PREDICTIVE MODELING OF LUNG CANCER RISK USING HYBRID ENSEMBLE MACHINE LEARNING TECHNIQUES

[1]MS.J.Lethisia Nithiya,[2]CH.Anil,[3]B.Srinivas Goud,[4]B.Raja Ravindra Reddy,[5]B.Nagendra Babu

ASST.Professor(CSE), UG Scholar, UG Scholar, UG Scholar, UG Scholar
Department of Computer Science & Engineering
Bharath Institute of Science and Technology, BIHER
173,Agaram Road , Selaiyur, Tambaram, Chennai, Tamil Nadu, India

Abstract : Lung cancer remains among the most fatal illnesses in the global community that is why the early diagnosis and the development of specific risk prediction techniques is so necessary. This work reports a predictive model to determine the degree of risk associated with lung cancerbasedonStackingEnsembleLearningstrategybased on the collection of four device basic cognitive process classifiers to enhance the performance of prediction. Demographic, lifestyle, and clinical variables of the dataset are age, gender, air pollution exposure, smoking habits, genetic risk, chronic lung disease, and such symptoms as fatigue, wheezing, and coughing of blood. Data preproces practices such as missing values,feature grading and coding of categorical variables were used to make sure that there was data consistency and quality. The suggested stacking model integrates the abilities of four foundation learners Support Vector Machine, Random Forest , K-Nearest Neighbors andLogistic Regression withameta-classifier in order toimprovetheoverall predictive power. This ensemble approach uses the heterogeneity of the individual models to both find linear and non-linear relationships in the data. Accuracy, precision, recall, and F1-score were used in determining the performance of the model. The enquiry findings indicate that the stacking classifier has great public presentation over individual classifiers in the ability to predict the levels of risk of lung malignant tumor. Analysis on feature importance indicated that the practice of smoking, genetic risk factors, and respiratory symptoms like wheezing are some of the most contributing predictors. The model developed can be described as categorizing the patients as Low, Medium, and High-risk, which is a strong and dependable tool of early detection and timely provision of medical care. This is an ensemble-based model which provides better generalization and accuracy, which is very appropriate in clinical decision support systems in the real world.

IndexTerms – Lung Cancer Risk Prediction, Stacking Ensemble Learning, Predictive Modeling, Early Detection, Classification Algorithms, Feature Engineering, Clinical Decision Support Systems

I. INTRODUCTION

Respiratory organ cancer is one of the major causes of deaths related it to cancer in the world and it contributes to high percentage of deaths in the world. Early diagnosis is essential in enhancing survival; nonetheless, the disease has very mild signs and non-specific risk factors, which make traditional diagnosis techniques ineffective in detecting it at its early stages. This has given rise to the growing interest in using advanced methods of computation to accelerate earlyprediction and diagnosis Over the last several years, the fast development of healthcare data and the achievements in artificial intelligence has opened the way to the implementation of device learning methods in the field of medical diagnostics. More specifically, predictive modeling has become a potent instrument in the process of determining patterns and relationships in large data sets, allowing estimating the probability of an individual developing lung cancer. These models are capable of including diverse factors that include demographic data, smoking, and exposure to the environment, genetic predisposition and clinical history. Ensemble techniques are machine learning methods that have shown high performance over individual machine learning models because they integrate the strengths of several algorithms. This is also improved by hybrid ensemble approaches that combine different models or determination trees, support vectors machines, and system networks to increase prediction accuracy, robustness, andgeneralization. This paper dwells on the creation of a hybrid ensemble device acquisition model to foretell lung malignant neoplastic disease risk. The proposed solution will offer a more accurate and reliable prediction system through the combination of several classifiers and optimizing their performance.This model canhelp medical workers to do early screening, risk assessment, and personalized treatment planning which will lead to lower mortality and better patient outcomes

II. LITERATURE REVIEW

Zhuang et al. (2024) [1] tested the stability of the imaging-based risk models of lung cancer by comparing the Sybil model through low-dose CT scan with various reconstruction parameters. They discovered that the outputs of prediction affected the changes in slice thickness and reconstruction kernels, which is a serious drawback of prediction outcomes in imaging-based models. This highlights the need to come up with more stable and generalized predictive systems to be used in learned profession organization pattern. Hossain Sarkar et al. (2024) [2] suggested a machine learning model that combines both clinical manifestations and etiological determinants of lung cancer. They compared several algorithm, such as Determination Tree, SVM, Random Forest and XGBoost, and discovered that Determination Tree model was the best in terms of accuracy (98.43) and AUC (0.983). Through

this research, the authors have shown that structured clinical data is effective in risk prediction of lung cancer. A hybrid stacking ensemble framework is a exemplary proposed by P. V. S. et al. (2025) [3] and consists of a combination of various device acquisition algorithms, including Random Forest, Slope Boost, SVM, and LightGBM. Their model had a 95.69%accuracy rate and SHAP explainability to enhance interpretability. The paper shows the significance of hybrid ensemble methods in improving the predictive performance, as well as making medical decisions transparent. Al Mamlook et al. (2020) [4] dedicated their attention to statistical investigation of animation in respiratory organ malignant neoplasm patients with the use of performance scoring systems, ECOG, and Karnofsky. They concluded that ECOG scores were the best predictors of survival risk than other measures. This paper offers a perspective of the conventional risk assessment procedures and their position in clinical prognosis. Singh and Taneja (2022) [5] have investigated the hazard factor of lung malignant neoplastic disease in non mokers, where the environment, work hazards, passive smoking, and genetic predisposition were reportedly relevant. Their article emphasizes that lung cancer is not only a disease of smokers, why it is important to include a variety of risk factors in the prediction models. Chitra et al. (2024) [6] have suggested an Optimized Weighting-Based Enhanced Neural Network (OWENN) of lung cancer detection with the help of CT images. Through the combination of preprocessing methods and optimization methods like Particle Swarm Optimization, their model increased the classification accuracy and efficiency. This paper proves how optimization methods can be used to improve the performance of deep learning. Goyal et al. (2024) [7] compared the different machine learning models such as the Random Forest, Naive Bayes, Gradient Boosting, and the Logistic Regression in lung cancer prediction. Their comparative analysis has revealed the advantages and disadvantages of both models, which assist a researcher in choosing appropriate algorithmsin a particular predictive task. Recent developments in the machine learning methods in the prediction of lung cancer have been reviewed by Swethaet al. (2025) [8], who presented high-performing models, including Rotation Forest and XGBoost with high AUC values. They state in their study that advanced ML algorithms can be used to enhance early detection and risk stratification. Fulga et al. (2025) [9] investigated the application of data visualization methods to determine the riskfactors that are central to lung cancer. They have found that passive smoking, alcohol drinking, and obesity are some of the factors that are strong predictors of lung cancer using correlation heatmaps and statistical plots. Weiet al. (2026) [10] have thoroughly reviewed the uses of artificial intelligence in the risk assessment of lung nodules. Their paper has discussed the classification of malignancies, predicting the metastasis, and tracking the progress of the disease, but also covered the problems of data scarcity, inability to interpret and predict, and generalization of AI models. Chen et al. (2023) [11] suggested a transformer-based modelonthe use ofelectronic claims records to predict lung cancer on a large scale. Their research indicated the utility of longitudinal healthcare data to detect risk of cancer at an early stage, but the model had moderate results in comparison withimaging-based systems. Zhao et al. (2025) [13] proposed the Sequential Multi-Instance Learning (SMILE) framework, which takes the form of various CTscans across time to make predictions of lung cancer risk without necessarily performing nodule manual annotation. Their method enhanced the performance of classification and minimized the workload of the radiologists. Rao andArshad (2023)[14] highlighted the applications of deep learning and convolutional neural networks (CNN) in early lung cancer diagnosis by means of medical images. Their model incorporated image preprocessing, image segmentation and feature extraction, which showed that structured pipelines are valuable in enhancing the accuracy of diagnostic processes. The system of the CNN-based lung cancer prediction was created by Indrakumari et al. (2024) [15] and incorporated into a web application. Through symptom analysis and risk factor analysis of the patients, this system was ableto offer an early prediction of risks and enhanced access to screening tools. Patel et al. (2021) [16] suggested a class-conscious CT radiomics and deep learning model to forecast the status of mutation and survival in non-small cell lung cancer without recurrence. They emphasized the prognostic role of imaging characteristics alongside the diagnostic ability in their study. Quadri and Vidyullatha (2025) [17] have reviewed machine learning and deep transformation methods of automated lung cancerdiagnosis. They have addressed the major issues of data scarcity, uninterpretability, and obstacles to clinical adoption and recommended the wayforward in the future research

III. METHODOLOGY

A. Data Collection The data applied in the research are demographic, lifestyle, and clinical factors that can be used to prognosticate the risk of respiratory organ malignant tumor. These are variables that comprise age, gender, exposure to air pollution, the use of cigarette, genetic predisposition, long-term lung illnesses, and symptoms such as fatigue, wheezing, and coughing of blood. The data was gathered using publicly available datasets of healthcare and in a tabular format that could be analyzed by a machine learning algorithm. It is important to have a diverse and representative data to enhance the ability of the model to generalize.

B. DataPreprocessing Preprocessing of data was performed in bidding to improve quality and consistency of the data set prior to training the model. There were missing valuesthat were dealt with by applying the right imputation methods to prevent loss of data. Gender and smoking status are collection changeant, and they were converted to numerical forms using label encoding or one-hot encoding. The feature scaling methods were used to make sure that all features add equal contribution to the model like standardization or normalization. This is an important step to enhance the model performance and convergence. Standardization

$$\text{Formula: } Z = (X-\mu)/\sigma \text{----(1) Normalization}$$

$$\text{Formula: } X' = (X- Xmin)/ (Xmax-Xmin)$$

C. FeaturesSelection and Analysis.-----(2) The process of attribute option was used to determine the most multipurpose characteristics that were used to predict the risk of lung cancer. The significance of each variable was assessed by statistical approaches and the feature importance techniques. Smoking habits, genetic risk and respiratory symptoms were found to be the strong predictors. Minimal irrelevant or redundant features will be reduced, which will assist in the efficiency of the model, decreased overfitting, and interpretability. Correlation

$$\text{Formula: } r = \Sigma[(X- \bar{X})(Y- \bar{Y})] / \sqrt{[\Sigma(X- \bar{X})^2 \Sigma(Y- \bar{Y})^2]} \text{-----(3)}$$

D. ModelDevelopment Several machine learning algorithms were used asbase learners in this study, which are Support Vector Machine, Random Forest (RF), K-Nearest Neighbors, and Logistic Regression. Each of the models was trained on the preprocessed data individually to get to know different patterns and relationships on the data. This is because these algorithms have seen to be efficient in classification process and can take up the linear and non-linear relationship. Stacking Ensemble

Technique An ensemble learning method of stacking was utilized to bring the merits of the underlying base models. The output of the base learners (SVM, RF, KNN and LR) in this approach are used as input features in a meta-classifier that makes the final prediction. This interim method is worse than the others in status of model inaccuracy and overall strength because it utilizes the diversity of various classifiers. The stacking method is useful in reducing the weaknesses of each modeland improving prediction.

Stacking Formula: Z=f(h1(X), h2(X), h3(X), ..., hn(X))-----(4)

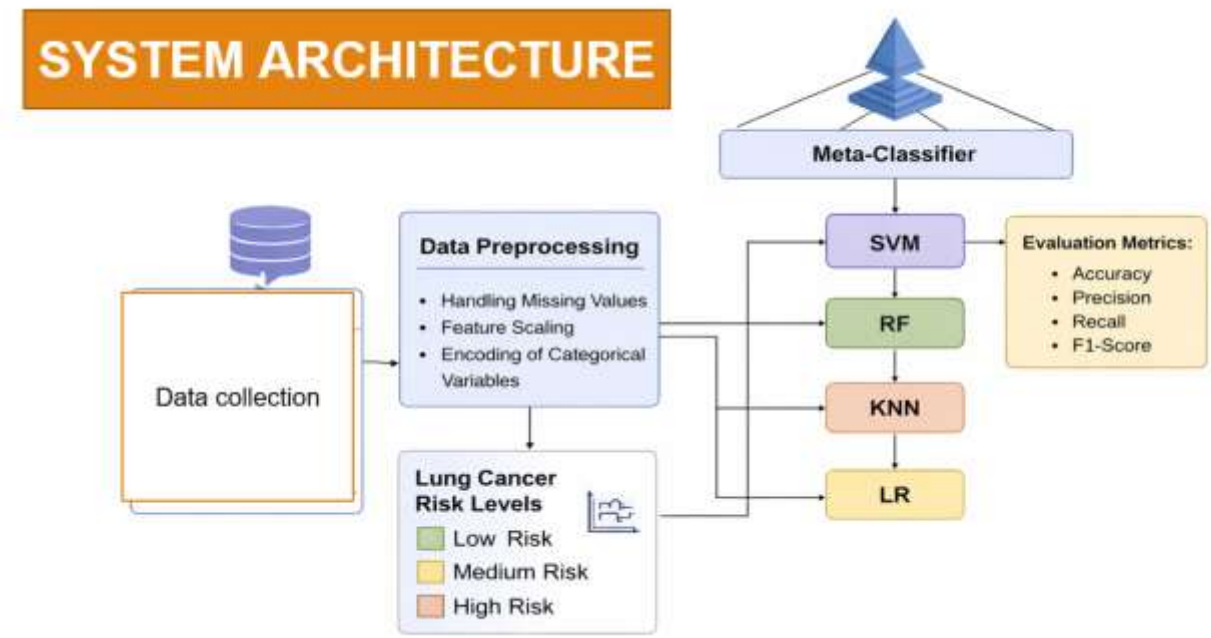
The data was separated into training and testing data to test the effectiveness of proposed model. In most cases, split ratio of 70:30 or 80:20 was applied to have enough data to train on and a few to be left to validate. The cross-validation methods were also used to minimize the bias and enhance reliability. The training set was used to train the models, and the assessment was done on the unknown test data to determine the generalization ability.

E. Performance Evaluation Metrics The models were measured in terms of standard classification measures, such as accuracy, precision, recall, and F1-score. The measures of accuracy are used to gauge the general correctness of the model, whereas the measures of precision and recall are used to measure the model in its capacity to recognize positive cases correctly. F1-score offers a balance between recall and precision, which gives a complete effect on the model performance. These were the metrics that were employed to compare the stacking ensemble model to individual classifiers.

Accuracy = (TP + TN) / (TP+TN+FP +FN)-----(5)

F. Risk Classification System The last model was intended to categorize the individuals under three risk groups Low, Medium, and High risk. This categorization will assist in determining those people who need urgent treatment and early diagnosing. The risk stratification system improves the practicality of the model in clinical decision support systems because it gives practical information to healthcare professionals

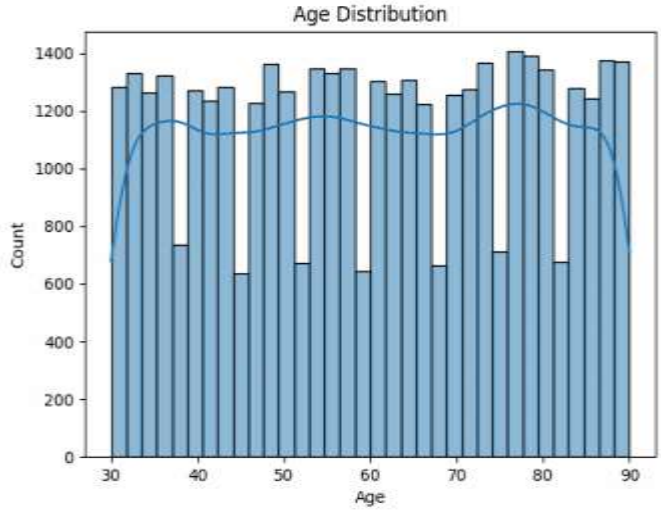
Architecture Diagram



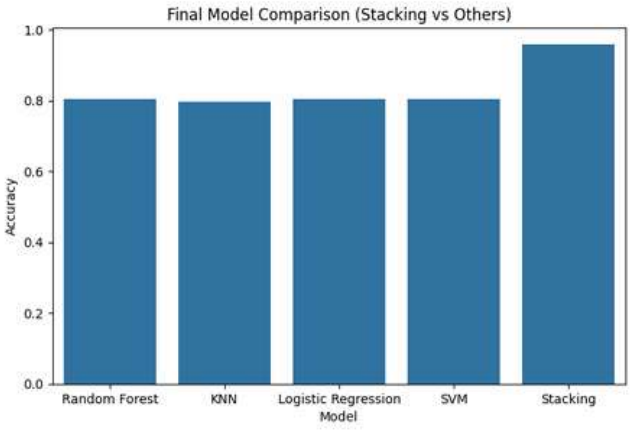
The risk prediction of lung cancer proposed system is an organized method that combines data processing, machine learning, and ensemble methods of providing valid classification. First, tthe data will be obtained, and it will comprise various demographic, lifestyle, and clinical variables such as age, gender, exposure to air pollution, smoking status, genetic predisposition, anxiety, peer pressure, chronic lung diseases, and other suitable symptoms. All these are essential in determining the potentiality of lung cancer and upon which predictive analysis is founded.Data is completely processed, and it is then inputted into machine learning algorithms to ensure the quality and consistency of data. The gaps in values are also addressed by the appropriate process of imputation in order to prevent loss of information. Encoding is the process of changing categories of variables into numbers, and feature scaling is the process of standardizing the value of all the attributes. These preprocessing functions are required to improve model efficiency, training accuracy and convergence.After preprocessing, a set of machine learning algorithms are used as base learners in order to learn various patterns in the data. They include Support Vector machine (SVM), random forest (RF), K-nearest neighbors (KNN) and logistic regression (LR). Such algorithms are trained individually upon the dataset, thus they are capable of learning different relationships between the input features and the risk of lung cancer. This is unlike the fact that although there are models that can be used to forecast linear associations, others can be used to further forecast non-linear associations that would otherwise be intricate.The stacking ensemble learning method is used to improve the output of prediction. In this approach, the existing approaches of the base learners are combined together and presented to a second-level model known as meta-classifier. The meta-classifier is an educated individual that is prepared on the most acceptable manner of incorporating the outcomes of the separate models, thereby adding the general correctness of forecasting and power. This hybrid method takes the merits of the two algorithms and reduces their demerits to the lowest possible.The data is split into training and testing data to assess the performance of the model effectively. The training data are used to train the models and the test data which is not knownis used to validate the models and determine their ability to generalize. Also, the cross-validation methods can be used to minimize bias and guarantee reliability in performance measurement. The model is

evaluated based on common metrics of evaluation, such as accuracy, precision, recall, and F1-score. They are the measures that paint the complete picture of the model efficacy in terms of predicting the risk of lung cancer. Accuracy is used to denote the total correctness and the measures of precision and recall are used to denote the measures of the model to detect correctly the positive cases. F1-score has a reliable score of the classification performance because it favors and recalls in a ratio.Finally, the developed model arranged the people into three levels of risks which are Low, Medium and High risk. This risk stratification program allows identifying persons who might need urgent treatment and facilitates clinical decision-making by means of identifying those at risk. Overall, the proposed stacking ensemble approach is a robust, accurate, and effective solution to the issue of risk prediction of lung cancer within the framework of the real-world healthcare setting.

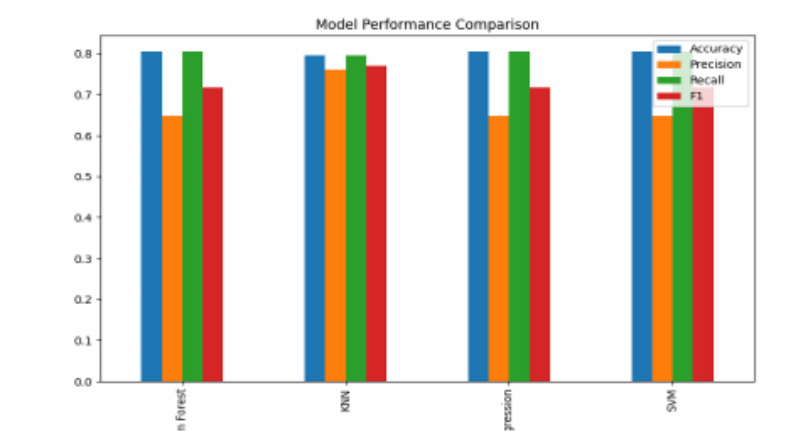
IV. RESULTANDDISCUSSION



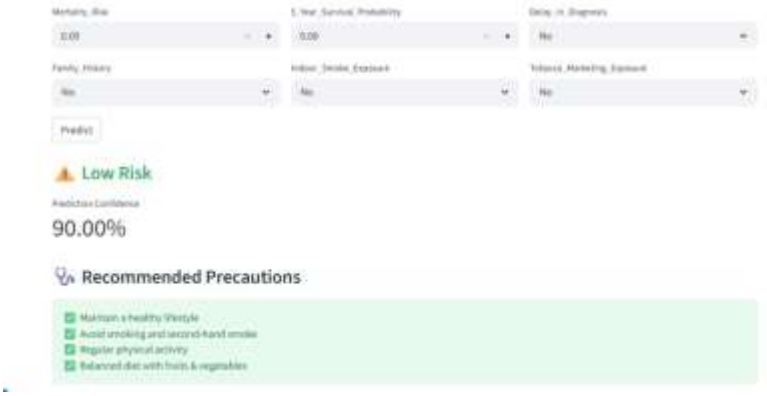
Theagepyramid chart represents the distribution of the age of the patients in the dataset which is about 30-90 years. Its distribution is rather homogeneous with a few differences among the age groups, which means that a dataset represents people representing a vast scope of ages. This heterogeneity is significant in the construction of a strong predictive model because the age is a major risk factor of lung cancer. The smooth curve of the density indicates that the data is not highly influenced by the age factor which implies that the model can beeasily generalized to other age groups.



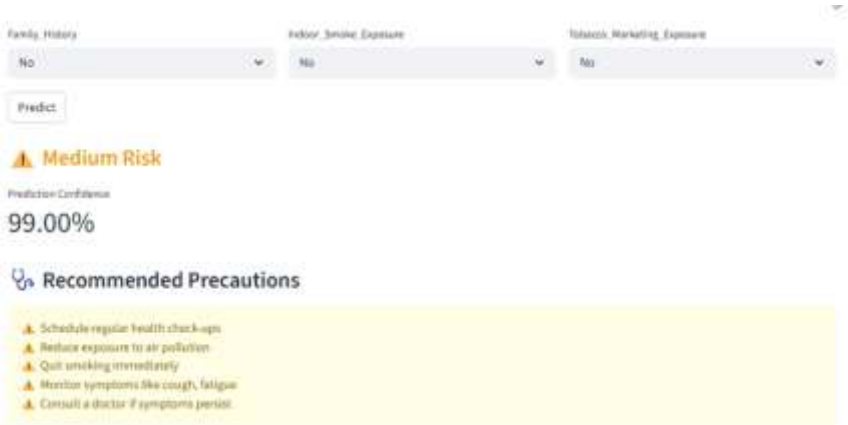
This figure compares accuracy of individual machine learning model with that of stacking ensemble exemplary. It is unsubtle that the stacking model is the most accurate of all the models considered as it rises above the rest of the models. Although the levels of performance of Random Forest, KNN, Logistic Regression, and SVM are close to 80 percent, the stacking method is much more effective in terms of accuracy of prediction. This shows that a combination of several modelscanincrease predictive ability



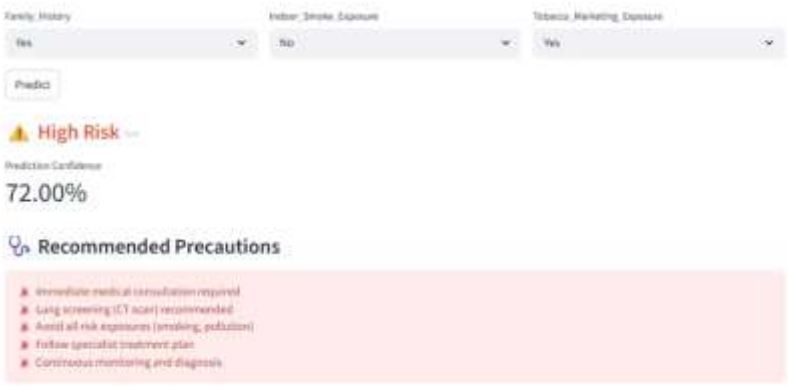
The model performance comparison chart analyses every classifier according to quality, precision, asking and F1-score. The findings demonstrate that the individual models have a reasonable performance but they vary in the value of their exactitude and recall. An example is that, KNN presents relatively balanced results in all the metrics, compared to other models, which might perform better on a single metric and worse on the other. This difference emphasizes the shortcomings of individual models and glorifies the benefit of employing ensemble techniques in order to obtain more consistent and dependable findings



This number shows the system interface in the case of a patient being termed as Low Risk. The model has a high confidence of prediction of 90 percent meaning it is very certain about what it is giving. Moreover, some preventive measures proposed by the system include having a healthy lifestyle, not smoking, exercising, and eating a balanced diet. This output demonstrates the practical applicability of the worthy in making not only the predictions but also taking action on health recommendations.



The medium risk prediction output indicates a high probability of moderate risk with 99 per cent confidence level of the classification. Some of the precautionary measures suggested by the system include frequent health examinations, minimizing air pollution, stopping smoking, and symptom monitoring. Such a degree of prediction plangstromys a very all-important role in the field of early intervention since it enables patients to undertake preventive measures before the situation maydeteriorate.



This value is the output in the case of a patient being considered as High Risk, and the prediction confidence is 72%. The system highly suggests urgent medical check-up, screening of the lungs (CTscans), elimination of harmful exposures, and continuous monitoring. This shows that the system is able to detect critical cases and offer urgent recommendations hence it is widely applicable in clinical decision support and early diagnosis

V. CONCLUSION

This paper introduced a stacking ensemble learning-based lung cancer risk assessment predictive model that combines various device learning classifiers, such as Activity Vector Machine, Random Forest, K-Nearest Neighbors (KNN), and Logistic Regression (LR). The main aim was to enhancethelevel of accuracy and reliability of the lung cancer risk prediction through integration of the strength of individual models to form coherent framework.The experimental findings proved that the stacking ensemble model had a much better performance and F1-score compared to the individual classifiers. The proposed model allowed the researcher to identify both bilinear and NON inear relationships within the data set and thus was better in predictive performance and generalization due to the diversity of different algorithms. The feature importance analysis also showed that the smoking habits, genetic orientation, and respiratory symptoms are important factors that can be used to identify the risk of lung cancer. Besides its ability to predict with high accuracy, the model is useful in the classification of individuals providing Low, Medium, and High-risk classes that can be used to make meaningful insights on early detection and clinical decision-making. The system also provides useful recommendations according to the estimated risk level, which makes it more useful in real-life healthcare implementation. This contributes to the fact that the suggested approach can be a necessary instrument helping medical workers to recognize high-risk patients and make timely interventions. Although the study is effective, it has a number of weaknesses such as reliance on the quality of the dataset and lack of real-time clinical data. The direction of work in the future can be the use of larger and more diverse datasets, the integration of time patient observation , and the implementation of more sophisticated deep learning methods to achieve even higher accuracy in prediction. Moreover, the applicability of the model in clinical settings can be enhanced by enhancing model interpretability and validation with clinical data collected in hospitals. To sum up, the suggested stacking ensemble-based predictive model is a strong, precise, and scalable answer to lung cancer risk assessment. The fact that it can provide effective predictions and actionable insights underscores its potential as a valuable decision support system to help diagnose patients early and achieve better patient outcomes

REFERENCES

[1] L. Zhuang, A. Yadav, G. H. Kim, S. M. H. Tabatabaei, A. Prosper, and W. Hsu, “Exploring the Impact of Acquisition and Reconstruction Parameters on an Imaging-Based Lung Cancer Risk Model,” 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society(EMBC),Orlando, FL,USA,2024,pp. 1– 5.
[2] M. M. H.Sarkar, S. Afrin, M. T. Reza, M. A. H. R. Bokshi,and S. S. Mim, “LungCancer Prediction and Risk Assessment: A Machine Learning Approach Integrating Symptoms and Etiological Factors,” 2024 IEEE 3rd International Conference on Robotics, Automation, Artificial-Intelligence and Internet-of-Things (RAAICON), Dhaka, Bangladesh, 2024, pp. 19– 23.
[3] P. V. S., P. R., D. N., and S. V. C., “Ensemble-Based Lung Cancer Risk Prediction with SHAP Explainability,” 2025 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT), Karaikal, India, 2025, pp. 1– 6.
[4] R. E. Al Mamlook, H. F. Bzizi, and S. Chen, “Evaluate Performance Risk Score in Patients Suffering from Lung Cancer Using Survival Analysis of Statistics,” 2020 IEEE International Conference on Electro Information Technology (EIT), Chicago, IL, USA,2020,pp. 145– 150.
[5] D. Singh and A. Taneja, “Assessment of Risk Factors of Lung Cancer in Non-Smokers in India,” 2022 7th International Conference on Intelligent Informatics and Biomedical Science (ICIIBMS), Nara, Japan,2022,pp. 7– 12.
[6] T. Chitra, S. V. Hemanth, S. Karthikeyan, V. R. Reddy, K. M. Banupriya, and G. Amirthayogam, “Predicting Lung Cancer Disease Using Optimized Weighting-Based Enhanced Neural Network Classification,” 2024 3rd International Conference on Artificial Intelligence for Internet of Things (AIIoT), Vellore, India, 2024,pp. 1– 6.
[7] A. Goyal, R. K. Shrivastava, M. Agarwal, and N. Joshi, “Enhanced Prediction of Lung Cancer Using Machine Learning,” 2024 7th International Conference on Contemporary Computing and Informatics (IC3I), Greater Noida,India, 2024, pp. 1478– 1483.
[8] T. Swetha, J. Sabeena, S. G, K. Hema, B. Rajani, and T. G. Krishna, “Lung Cancer Risk Prediction: A Machine Learning Approach,” 2025IEEEInternationalConference onAdvanced Computing Technologies (ICACT), Tirupati, India, 2025, pp. 367– 372.

[9] F. Fulga, D. Velciov, and A. Topirceanu, "On the Impact of Data Visualization in Understanding Risk Factors for Lung Cancer Detection," 2025 IEEE 19th International Symposium onAppliedComputational IntelligenceandInformatics(SACI), Timisoara,Romania, 2025,pp. 139– 144.
[10] Y. Wei et al., "Review of Artificial Intelligence in Lung Nodule Risk Assessment," IEEE Reviews in Biomedical Engineering, vol. 19, pp. 412– 427, 2026.
[11] H.-Y. Chen, H.-M. Wang, C.-H. Lin, R. Yang, and C.-C. Lee, "Lung Cancer Prediction Using Electronic Claims Records: A Transformer-Based Approach," IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 12, pp. 6062– 6073, Dec. 2023.
[12] O. B. S. Sai, D. Vighnesh, N. N. Varshitha, P. Nandhini, and I. T. J. S., "A Multifaceted Approach to Lung Cancer Diagnosis Using Machine Learning Techniques," 2024 4th International Conference on Sustainable Expert Systems (ICSES), Nepal,2024,pp. 658– 663.
[13] W. Zhao etal., "Lung Cancer Screening Classification by Sequential Multi-Instance Learning (SMILE) Framework With Multiple CT Scans," IEEE Transactions on Medical Imaging, vol. 44, no. 8, pp. 3151– 3161,Aug. 2025.
[14] B. D. Rao and M. Arshad, "Early Detection of Lung Cancer Using Machine Learning Technique," 2023 International Conference on Computer Communication and Informatics(ICCCI), Coimbatore,India, 2023,pp. 1– 5.
[15] M. Indrakumari, A. Kumar, and A. Kumar, "Lung Cancer Detection Using CNN," 2024 1st International Conference on Advances in Computing, Communication and Networking (ICAC2N), Greater Noida, India, 2024, pp. 361– 364.