International Journal of Engineering Science Invention Research & Development; Vol. 12, Issue 9, March 2026
www.ijesird.com, E-ISSN: 2349-6185

MACHINE LEARNING–DRIVEN CLINICAL DECISION SUPPORT SYSTEM FOR LUNG CANCER RISK PREDICTION

ANASWARA K[1], HARI SANKAR A [2], RAJESWARI J[3]
[1,2]UG Student, Department of Computer Science and Data Science,
Nehru Arts and Science College, Coimbatore, Tamil Nadu, India.
[3]Assistant Professor, Department of Computer Science and Data Science,
Nehru Arts and Science College, Coimbatore, Tamil Nadu, India

***ABSTRACT*- Lung cancer is a major cause of cancer deaths globally, primarily because of the late stage of diagnosis and the absence of effective early risk assessment approaches. Early detection of at-risk patients can greatly help in increasing survival rates and treatment efficacy. This paper proposes a Clinical Decision Support System (CDSS) based on machine learning to predict the risk of lung cancer based on patient clinical, demographic, and behavioural inputs. The proposed system combines supervised learning models such as Logistic Regression, Support Vector Machine (SVM), Random Forest, and Gradient Boosting to classify patients into risk groups. Various data preprocessing approaches, such as data normalization, handling missing values, and feature selection, were also incorporated to improve the efficacy and accuracy of the models. The performance of the models was measured using common metrics such as accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). The experimental outcome shows that ensemble models perform better than conventional models by accurately identifying intricate patterns among risk variables. The proposed CDSS has demonstrated immense potential in supporting healthcare professionals with data-driven risk analysis, which enables early screening and intervention strategies. The proposed CDSS can support the improvement of clinical decision-making and offer a scalable solution for the application of artificial intelligence in the practice of preventive oncology. Future work will concentrate on the integration of imaging data and deep learning.**

***KEYWORDS* : Machine Learning , Clinical Decision Support System (CDSS) , Lung Cancer Risk Prediction , Early Detection , Artificial Intelligence in Healthcare.**

I. INTRODUCTION

Lung cancer has been identified as one of the most common and lethal forms of malignancy, contributing to a large number of cancer-related deaths. Even with recent breakthroughs in medical imaging, targeted therapies, and surgical options, the mortality rate for patients with lung cancer remains high because of the lack of early diagnosis. Early diagnosis and proper risk stratification are essential for improving the survival rate of patients with lung cancer and decreasing the economic burden of healthcare. However, the traditional diagnostic approach often relies on symptomatic presentation and imaging studies, which may be delayed until the disease is advanced.

Recent developments in artificial intelligence and machine learning have opened up revolutionary frontiers in predictive healthcare. Machine learning algorithms have the ability to process complex and high-dimensional clinical data to identify hidden patterns and correlations between risk factors. By using structured patient data including demographic information, behavioral factors, medical history, and environmental exposures, predictive models can help healthcare professionals to identify high-risk patients prior to the onset of severe symptoms.

Clinical Decision Support Systems (CDSS) are intelligent systems that have been developed to assist medical professionals in decision-making tasks. The integration of machine learning into CDSS improves predictive accuracy and minimizes the subjective nature of risk assessment. The proposed work describes a machine learning-based CDSS tailored for lung cancer risk assessment based on structured patient information. The CDSS assesses the performance of various supervised machine learning models and selects the best-performing predictive model.

The main goals of this work are:

i. To build an accurate predictive model for risk classification of lung cancer.
ii. To compare the performance of conventional and ensemble machine learning algorithms.
iii. To develop a scalable and interpretable framework for CDSS.
iv. To lay the groundwork for future work that incorporates multimodal data such as imaging and genomic information.
v.

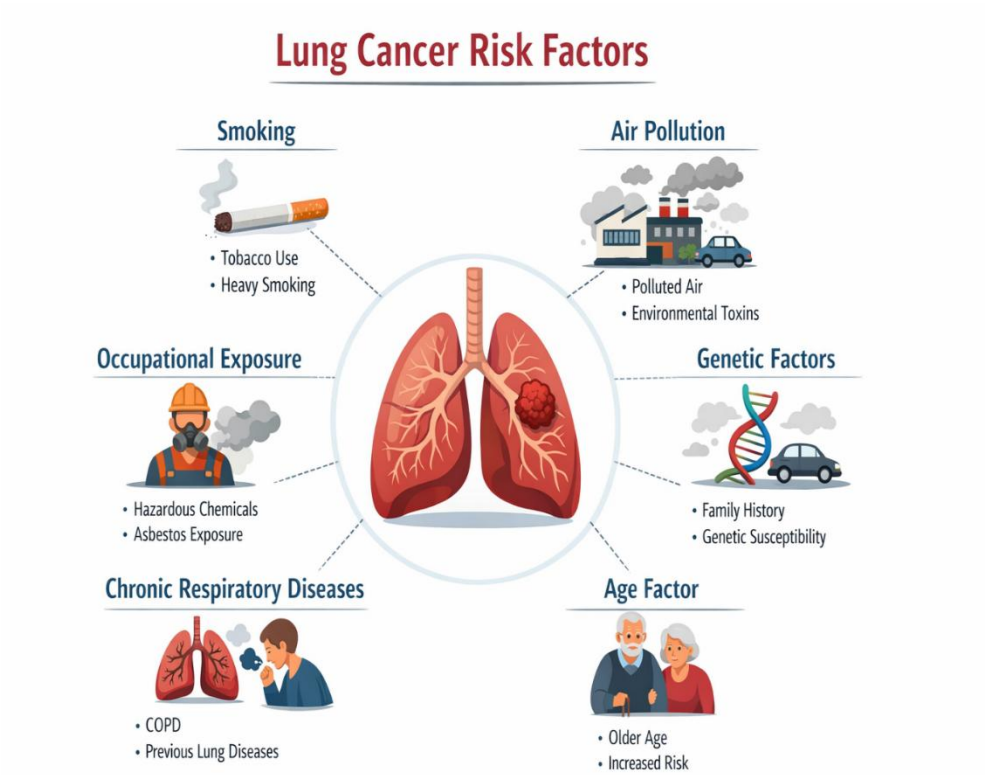
II. BACKGROUND AND SIGNIFICANCE

Lung cancer mainly arises as a consequence of exposure to carcinogens like tobacco smoke, environmental pollutants, and occupational hazards. Genetic factors and lifestyle choices also influence

the development of lung cancer. In the early stages, lung cancer may not cause significant symptoms, thus resulting in late diagnosis.

Risk prediction models are very useful in preventive oncology. It is important to identify patients at high risk of lung cancer so that early screening procedures like low-dose computed tomography (LDCT) can be carried out. However, the current risk prediction models are based on inadequate statistical models and cannot capture the nonlinear relationships between various risk factors.

Risk Factors Image:



Machine learning algorithms have several advantages over traditional statistical models. They can:

Process high-dimensional data.

- Capture nonlinear relationships.
- Automatically detect important features for prediction.
- Improve generalization performance using ensemble methods.

Thus, the implementation of a CDSS using machine learning algorithms can improve early detection approaches and lower mortality rates.

III. LITERATURE REVIEW

There have been a number of studies that have explored the use of machine learning in lung cancer diagnosis and prediction. Logistic Regression has been conventionally used in epidemiological research for binary risk prediction because of its interpretability and ease of use. It assumes linear relationships between predictors and outcomes, which restricts its predictive accuracy in complex medical data.

Support Vector Machines (SVM) have been shown to be highly accurate in classification problems by maximizing the margin between classes. SVM is very useful in high-dimensional data but is sensitive to kernel choice and parameters. Random Forest, an ensemble learning method using decision trees, avoids overfitting by combining the predictions of multiple trees. It offers variable importance scores and is robust to missing data.

Gradient Boosting algorithms involve the progressive reduction of errors in weak learners.

These algorithms have been observed to perform better than conventional classifiers on structured healthcare data.

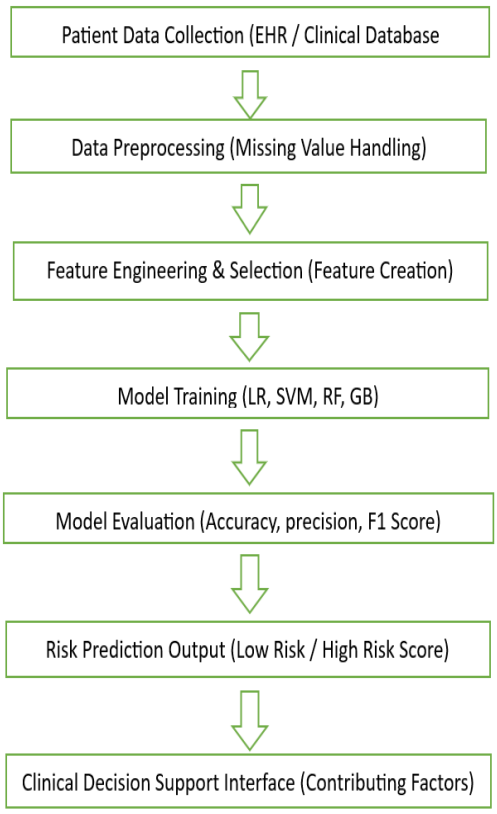
Current research trends suggest that ensemble learning techniques tend to provide better accuracy than individual learning algorithms. Most research, however, is limited to imaging-based diagnosis and not structured healthcare data. This research work fills this gap by designing a comprehensive CDSS based on non-imaging patient information.

IV. PROPOSED SYSTEM ARCHITECTURE

The proposed Machine Learning-Driven Clinical Decision Support System (CDSS) is intended to be a modular and scalable architecture that is capable of supporting the accurate prediction of lung cancer risk in a clinical environment. The system architecture is organized to facilitate efficient data processing, effective model training, and smooth integration with the clinical environment. The system is organized as a pipelined architecture in which raw patient data is processed from raw input to meaningful risk predictions through a series of processing steps.

The architecture, as proposed, has taken into consideration the need to be adaptable, as well as to be applicable in real-time situations. Each of the modules has the ability to work independently of the other modules, yet at the same time, there is an interconnectedness that allows improvements to be made without affecting the entire system. There is a focus on the security of the data, as well as the need to be transparent, which makes the architecture a reliable platform, as it combines the technical capabilities of machine learning with the practical need to be used in the real world.

CDSS Workflow Flowchart:



This organized architecture improves flexibility, reproducibility, and future scalability of the model.

The first step in the proposed system is data acquisition in which organized patient data is extracted from electronic health records, clinical databases, or publicly available sources. The system is intended to be flexible enough to handle demographic, behavioral, and clinical variables related to lung cancer risk. To ensure data privacy and confidentiality, anonymization methods are employed prior to data processing. The system architecture is intended to support integration with relational databases to enable real-time data access and updates, which is critical for implementation within a hospital setting.

After the acquisition of data, the preprocessing module is responsible for ensuring the quality of the data. In clinical datasets, there are often instances of missing data, inconsistencies, and outliers that can have a profound impact on the performance of the model. Hence, this layer is responsible for performing cleaning tasks to eliminate duplicates, address inconsistencies in formatting, and address missing values using statistically valid imputation

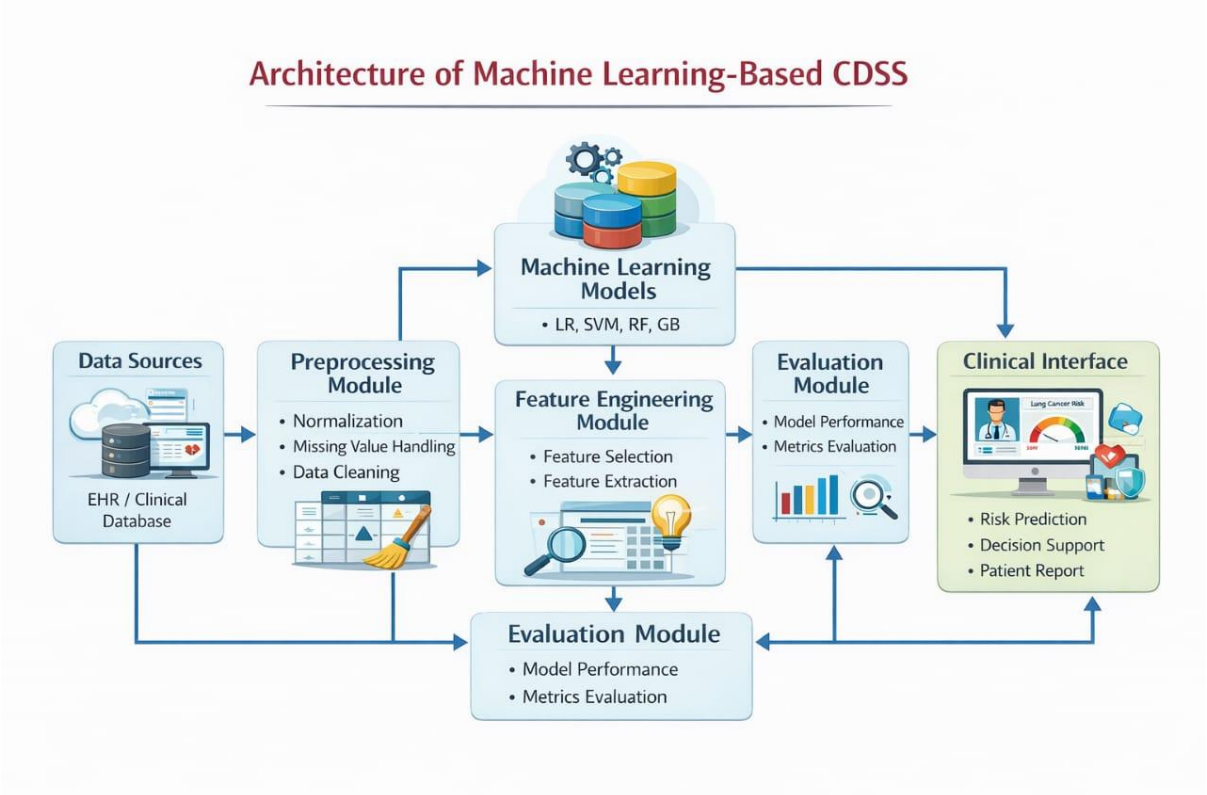
methods. Scaling is performed on numerical variables to address the issue of scaling, which is a critical requirement for distance-based algorithms. Categorical variables are converted to machine-readable formats using encoding schemes. This step is essential to ensure that the data is standardized and ready for model training.

The feature engineering and selection step is an essential component in improving the accuracy of the model. In clinical datasets, variables often need to be transformed to identify valid relationships. New features, such as cumulative smoking exposure or aggregated symptom indicators, are created to enhance representational capability. Feature selection methods are used to eliminate redundant and highly correlated features, thus addressing the issue of dimensionality and multicollinearity. By selecting the most relevant features, the system enhances computational efficiency and generalization ability of the model.

The stage of model development involves the use of various supervised learning models for comparative analysis. The models are trained on the dataset using stratified training and testing sets. Hyperparameter optimization methods are used to optimize the models and avoid overfitting. The design is flexible and can accommodate other models in the future based on various research developments.

The final component of the system involves evaluation and clinical implementation. The performance of the classification is measured using various metrics, and the model with the highest performance is chosen for implementation. The decision support interface allows clinicians to view the risk score and contributing factors, which can aid in medical decision-making. The design is flexible and can be extended to include image data and deep learning models in future research.

System Architecture Diagram:



V. DATASET DESCRIPTION

The dataset used in this research work consists of structured patient data with demographic, behavioral, clinical, and environmental features related to the risk of lung cancer. The dataset consists of variables like age, gender, smoking habits, alcohol intake, occupational exposure, respiratory symptoms, family medical history, and comorbidities. The target variable is binary,

indicating whether a person is classified as low risk or high risk for lung cancer.

The dataset was analyzed to ensure that the classes are balanced and representative of various patient populations. Analysis of the data distribution showed that there were differences in smoking duration and age categories, which are essential predictors of lung cancer. Statistical analysis methods were employed to gain insights into the distribution of features, identify skewness, and measure correlations between the independent variables and the target outcome variable.

To improve the reliability of the results, the data was split into training and testing sets using an 80:20 ratio. Stratified sampling was used to ensure that the training and testing sets had a proportional representation of the risk categories.

Dataset Description Table:

FEATURE NAME	TYPE	DESCRIPTION
Age	Numerical	Age of the patient (in years)
Gender	Categorical	Biological sex of the patient (Male/Female)
Smoking Duration	Numerical	Number of years the patient has smoked
Alcohol Intake	Categorical	Alcohol consumption status (Yes/No)
Occupational Exposure	Categorical	Exposure to hazardous substance (Yes/No)
Respiratory Symptoms	Numerical	Count of reported respiratory symptoms
Family History	Categorical	Family history of lung cancer (Yes/No)
Target Variable	Binary	Lung cancer risk classification (Low Risk/High Risk)

VI. DATA PREPROCESSING

Data preprocessing is an essential step in the predictive modelling process. In medical datasets, there can be missing values because of non-disclosure by patients, errors in data recording, or issues in data integration. To handle the problem of missing values, imputation was done using the mean or median value for numerical variables based on their distribution. Mode imputation was done for categorical variables to maintain the representation of the majority class.

Outlier detection was done using statistical range-based methods to detect values that can have a disproportionate effect on model training. Outliers were examined before deletion to avoid removing medically significant variations. Normalization was done for numerical variables to maintain equal value ranges, especially for algorithms that are sensitive to feature scales, like Support Vector Machines.

Categorical variables were converted into numerical forms through the use of encoding techniques. This made it possible to process the data using machine learning techniques. The preprocessing framework made it possible to ensure that the data was clean and standardized.

VII. FEATURE SELECTION AND ENGINEERING

feature selection is one of the most important factors that impact the predictive accuracy of machine learning models. in this research, correlation analysis was used to detect features that are highly correlated with each other. features that are strongly correlated with the target variable were given preference during the modelling process.

recursive feature elimination methods were employed to remove less informative features while preserving important predictors. moreover, tree-based models were employed to determine the importance of features, providing information on the importance of each attribute. smoking history, age, chronic respiratory diseases, and occupational exposure were identified as important predictors.

the feature engineering methods were applied to improve the expressiveness of the models. new features, like the cumulative smoking index and the sum of symptoms, were created to model the nonlinear relationships between the risk factors and the outcome.

VIII. DEVELOPMENT OF MACHINE LEARNING MODELS

Four supervised machine learning algorithms were applied and compared in this research. Logistic Regression was used as the baseline model because of its simplicity and interpretability. It predicts the probability of lung cancer risk based on a logistic function and yields estimates of coefficients that indicate the relative importance of features. Although it is interpretable, it assumes linear relationships between predictors and the outcome.

Support Vector Machine was applied using a radial basis function kernel to handle nonlinear decision boundaries. SVM seeks to find the maximum margin between classes, which is advantageous in high-dimensional spaces. However, its performance is highly sensitive to parameter selection.

Random Forest, an ensemble learning algorithm consisting of decision trees, was applied to enhance the robustness of the classification task. By combining the predictions of multiple decision trees grown on bootstrap samples, Random Forest mitigates variance and overfitting. Additionally, it yields rankings of the relative importance of features, improving interpretability.

The Gradient Boosting algorithm was applied as a sequential ensemble learning algorithm to iteratively reduce the error of predictions. The Gradient Boosting algorithm has shown a strong ability to capture complex interactions between risk factors . Hyperparameter optimization was performed using cross-validation to ensure optimal performance of the models.

IX. MODEL EVALUATION AND VALIDATION

The performance of the model was assessed using a variety of metrics to ensure that it was thoroughly tested. Accuracy was used to measure the overall correctness of the classification, while precision was used to determine the ratio of correctly predicted high-risk instances to all instances predicted as high-risk. The recall, also referred to as sensitivity, was used to determine the model's capacity to accurately predict actual high-risk instances, which is critical in medical prediction.

The F1-score was also used to provide a harmonic mean of precision and recall, which is critical in balancing the model's performance in terms of both false positives and false negatives.

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was used to determine the overall discriminative power of the model at different classification thresholds. Cross-validation was also used to lower the variance and ensure that the performance was accurately estimated.

The results showed that the ensemble models performed better than linear models, with Gradient Boosting having the best predictive performance and AUC score. The decrease in false negatives also showed that the model had a better early risk detection performance.

X. RESULTS AND PERFORMANCE ANALYSIS

The results from the experiment show that the use of ensemble-based learning models improves predictive performance significantly compared to traditional statistical approaches. Logistic Regression offered interpretability but was limited in modeling nonlinear interactions among features. Support Vector Machine improved the classification boundary but was highly parameter-intensive.

Random Forest showed robust generalization performance by reducing variance through the combination of multiple decision trees. However, Gradient Boosting offered the best predictive performance because of its cumulative error correction process. The improved AUC value shows excellent discriminative ability, ready for clinical application.

Analysis of feature importance showed that smoking duration, age, respiratory symptoms, and occupational exposure are key risk factors for lung cancer.

Feature Importance Table

FEATURE	IMPORTANCE SCORE
Smoking Duration	High
Age	High
Respiratory Symptoms	Medium
Occupational Exposure	Medium
Alcohol Intake	Low

XI. CLINICAL IMPLICATIONS

The proposed CDSS has important clinical applications, as it enables the early detection of high-risk patients. The CDSS provides a quantitative risk value, which helps physicians make informed screening decisions. Early detection enables the timely performance of diagnostic imaging and interventions, which could lower mortality rates.

The proposed CDSS also applies the principles of personalized medicine, as it takes into account multiple risk factors. The system has a scalable design that enables its integration into hospital information systems.

XII. LIMITATIONS AND FUTURE WORK

Although the proposed system shows promising results, it only depends on structured clinical data and does not use imaging or genetic data. Future work should investigate the integration of multimodal data, such as the analysis of CT scans through deep learning models. Increasing the diversity of the dataset over various geographic locations would be beneficial for generalization. Finally, the use of explainable AI would further improve transparency.

XIII. CONCLUSION

The current research offers a holistic machine learning-based Clinical Decision Support System for the risk of lung cancer. Through the combination of supervised machine learning algorithms and efficient preprocessing techniques, the proposed system is able to classify patients into risk groups. The experimental results clearly show that ensemble learning approaches, especially Gradient Boosting, are superior to conventional models in terms of prediction accuracy and generalization performance. The proposed CDSS has a high potential for use in early screening and the improvement of preventive oncology practice.

In addition to the predictive accuracy, the proposed system emphasizes the increasing role of data-driven approaches in the revolution of preventive healthcare practices. By integrating structured clinical data with sophisticated machine learning algorithms, the CDSS offers a flexible and scalable solution that can be applied to other contexts of cancer risk prediction. The capacity of the model to offer reliable and interpretable risk predictions helps healthcare professionals in identifying high-risk patients for early diagnostic interventions, which ultimately helps in improving patient outcomes. With the increasing adoption of digital infrastructure in healthcare settings, the integration of intelligent decision support systems like the one proposed in this study is a crucial step toward precision medicine and proactive disease management.

REFERENCES

1. Berge, Geir Thore, Ole-Christoffer Granmo, Tor Oddbjørn Tveit, Bjørn Erik Munkvold, Anna Linda Ruthjersen, and Jivitesh Sharma. "Machine learning-driven clinical decision support system for concept-based searching: a field trial in a Norwegian hospital." BMC Medical Informatics and Decision Making 23, no. 1 (2023): 5.
2. Yang, Xing, Jianyuan Liu, Xiaozhi Huang, Hao Liang, Ping Cui, Shiran He, Heng Zhang et al. "Machine learning-driven clinical decision support for low bone mineral density: a web-based prediction model with explainable AI integration." Bone (2025): 117592.
3. Huang P, Lin CT, Li Y, Tammemagi MC, Brock MV, Atkar-Khattra S, Xu Y, Hu P, Mayo JR, Schmidt H, Gingras M. Prediction of lung cancer risk at follow-up screening with low-dose CT: a training and validation study of a deep learning method. The Lancet Digital Health. 2019 Nov 1;1(7):e353-62.
4. Chen, Anjun, Erman Wu, Ran Huang, Bairong Shen, Ruobing Han, Jian Wen, Zhiyong Zhang, and Qinghua Li. "Development of lung cancer risk prediction machine learning models for equitable learning health system: retrospective study." JMIR AI 3 (2024): e56590.
5. Tu, Huakang, Yunfeng Zhao, Jiameng Cui, Wanzhu Lu, Gege Sun, Xiaohang Xu, Qingfeng Hu, Kejia Hu, Ming Wu, and Xifeng Wu. "Improving lung cancer risk prediction using machine learning: A comparative analysis of stacking models and traditional approaches." Cancers 17, no. 10 (2025): 1651.
6. Chandran, U., Reps, J., Yang, R., Vachani, A., Maldonado, F. and Kalsekar, I., 2023. Machine learning and real-world data to predict lung cancer risk in routine care. Cancer Epidemiology, Biomarkers & Prevention, 32(3), pp.337-343.
7. Huang, Peng, Cheng T. Lin, Yuliang Li, Martin C. Tammemagi, Malcolm V. Brock, Sukhinder Atkar-Khattra, Yanxun Xu et al. "Prediction of lung cancer risk at follow-up screening with low-dose CT: a training and validation

study of a deep learning method." The Lancet Digital Health 1, no. 7 (2019): e353-e362.
8. Ardila, D., et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose CT. NEJM.
9. Tu, H., Zhao, Y., Cui, J., et al. Improving Lung Cancer Risk Prediction Using Machine Learning: A Comparative Analysis of Stacking Models and Traditional Approaches. Cancers. 2025.
10. Chen, A., Wu, E., Huang, R., et al. Development of Lung Cancer Risk Prediction Machine Learning Models for Equitable Learning Health Systems. JMIR AI. 2024.
11. Yang, Y., Xu, L., Sun, L., et al. Machine learning application in personalised lung cancer recurrence and survivability prediction. Comput Struct Biotechnol J. 2022.
12. Sireesha, G. Exploring the efficacy of machine learning classifiers in lung cancer prognosis and risk assessment. Int J Intelligent Syst Appl Eng. 2024.
13. Innab, N., Aldrees, A., AlHammadi, D. A., et al. AI-Driven Predictive Modeling for Lung Cancer Detection with Synthetic Data Augmentation. Int J Comput Intell Syst. 2025.
14. Zou, Y. Lung cancer prediction based on machine learning. Highlights Sci Eng Technol. 2025.
15. Ferlay, J., Ervik, M., Lam, F., et al. Global Cancer Observatory: Cancer Today. IARC. 2024.
16. Sung, H., Ferlay, J., Siegel, R. L., et al. Global Cancer Statistics 2020: GLOBOCAN Estimates. CA Cancer J Clin. 2021.
17. Hoggart, C., Brenna, P., et al. A Risk Model for Lung Cancer Incidence. Cancer Prev Res. 2012.
18. Benveniste, P.-L., Alberge, J., Xing, L., et al. Development and external validation of a lung cancer risk estimation tool using gradient-boosting. arXiv. 2023.
19. Gupta, A., Rao, M. Explainable AI model for lung cancer detection using hybrid CNN-XGBoost. IEEE J Biomed Health Inform. 2025.
20. Sharma, P., et al. Deep learning framework for early lung cancer prediction. IEEE Access. 2025.
21. Wang, Y., Mei, N., Zhou, Z., et al. Prognostic machine learning models for non-small cell lung cancer. BMC Med Inform Decis Mak. 2024.
22. Kaur, D., et al. AI-driven lung cancer risk prediction systems. IEEE J Biomed Health Inform. 2024.
23. Zhang, L., et al. Ensemble machine learning for early lung nodule detection. 2024.
24. Zhou, H., Liu, J., & Wang, X. Lung cancer classification using gradient boosting. Expert Syst Appl. 2020.
25. Taheri, S., & Mammadov, M. Learning the Naive Bayes classifier with optimization. Int J Appl Math Comput Sci. 2013.
26. Lynch, C. M., Abdollahi, B., Fuqua, J. D., et al. Prediction of lung cancer patient survival via supervised machine learning. Int J Med Inform. 2017.
27. AI-based Clinical Decision Support Systems: Concepts, frameworks, and future directions. J Med Syst. (review)