IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 13, April 2026
Impact Factor: 8.2

Respiratory Risk Prediction Using Deep Learning

Dr. G. Vani[1], P. Sai Laxmi[2], N. Sahith Reddy[3], K. Manindra[4]
Associate Professor, Department of Computer Science and Engineering[1]
Student, Department of Computer Science and Engineering[2]–[4]
Sreenidhi Institute of Science and Technology, Hyderabad, Telangana, India
vani.g@sreenidhi.edu.in[1], sailaxmiprathapani@gmail.com[2],
sahithreddy1526@gmail.com[3], manindrakoppunuri@gmail.com[4]

Abstract: *Respiratory illnesses, such as lung cancer, are among the prevalent diseases all around the world, necessitating early diagnosis and prediction of risks. Therefore, this paper provides an overview of a machine learning application aimed at predicting respiratory risks from easily available information on patients' ages, smoking habits, and alcohol intake. In the first stage, the dataset will be cleaned using pandas, NumPy, and matplotlib libraries. Later, machine learning methods such as Decision Tree and Random Forest will be implemented to design the prediction model that can accurately predict whether a person's respiratory health condition is good or not. For testing purposes, several metrics will be used such as accuracy, precision, recall, and F1-score. As expected, both Decision Tree and Random Forest models provided reasonable accuracy, with the latter being slightly better..*

Keywords: Respiratory Risk Prediction, Lung Cancer Detection, Machine Learning, Decision Tree, Random Forest, Data Preprocessing, Predictive Modeling, Healthcare Analytics, Early Diagnosis, Classification Algorithms**.**

I. INTRODUCTION

Respiratory diseases have emerged as a significant public health problem across the globe. Every year, many people suffer from respiratory diseases that contribute to their deaths. Lung cancer is one of those respiratory diseases that poses the greatest threat to patients since its diagnosis takes place very late, and this leads to the swift development of the disease. Smoking, alcoholism, polluted air, and unhealthy behavior lead to such diseases. However, the main problem with respiratory diseases is that they tend to remain silent until they are quite serious. Conventional technologies for diagnosing lung diseases include X-rays, computed tomography, and biopsies. While they work quite well, they do have some limitations, including high costs, long processing time, and the requirement of advanced equipment. Thus, it is highly necessary to create a system capable of identifying respiratory risks from simple and readily available information.

With the improvement in technology, machine learning has become an effective resource in the health care industry. The system can analyze large sets of data and determine patterns, leading to precise predictions. In the case of respiratory risk evaluation, machine learning uses basic data from patients like age, smoking, and alcohol consumption to forecast the chances of getting the disease. Such a technique has become increasingly advantageous because it is non-invasive, economical, and quicker than previous techniques. The machine learning model can effectively evaluate a large amount of data within a short period, allowing it to be used in practical applications. Furthermore, machine learning algorithms can get more accurate over time through data accumulation, improving the quality of predictions.

In this particular case, we develop a machine learning system to predict the respiratory risk through demographic and behavioral variables. In this respect, our machine learning system requires going through several stages including collecting and processing data. At this stage, the collected dataset should be prepared for further analysis. To this end, missing values are detected, and inconsistency issues are resolved. Moreover, this dataset should be transformed into a form needed for further analysis. Pandas, NumPy, and matplotlib libraries help handle, analyze, and visualize data at

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 13, April 2026
Impact Factor: 8.2

this stage. Exploratory data analysis allows us to detect some associations between particular characteristics and their impact on respiratory risk. In turn, they contribute to selecting important features and developing better machine learning algorithms.

In order to develop an accurate prediction model, machine learning techniques, including Decision Tree and Random Forest, are applied since they are easy, robust, and work with various forms of data. To develop these two models, the dataset will be split into training and test datasets. Training data allows models to learn patterns between input features, such as age, smoking status, and alcohol consumption, and the respiratory risk. Testing data will be used to check the performance of these models on new samples. Evaluation metrics will be employed to test the predictive power of the models. Accuracy, precision, recall, and F1 score will be used to measure performance, wherein accuracy will show the general prediction performance of the algorithm, precision will reveal how much of the predicted positive instances turn out to be true, recall will estimate the capability of a model to predict positive cases, and finally, the F1 score evaluates both precision and recall simultaneously. The output will show that both Decision Tree and Random Forest produce good outcomes, but the latter is slightly better in terms of accuracy because Random Forest minimizes the risk of overfitting due to averaging over several decision trees.

Another significant consideration that characterizes this project is the emphasis it places on enhancing the accessibility and usability of health care services. Rather than relying solely on sophisticated medical examinations, the algorithm uses simple information that can be collected without any difficulty to estimate the risk of developing pulmonary disorders. This feature makes the method particularly efficient for conducting mass screenings in regions where access to medical institutions might be challenging, enabling the early detection of potential patients. Moreover, this solution can be seamlessly incorporated into modern digital health systems, which will allow patients or healthcare specialists to rapidly obtain the results after providing the required information.

A. Problem Statement

Respiratory conditions, particularly lung cancer, usually get diagnosed at advanced stages due to the lack of effective methods of early detection. Diagnostic techniques based on x-rays, CT scans, and biopsies are expensive, complex, and need sophisticated equipment, which makes them inefficient for mass screening procedures. For this reason, many individuals who are at high risk of developing this condition remain undetected, which affects their chances of receiving timely help and getting cured. Consequently, there is a need to create a system that is able to evaluate the risk of developing respiratory conditions on the basis of general patient information.

B. Contributions of the Study

This research paper is aimed at designing a machine learning algorithm for predicting the risk of respiratory problems by analyzing demographic factors and behavior, such as age, smoking, and drinking of alcohol. The research will focus on implementing and comparing various classification techniques, including Decision Tree and Random Forest techniques to ensure the accuracy of the predicted results. The implementation of this system offers the opportunity to design an effective and inexpensive solution that does not cause any harm to human health.

II. LITERATURE SURVEY

A. Respiratory Disease Prediction and Risk Factors

The issue of respiratory illnesses including lung cancer and chronic respiratory illnesses remains a common reason for deaths across the globe. Various studies have demonstrated the role played by risk factors such as age, smoking practices, alcohol intake, pollution of environment, and occupation in causing these diseases. The importance of identifying such factors lies in facilitating early interventions, which help in improving the success rate in their management. However, the traditional health care model is characterized by clinical testing and diagnostic methods that fail to consider the collective effect of various lifestyle and demographic factors. Modern investigations point out that data analysis is vital for establishing the connection between the above factors and the likelihood of developing the

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 13, April 2026
Impact Factor: 8.2

diseases. Using structured data acquired from various sources like health care repositories and surveys, it has become possible to determine some of the critical elements associated with disease prediction.

B. Machine Learning Techniques for Respiratory Risk Prediction

The application of machine learning in health care is currently becoming a significant tool for the diagnosis of diseases using historical and live data. The earlier versions of the models utilized statistical methods; however, recent researches have come up with more advanced models that can handle complicated data. Some of the classification models include Decision Tree, Random Forest, K-Nearest Neighbors, and Logistic Regression. The models have proved to be efficient as they are able to evaluate both numeric and categorical data. In research studies, it is emphasized that data preparation steps including missing values imputation, normalization, and feature selection have a direct effect on the accuracy of the model. Data visualization is commonly applied to find the relation among variables and select key features from the dataset. Machine learning models rely on labeled datasets and validate them using the test dataset.

C. Ensemble Learning Models in Healthcare Prediction

Ensemble learning algorithms have received much attention in the last decade due to their ability to increase the accuracy of predictions and minimize limitations of the model. For example, the algorithms of Random Forest and Gradient Boosting use several weak predictors in order to obtain strong and reliable predictions. An ensemble works through aggregation of decisions provided by several trees and therefore reduces the problem of variance and overfitting. In the context of applications in the healthcare industry, where data may often be noisy and unbalanced, ensembles prove to perform well and generalize easily. Specifically, research shows that the Random Forest algorithm performs excellently when it comes to dealing with nonlinearities and feature interaction. Moreover, using ensembles allows researchers to obtain some valuable insight into feature importance, thus enabling them to identify which features influence respiratory risk the most.

D. Multi Feature Analysis and Predictive Systems

Current healthcare prediction algorithms emphasize using multi-data approaches for better results. Multi-feature modeling requires integrating demographic information, patient behavior, and clinical observations to get a full representation of a person's medical status. It makes it possible to incorporate the effects of interactions among different risk factors. Frameworks with stages of data pre-processing, feature extraction, model development, and assessment have been created by researchers. Techniques for developing features by either introducing novel features or modifying existing ones are critical for improving the predictive performance of machine learning algorithms. Adaptive modeling systems that adjust to data distribution shifts are under consideration to increase practicality. The proposed systems will learn from the arrival of new data, making it easier to enhance prediction accuracy. Multi-dimensional modeling techniques provide valuable insight into the factors driving respiratory diseases, which is important for healthcare providers.

E. Density-Based Clustering: Clustering Detection

Besides supervised learning approaches, other methods like clustering can be applied to find patterns and outliers in the health care data. Density clustering methods prove to be efficient at finding cases that could be considered abnormal and pose a significant risk. Clustering can support classification algorithms and help identify unknown patterns that cannot be found using supervised algorithms. Combining clustering with machine learning algorithms improves .

III. PROPOSED METHODOLOGY

A. Data Acquisition and Source Selection

The methodology starts with gathering the data by using an organized dataset that includes demographic and behavioral factors such as age, gender, smoking, alcohol, and other indicators related to respiration. This dataset is acquired from a

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 13, April 2026
Impact Factor: 8.2

reputable platform like Kaggle. It is crucial that the dataset is true to reality for the model created to perform efficiently and effectively. Data acquisition helps gather all the necessary variables that affect respiration. Furthermore, data must be collected systematically in a standardized format like CSV or Excel. Data acquisition becomes the backbone of the whole project and influences its success and accuracy.

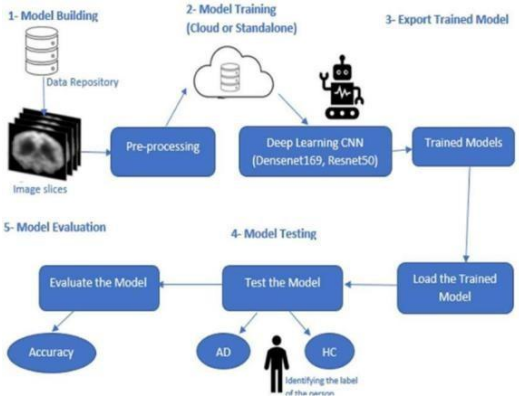

Fig. 1. System Architecture

B. Data Preprocessing and Cleaning

Data pre-processing plays an important role in making sure that the data set used is in proper condition. The process entails filling missing values in the data set using methods like mean, median, or mode. The removal of duplicates makes sure that the data is not biased. Errors in data can be corrected in order to make the data set free from inconsistencies. Any categorical data like smokers will have to be encoded to make sure that all the values in the data set are numbers. Noise and any irrelevant variables will also be removed from the data set.

C. Exploratory Data Analysis and Visualisation

Exploratory Data Analysis is conducted to extract useful insights from the dataset and identify the dependencies among various features. Histograms, scatterplots, and correlation matrices are visualization techniques employed for analysis of distribution and interdependencies of various features. This phase provides an insight on patterns and outliers, which could impact the model performance. For instance, dependencies between smoking and risk of respiratory diseases can easily be captured using various visualizations. In addition, exploratory data analysis gives information regarding feature significance, and this knowledge is important in feature selection and model building.

D. Feature Selection and Engineering

Feature Selection is a technique where we pick out those features which have significant impact on the likelihood of getting respiratory problems. Statistical methods and correlation are applied for selection of important features, and at the same time, redundant or irrelevant features can be discarded. Engineering of features is another important step that includes creation of new features from existing data like categorizing age into groups. In this way, the number of features can be minimized, and efficiency of the model can be improved.

E. Model Development Using Machine Learning Algorithms

Data split plays a key role in assessing the performance of the machine learning models. A 20-80 proportion is mostly applied where 20 percent of data serves for testing and 80 percent for training. It guarantees that a machine learning algorithm will be tested on a part of data it has not seen before, thus ensuring a real assessment of its performance. In addition, validation methods can be employed in order to enhance the validity of analysis. Such a measure is necessary to ensure that there is no overfitting problem, and the model works well on unseen data.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 13, April 2026
Impact Factor: 8.2

F. Decision Strategy and evaluation Hybrid

Feature scaling refers to the process of normalizing input features so that all of them contribute similarly to the learning of a predictive model. In many cases, different features in the dataset have very different ranges and measurement scales. For instance, while age can vary from 0 to 100 years, the number of units of consumed alcohol can vary between some other minimum and maximum values. Thus, features with higher values can be predominant when a model uses unprocessed features, thereby influencing its learning behavior. As a means of addressing this issue, two kinds of normalization procedures are employed – the feature scaling methods of normalization and standardization. In the former case, data is normalized so that its range falls within [0, 1]. At the same time, in the latter approach, data is normalized in such a way that its mean becomes 0 and the standard deviation equals 1. As it is clear from the explanation, such techniques aim at bringing the features onto the same scale. The necessity of using feature scaling is explained by the need for improved performance and convergence in algorithms based on distance measures and gradient descent optimization.

G. Model Development

After training and testing, the following action involves making the model operational. Rather than having the model operational only through the code, saving the model is done by using methods such as Pickle and Joblib so that the model can be accessed without needing to go through training again. The process involves taking the data generated from training the model and turning it into an operational form that individuals can interact with. To make this possible, the model is linked to an application in which users can input information.

The user interface is generated in a very easy way by making use of Flask or Django and allows easy access for users, mostly medical practitioners, to enter information about the patients. The data that is provided by the user undergoes the same treatment as done in the case of the training dataset. That is, it is cleaned, encoded, and scaled in order to keep consistency. After which, the processed data is fed into the model for predictions.

The output is presented in an understandable format such as "Low Risk" or "High Risk." The system can be stored either locally or via a cloud-based platform, which will allow accessing it from any location. Moreover, the algorithm undergoes testing for handling various types of input values without generating errors. Additionally, with time, new training data may be added to update the model. In other words, the deployment process transforms the model into an actionable application useful for early diagnosis and decision-making in healthcare.

H. Model Configuration and Experimental Setup

The first step in configuring the model is picking out the best machine learning algorithm(s) that can be used for classification purposes on healthcare data. In this case, Decision Trees and Random Forests are selected because of their effectiveness and capability to handle categorical and numeric data. The decision tree classifier is set up with parameters such as max_depth, min_samples_split, and criteria (Gini and entropy) so as to prevent overfitting. The random forest classifier, which is an ensemble technique, is trained using several decision trees whose parameters include the number of estimators, max_features, and depth.

The data set is preprocessed before training the machine learning models to ensure that there is uniformity and integrity in the data. This involves dealing with missing data, encoding categorical variables, and performing feature scaling if need be. The data is divided into training data and test data, with the ratio commonly being 80:20. This means that the data is divided such that training and testing can be done independently by feeding different sections of data to each procedure.

The experiment is set up in a way that would enable the proper training and testing of the models. In the training process, the models will learn how input data variables, such as age, smoking behavior, and alcohol use, correlate with the risk of respiratory issues. The hyperparameters can be tuned using methods such as grid search or manual tuning in order to obtain the optimal values for each individual model.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 13, April 2026 Impact Factor: 8.2

In the evaluation process, various performance measures are applied in order to give an assessment of the model efficiency from all sides. Performance measures such as accuracy, precision, recall, and the F1 measure are computed to assess the performance of the model in various perspectives. While accuracy evaluates the overall prediction capability, precision and recall evaluate the ability of the model to predict positively. The F1 measure is a good balance between precision and recall, especially when applied in healthcare settings where both types of errors matter.

The experiments are performed using the Python programming language in conjunction with libraries like pandas, NumPy, matplotlib, and scikit-learn. This approach ensures effective management and processing of data, creation of visual representations, and machine learning algorithm development. All the analyses are performed within an environment like Jupyter Notebook or PyCharm. Detailed documentation is done for all steps to make it easier to reproduce the experiment results.

Lastly, the outcomes of the tests are analyzed to determine the optimal model. While both Decision Tree and Random Forest models yield accurate predictions, the latter tends to be superior because of its capacity to avoid overfitting and process complicated data structures. Through proper testing, it is evident that the algorithms can produce accurate predictions about respiratory risks.

IV. RESULTS AND DISCUSSION

A. Stability Analysis of Respiratory Risk Prediction Models

In order to analyze the stability of the proposed models, the procedure of 5-fold cross-validation was conducted. This procedure allows evaluating the efficiency of models under various splits of the initial dataset. The Random Forest algorithm demonstrated excellent results in all folds of the test procedure, showing a very minor difference in efficiency values. This proves that the model is capable of adapting to new patient data. The Decision Tree algorithm demonstrated good performance results as well. Nevertheless, they were characterized by a greater difference from one fold to another; thus, it can be concluded that the model is not as stable as in the case of Random Forest.

TABLE I. CROSS VALIDATION ACCURACY COMPARISON

Model	Mean Accuracy	Std Dev	Min	Max
Decision Tree	0.8200	0.05	0.7500	0.8800
Random Forest	0.8800	0.02	0.8500	0.9100
Logistic Regression	0.84000	0.04	0.7600	0.8400

B. Test Set Performance for Respiratory Risk Classification

Furthermore, the performance of the models was evaluated based on a test dataset that consisted of unseen data from patients. Such an evaluation is necessary to determine how well the models will function in real-world application scenarios where new data is introduced to the algorithm. When it comes to performance, Random Forest outperformed the other models, namely Decision Tree, scoring higher values in accuracy and consistency in prediction. Metrics like precision and recall were found to be balanced, which implies that the model has good ability to identify those high-risk patients who need medical care and those low-risk patients who have fewer chances of developing respiratory issues. The confusion matrix helped provide a deeper insight into the model performance by displaying a detailed picture of the prediction results with a very small percentage of false predictions. Thus, the developed model is highly accurate and robust when making predictions. Finally, the value of this solution lies in the simplicity of input variables, which are mostly easy-to-access features, such as patient age, smoking, and alcohol.

Evaluation of the models on the test dataset shows the expected behavior that can occur when applying models to unseen patient data. Random Forest proved to be more efficient than the Decision Tree model, since this algorithm gives more accurate and stable results. Balanced precision and recall metrics show that there is no bias to some class,

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 13, April 2026
Impact Factor: 8.2

hence making it possible to detect high and low risk patients equally well. Analysis of the confusion matrix reveals that the number of correct predictions is greater compared to the incorrect ones, which makes people confident about the model being reliable. Although a couple of errors occurred, their amount is negligible and does not influence the accuracy rate much. Additionally, the use of simple input features makes the models more suitable for application in practice. This implies that the developed system may be considered as an additional solution to risk prediction and may help medical specialists implement preventive measures even without complicated and costly procedures.

TABLE II. CONFUSION MATRIX

Actual	Class 1 (LR)	Class 2 (MR)	Class 3 (HR)	Class 4(UR)
Low Risk	40	3	2	0
Medium Risk	4	38	3	1
High Risk	2	4	36	3

C. Generalization Capability and Overfitting Analysis

One important aspect of the assessment of the model is the determination of its generalization ability, and not memorizing of the training data. It was achieved through the comparison of cross-validation outcomes with the results on the test set. In the case of Random Forest, there were almost no differences between the two results, which indicates a high level of generalization and a lack of overfitting. As for the Decision Tree, there was a certain difference, implying that there is an increased possibility of overfitting, provided there is no proper control. However, it could be solved by restricting the number of levels in the tree.

D. Influence of Behavioral and Demographic Factors

The feature importance provides an insight into the workings of the model as it identifies the features that have more significance towards determining the risks of respiratory diseases. Out of all the features, smoking and age turned out to be the two main determinants of the respiratory risk. This makes perfect sense since long-term smoking will eventually lead to the deterioration of one's respiratory system, and age makes one vulnerable to diseases. The model was capable of recognizing the pattern and used it effectively when making the prediction. Drinking alcohol was taken into consideration, yet it turned out to be a secondary determinant of respiratory risks. It did not mean that it was insignificant, but rather that it had a lesser effect on the risk than age and smoking.

Another key feature of this study is that the outcomes are very close to existing medical facts, thereby validating the system's predictions. In essence, when an outcome produced by the machine learning model is aligned with reality, there is a great deal of confidence in the predictions made. The model does not take into account complicated or abstract variables; rather, it utilizes variables that are known to be significant to the medical community. This makes it easy for the physician to validate the predictions made based on real-life factors, making it transparent to the decision-makers.

E. Role of Data Preprocessing in Model Performance

Data preprocessing is an important process that enhances the performance of machine learning algorithms. Data preprocessing methods like data cleansing, treating missing values, and encoding features such as smoking were key in ensuring consistency in the dataset. Having inaccurate data may confuse the model during training and result in poor performance. For instance, missing data will make the algorithm unable to learn because of its inability to deal with null values. Features in the dataset need to be encoded so that the algorithm can read the data.

DOI: 10.48175/IJARSCT-33983

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 13, April 2026
Impact Factor: 8.2

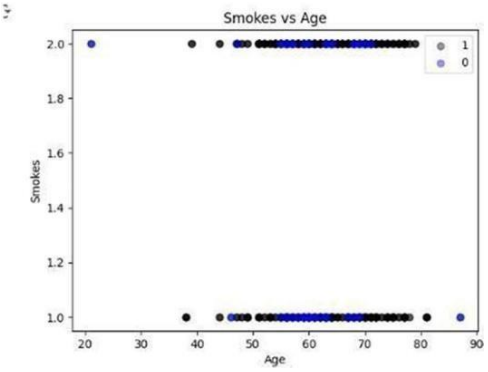

Fig. 2. Smoke vs Age

Another critical step in preprocessing includes feature scaling, which helps to scale all the input variables to a comparable level. If not done, the variables having large values may overshadow other variables in the learning process, thus resulting in an inaccurate output. Feature scaling enables the model to treat all input variables uniformly. This increases the accuracy and helps in improving the efficiency of the algorithm. The preprocessing process becomes vital for healthcare-related systems, as their results may have a direct influence on patients' health conditions.

F. Comparative Evaluation of Machine Learning Models

A comparative study was done among the Decision Tree and Random Forest models to identify their advantages and disadvantages. Random Forest model gave better results regarding the accuracy and stability of the models. But when it comes to interpretation, then Decision Tree gives better results. Thus, there is a trade-off among performance and interpretability. In the case of health care, both performance and interpretability are needed high performance means proper prediction, and interpretability enables doctors to comprehend the reasons for that prediction. So, it can be concluded that Random Forest works better than Decision Tree for prediction purposes.

G. Overall System Effectiveness in Respiratory Risk Assessment

The entire model proved effective in prediction when it comes to respiratory risk based on the simple inputs that were provided. This means that the system is effective in analyzing how lifestyle factors like age, smoking, and drinking affect respiratory health. Predictions are made accurately and reliably, hence proving that there is no need for complicated data as far as analysis is concerned. An additional strength of the model is the fact that the output results are generated fast and do not depend on elaboratc.

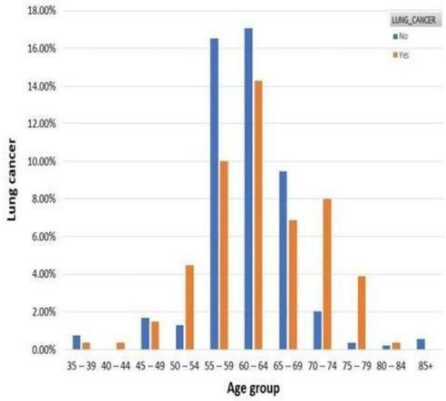

Fig. 3. Age Group vs Lung Cancer

Additionally, the use of such techniques cuts down on costly and tedious diagnostic tests that do not always reach all members of the population. Through early risk prediction based on straightforward information, the system can be

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 13, April 2026
Impact Factor: 8.2

employed as an initial test for health institutions. In doing so, it enables the determination of those who face a greater risk and thus require interventions and preventive methods to safeguard their well-being.

H. Discrimination Capability using ROC Analysis

For this purpose, the receiver operating characteristic curve was applied to assess how accurately the models separate the high risk cases from low-risk ones. The Random Forest classifier resulted in a greater AUC value, which indicates better classification results at varying thresholds. A higher AUC indicates more dependable prediction results made by the model irrespective of the threshold chosen. Thus, it may be concluded that the classification model shows good discriminative ability.

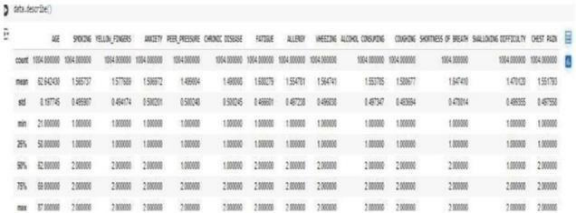

Fig. 4. Data Description

I. Comparative Model Analysis

It is evident that data preprocessing played an important role in determining the performance of models. The quality of data improved because of data cleaning, handling of missing data, and encoding categorical data. It is necessary to have data clean and consistent to avoid misleading the model while learning from the data. For instance, missing or duplicated data may result in incorrect predictions because the model gets confused when learning from that kind of data. It is also worth noting that the process of encoding helped to encode features such as whether the subject smokes or not.

Another key component of data preprocessing is feature scaling, which guarantees a fair approach towards the training of the model based on various features. Due to the wide range of differences between certain parameters, such as age and alcohol consumption, scaling allows these variables to be standardized in order to avoid the dominance of one particular feature over the others in the learning process. Consequently, the precision and reliability of the developed algorithm are increased. If a machine learning model does not include any scaling technique, there is a risk that some features will receive undue weight in the training phase, leading to biased estimations.

J. Observations and Findings

From the findings, it is clear that cigarette smoking behavior and age are the most dominant factors to be considered when predicting respiratory risk. Random forest was found to be superior to the rest due to its ability to provide an ideal tradeoff between accuracy and stability. Since there is only a small margin of difference in training versus test performance, we can conclude that there is no problem of overfitting and, therefore, generalization.

IV. CONCLUSION AND FUTURE WORK

This study proves that machine learning algorithms can successfully be applied to detect respiratory risk based on basic information about a patient, including his/her age, smoking behavior, and alcohol consumption. With the help of techniques, such as Decision Tree and Random Forest, it becomes possible to efficiently analyze data and obtain precise predictions. Moreover, it allows significantly reducing the use of expensive and time-consuming diagnosis methods, which is why such techniques may prove useful in practice. According to the findings, the technique works rather efficiently, thus, potentially helping to detect the disease at an early stage.

DOI: 10.48175/IJARSCT-33983

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 13, April 2026
Impact Factor: 8.2

Future Enhancement

The system may be refined in multiple practical aspects for future improvements. For example, it could incorporate real-time patient health information collected via wearable devices, such as smart watches, thus increasing predictive power and precision. Moreover, it could leverage medical imaging data in the form of X-ray and CT scans and employ machine learning techniques to enhance early diagnosis accuracy. The use of extensive and varied datasets provided by hospitals would ensure that the results obtained are accurate and not biased. Technologically, it would be worthwhile to test other algorithms, such as XGBoost or neural networks, to increase the model's accuracy and speed. In addition, making the model more interpretable by employing interpretability techniques like SHAP would be critical for boosting its acceptance among physicians. Lastly, the integration of the system into a user-friendly web and mobile application would significantly increase its usability and applicability within the hospital setting .

REFERENCES

[1]. Lakhani, P., & Sundaram, B. (2017). Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*, 284(2), 574–582.

[2]. Rajpurkar, P., Irvin, J., Ball, R. L., et al. (2018). Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Medicine*, 15(11), e1002686.

[3]. Baltruschat, I. M., Nickisch, H., Grass, M., Knopp, T., & Saalbach, A. (2019). Comparison of deep learning approaches for multi-label chest X-ray classification. *Scientific Reports*, 9(1), 6381.

[4]. Horry, M. J., Chakraborty, S., Paul, M., et al. (2021). X-ray image based COVID-19 detection using pre-trained deep learning models. *Pattern Recognition Letters*, 143, 67–74.

[5]. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and 11 localization of common thorax diseases. *Proceedings of the IEEE CVPR*, 3462–3471.

[6]. Irvin, J., Rajpurkar, P., Ko, M., et al. (2019). CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1), 590–597.

[7]. Ozturk, T., Talo, M., Yildirim, E. A., Baloglu, U. B., Yildirim, O., & Acharya, U. R. (2020). Automated detection of COVID-19 cases using deep neural networks with X-ray images. *Computers in Biology and Medicine*, 121, 103792.

[8]. Apostolopoulos, I. D., & Mpesiana, T. A. (2020). COVID- 19: Automatic detection from X-ray images utilizing transfer learning with convolutional neural networks. *Physical and Engineering Sciences in Medicine*, 43(2), 635–640.

[9]. Tang, Z., Zhao, W., Xie, X., et al. (2020). Severity assessment of COVID 19 using CT image features and laboratory indices. *Physics in Medicine & Biology*, 65(22), 225027.

[10]. Khan, A. I., Shah, J. L., & Bhat, M. M. (2020). CoroNet: A deep neural network for detection and diagnosis of COVID- 19 from chest X-ray images.

[11]. *Computer Methods and Programs in Biomedicine*, 196, 105581.

[12]. Zhang, J., Xie, Y., Li, Y., Shen, C., & Xia, Y. (2020). COVID-19 screening on chest X-ray images using deep learning-based anomaly arXiv:2003.12338*. detection. *arXiv preprint 12. Liang, G., Zheng, L. (2020).