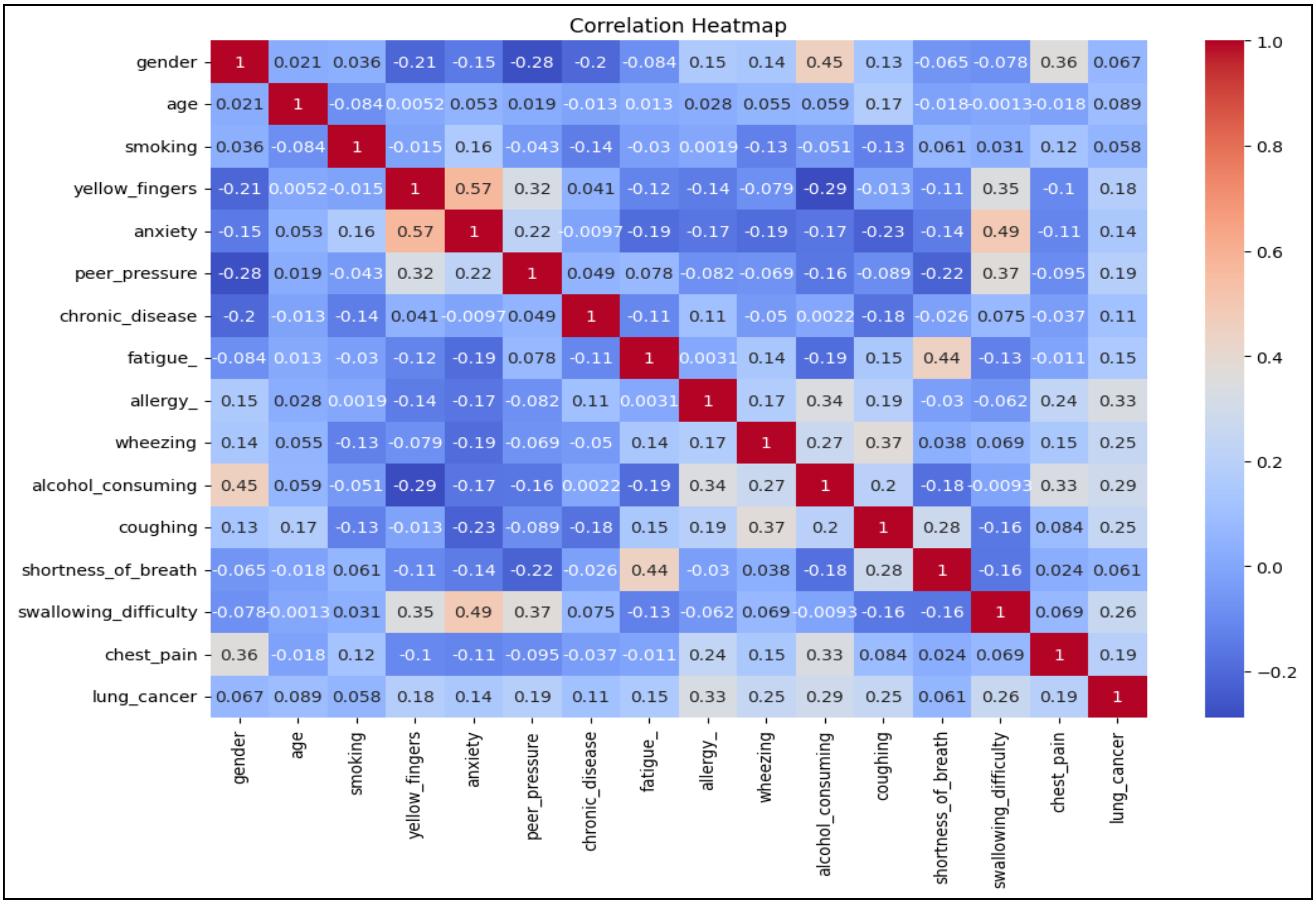

Fig 2 Correlation Heatmap

➢ *Impact of Smoking on Lung Cancer:*

Fig 3 shows the bar that display relationship between smoking and the occurrence of lung cancer. It also shows that among individuals in this study, those who smoke (1, 2 for two-level smoking intensity) have a much higher incidence of lung-cancer (Yes) compared to non-smokers. This is in line with an already established relationship between tobacco use and lung cancer and justifies whether smoking should be included as significant predictor variable in the models.

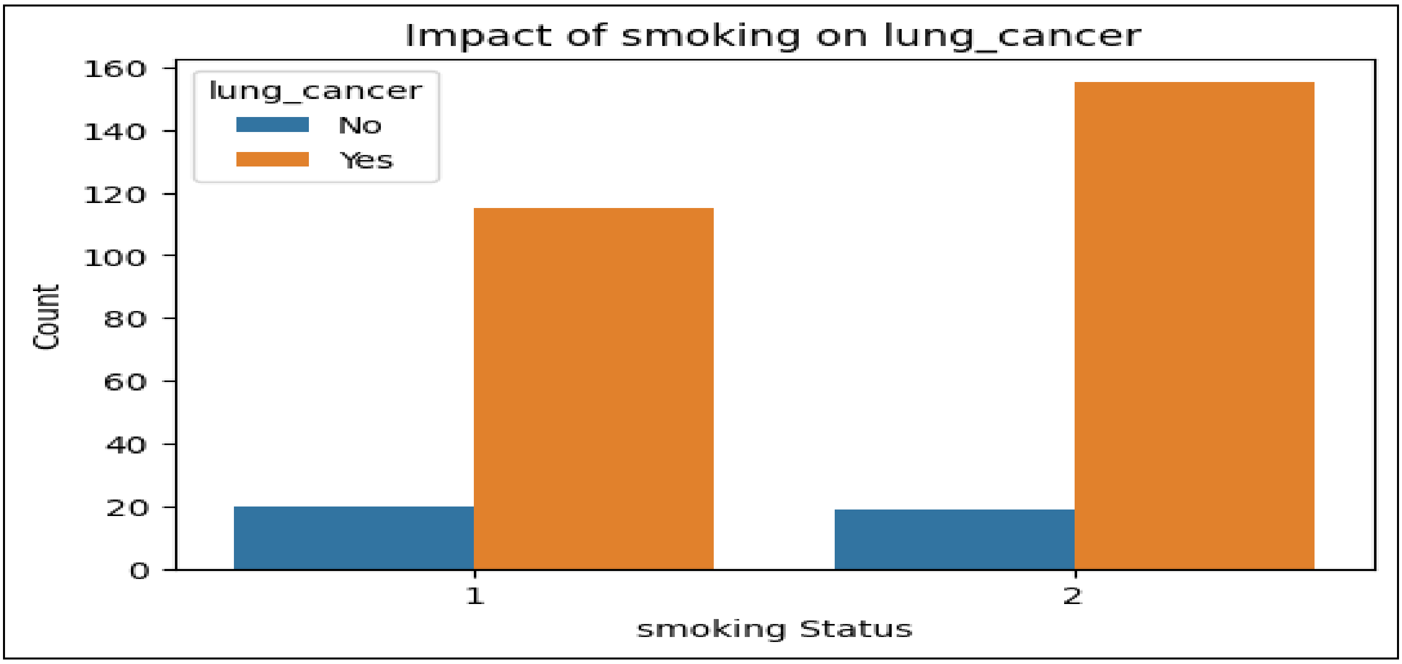

Fig 3 Impact of Smoking on Lung Cancer

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr1335

- *Model Performance Evaluation*

Table 1 shows the performance comparison of different machine learning models for lung cancer prediction reveals that the Random Forest, Logistic Regression, and Support Vector Machine (SVM) models thus raise their heads above the others with 96.7% accuracy, 98.3% precision, 98.3% recall, and 0.983 F1-score, as shown by the performance comparison done in Table I. The results indicate that these models indeed do a great job at identifying positive cases (patients with lung cancer) and incur very few false positives. Such consistency across all these metrics suggests a good prediction capability that is indeed a balance, an implication of the reliable nature of these models in the real world, where medical diagnosis considering accuracy and reliability will not be compromised.

Table 1 Model Comparison

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.952	0.983	0.967	0.975
Random Forest	0.967	0.983	0.983	0.983
Logistic Regression	0.967	0.983	0.983	0.983
Naive Bayes	0.952	0.983	0.967	0.975
Support Vector Machine (SVM)	0.967	0.983	0.983	0.983

On the contrary, the Decision Tree and Naive Bayes models have slightly lower accuracy, that is, 95.2% and 0.975 F1-score respectively. Although performing quite well, the slight drop in recall and F1 scores compared to the leading models is indicative that these models are probably less able to identify all true lung cancer cases. In addition, these models also meet the requirements and could therefore be considered in a situation where computation simplicity or model interpretability is preferred. Overall, these different models seem to have potential to contribute to the early detection of lung cancer; however, in this case, the popularity of the model seems to reside with the ensemble and linear ones.

- *Streamlit Interface:*

The interface of streamlit app has 4 sections: Home, Data Exploration, Model Performance, Prediction.

Fig 4 shows the home page that displays at the top the application title "Lung Cancer Prediction System" which is actually an overview of the application and informs the user that this application employs machine learning using a Random Forest classifier for predicting the likelihood of lung cancer based on risk factors and symptoms. It also has a "How to use this app" portion containing the steps in numbers to navigate to the different sections. Notably, an important notice about the nature of this tool is stated at the bottom to inform users of this educational tool so that they would consult health professionals for proper diagnosis and treatment. The navigation sidebar shows "Home" selected and would allow access to the remaining three sections.

In the Data Exploration section, users are provided insights into the dataset that trained the prediction model. This interactive page shows the raw dataset in combination with statistical summaries to assist in comprehending the data pattern. Some of the visualizations are crafted to represent distributions in charts, histograms, and correlation heatmaps that will allow users to examine relationships between factors including age, gender, smoking status, and lung cancer occurrence. Hence, this section enables both technical and non-technical users to grasp the underlying causes of lung cancer risk without the requirement of any prior knowledge of data analysis.

Through a collection of intensive metrics and visuals, the section on Model Performance undermines revealing exactly how well the predicting device operates. It presents at-a-glance indicators that comply with the structured formats of the accuracy, precision, and recall values of the Random Forest model. Further, users can find and digest the detailed classification report and view how the model performs across different prediction scenarios by visualizing the confusion matrix. The feature importance chart provides a ranked visualization of the most important factors in the model's input to predictions thus giving a clinical feel to the machine learning approach as well as encouraging confidence in the foundations of the system.

Within the Prediction section, a user interface is displayed where the user shall enter his/her health information to know about the risks for lung cancer. It uses a two-column form with yes/no radio buttons for the symptom/risk factor questions, plus an age slider. The provided information can then be submitted through the "Get Prediction" button, after which the user receives clear results showing risk levels in a percentage and in intuitive horizontal gauge visualizations with red and green color coding. It is reminded that, regardless of the result, proper medical disclaimers explain that the prediction is for educational purposes and is not a replacement for professional medical advice.

Fig 5 shows the page that allows users to enter their health information using a combination of radio buttons and slider controls. The design follows a two-column format with questions regarding different risk factors and symptoms. The left column includes questions on gender (male selected), age (slider set to 38), smoking status (yes selected), yellow fingers (no selected), anxiety (yes selected), peer pressure (no selected), chronic disease (yes selected), and fatigue (yes selected). The right column shows questions on allergies (no selected), wheezing (no selected), alcohol consumption (yes selected), cough (yes selected), shortness of breath (yes selected), difficulty swallowing (no selected), and chest pain (no selected). The button which calls for prediction is at the bottom.

In Fig 6 the outcome page appears after the user submitted the information for an analysis. In a bright red alert

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr1335

box, 'High risk of lung cancer detected. Probability: 81.00%' appears. A note following states that this is not a diagnosis and recommends professional medical consultation. The main view is a horizontal bar chart that reads 'Lung Cancer Risk Probability'; a large red portion (81.00%) stands for the probability of cancer risk, and the smaller green section (19.00%) is the probability of no cancer. The x-axis is laid out from 0.0 to 1.0.

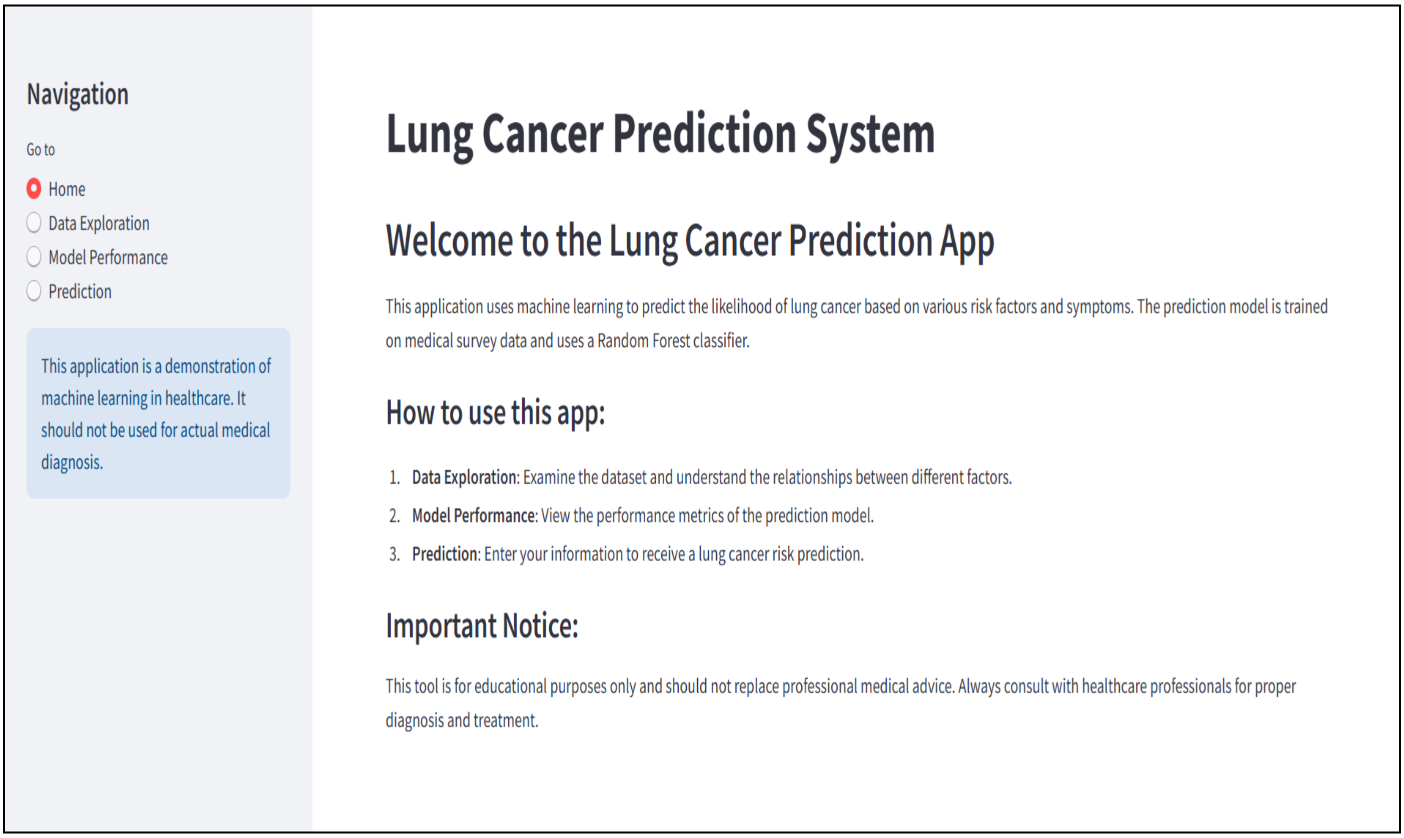

Fig 4 Home Page

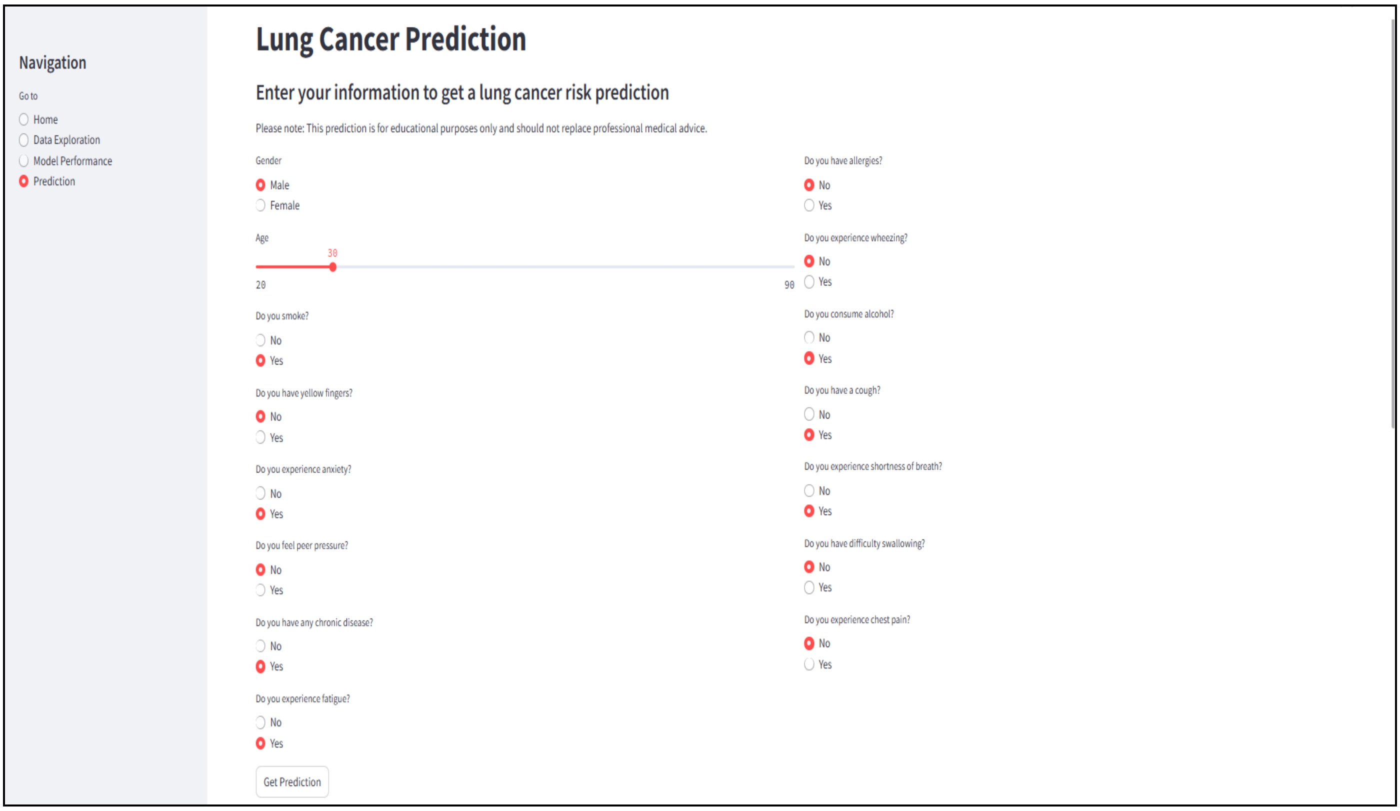

Fig 5 Prediction Input

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr1335

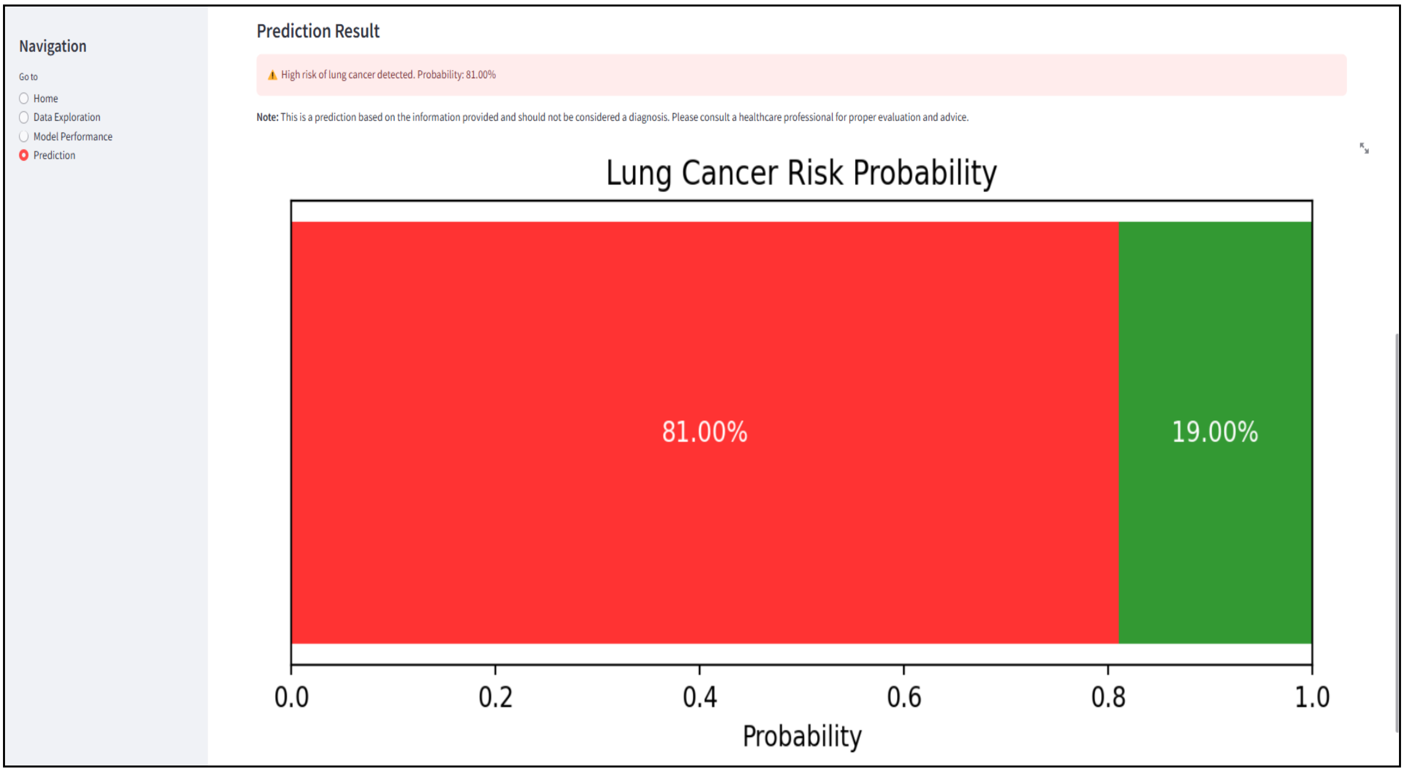

Fig 6 Prediction Result

V. CONCLUSION

Hence, the present study provides an evidence for the efficacy of common machine learning algorithms in case of lung cancer prediction based on data from survey questionnaires including demographic and health-related features. Out of the chosen models - Decision Tree, Random Forest, Logistic Regression, Naive Bayes, and Support Vector Machine (SVM), rest of them including Random Forest, Logistic Regression, and SVM were the most accurate models yielding an F1-score of 0.983. The consistently high performance across these models accentuates the utility of machine learning as a reliable diagnostic aid capable of identifying individuals at a very low probability of being falsely diagnosed as either high risk or low risk. The study thus marks an emphatic reiteration of the need for data driven approaches into health care to strengthen their ability to ease early detection and lead to timely and appropriate intervention, as such needle moving patient outcomes with lung cancer.

The great future holds in store a plethora of opportunities through which this research could be improved. Larger and more heterogeneous datasets-along clinical, genetic, and imaging lines-may thereby be included in future studies for the better generalization of models and capturing of more complex patterns. Therefore, it could also entail the development of real-time prediction systems with integration into management solutions for hospitals or mobile health applications so they could be available to the general public as proactive screening tools. This incorporation of explainable artificial intelligence (XAI) techniques would be geared toward transparency and trust in the decision-making process, allowing clinicians to comprehend why certain model predictions were given. Eventually, these will all play an important role in creating more personalized, effective health solutions, strengthening the current role of AI in modern medicine.

ACKNOWLEDGMENT

In this section, we wish to express our sincere recognition and appreciation to Prof. Ganesh Shelke for his constant counseling, motivation, and support towards this project. His understanding and expertise were of great value in organizing the structure and outcomes of this project work.

Of course, we would also like to thank the Vishwakarma Institute of Information Technology, Pune for the project infrastructure and provisions made available to us enabling the project to be carried out successfully.

REFERENCES

[1]. Narvekar, S., Shirodkar, M., Raut, T., Vainganker, P., Kumar, K. M. C., & Aswale, S. (2022). A survey on detection of lung cancer using different image processing techniques. *2022 3rd International Conference on Intelligent Engineering and Management (ICIEM),* 13–18. https://doi.org/10.1109/iciem54221.2022.9853190

[2]. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73. Nadkarni, N. S., & Borkar, S. (2019). Detection of Lung Cancer in CT Images using Image Processing. In Goa College of Engineering, *Proceedings of the Third International Conference on*

Trends in Electronics and Informatics (ICOEI 2019) (p. 863).

[3]. Shafiq, S., Asghar, M. A., Amjad, M. E., & Ibrahim, J. (2022). An Effective Early Stage Detection of Lung Cancer Using Fuzzy Local Information cMean and GoogLeNet. *New*, 1–6.

[4]. Vikruthi, S., Thippagudisa, K. B., Basha, P. H., Vamsi, P., Chandra, G. M., & Kiran, V. S. (2023). Performance Analysis of Gene Expression Profiles of Lung Cancer Prediction using JR Algorithm. *New*, 1731–1736.

[5]. B, K. S. (2020). Prediction of lung cancer using Convolutional Neural Network (CNN). *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(3), 3361–3365.

[6]. Bhargav, A. L., & Ashokkumar, C. (2024). AI-Driven Insights: A Survey on Innovative Approach for Lung Cancer Prediction Utilizing Machine Learning and Deep Learning Methods : Lung Cancer Prediction Utilizing Machine Learning and Deep Learning Methods. *New*, 1152–1158.

[7]. Baker, Q. B., Gharaibeh, M., & Al-Harahsheh, Y. (2021). Predicting Lung Cancer Survival Time Using Deep Learning Techniques. *Ew*, 177–181.

[8]. K, A., T, R., B, S., & M, K. (2024). Real-Time lung cancer detection using GAN-Enhanced LDCT imaging and CNN-Based classification for clinical accuracy. *2021 International Conference on System, Computation, Automation and Networking (ICSCAN)*, 1–6.

[9]. S. H. Begum, M. I. Baig, M. A. Hussain and M. A. Muqeet, "A Lightweight Deep Learning Model for Automatic Diagnosis of Lung Cancer," *2022 IEEE 2nd International Conference on Mobile Networks and Wireless Communications (ICMNWC)*, Tumkur, Karnataka, India, 2022

[10]. A. Jawaid, J. Hafeez, S. Khan and A. Ur Rehman, "Lung Cancer Detection using Artificial Neural Network on Android," *2023 Global Conference on Wireless and Optical Technologies (GCWOT)*, Malaga, Spain, 2023, pp. 1-8, doi: 10.1109/GCWOT57803.2023.10064658.

[11]. A. A. Sagaya Priscilla and R. Balamanigandan, "Hybrid Artificial Neural Networks for Lung Cancer Detection," *2024 International Conference on Sustainable Communication Networks and Application (ICSCNA)*, Theni, India, 2024, pp. 757-761, doi: 10.1109/ICSCNA63714.2024.10864149.