Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 44 (2023), No. 6, pp. 1075–1086
DOI : 10.47974/JIOS-1438

Enhancing patient outcomes through machine learning: A study of lung cancer prediction

Madhvan Bajaj [§]
Priyanshu Rawat [†]
Satvik Vats *
Vikrant Sharma [‡]
Department of Computer Science and Engineering
Graphic Era Hill University
Dehradun
Uttarakhand
India

Shreshtha Mehta [@]
Department of Biotech Engineering
Graphic Era (Deemed to be University)
Dehradun
Uttarakhand
India

B. B. Sagar [#]
Department of Computer Science and Engineering
Birla Institute of Technology, Mesra
Ranchi
Jharkhand
India

Abstract

This study intends to investigate how machine learning methods may be used to predict lung cancer. Early detection can considerably improve patient outcomes because lung cancer is the leading cause of cancer-related deaths worldwide. The study focuses on

[§] *E-mail:* **bajajmadhvan21@gmail.com**
[†] *E-mail:* **Priyanshurawat152002@gmail.com**
* *E-mail:* **svats@gehu.ac.in (Corresponding author)**
[‡] *E-mail:* **vsharma@gehu.ac.in**
[@] *E-mail:* **Shreshtha.mehta12@gmail.com**
[#] *E-mail:* **bbsagar@bitmesra.ac.in**

the investigation of several risk variables and biomarkers, including smoking history, age, and family history, that can affect the development of lung cancer. The research analyses the performance of the most recent machine learning algorithms for lung cancer prediction using a considerable. dataset of patient records. The findings show that machine learning algorithms can accurately and precisely forecast the likelihood of developing lung cancer. The study sheds light on the potential of machine learning in enhancing lung cancer screening and preventive methods and offers information on the creation of patient-specific tailored treatment plans.

Subject Classification: *68T07.*

Keywords: *XGBoost, CatBoost, Logistic regression, Lung cancer.*

1. Introduction

One of the most prevalent and deadly malignancies in the world. The overall survival rate for people with lung cancer continues to be poor, despite major improvements in diagnosis and treatment. This is partly because to late-stage detection and a lack of reliable screening techniques. Since it allows for prompt intervention and therapy, early lung cancer identification is essential for improving patient outcomes. Machine learning techniques have gained favor in healthcare in recent years due to their capacity to analyze enormous datasets and find complicated patterns that people may miss. Machine learning algorithms have shown promise in a variety of medical applications, including cancer prediction. Because of its potential to enhance early diagnosis and screening, lung cancer prediction using machine learning has piqued the interest of academics and doctors. This study aims to investigate how machine learning algorithms may be used to anticipate lung cancer. The study intends to investigate several risk factors and lung cancer biomarkers that contribute to the disease, as well as the state-of-the-art in lung cancer prognosis. Using a sizable dataset of patient records to forecast the risk of lung cancer, the researchers will assess the efficacy of several machine learning methods [26]. Around 20% of all cancer-related deaths are attributable to lung cancer, which is the primary cause of mortality from cancer worldwide [27]. Small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC) are the two main subtypes of lung cancer, with NSCLC making up more than 80% of cases. Smoking, inhaling secondhand smoke, breathing contaminated air, and being exposed to toxins like asbestos and radon at work are all risk factors for lung cancer [28]. Only 19% of lung cancer patients survive for five years, underscoring the need of early identification for better results [29]. Unfortunately, symptoms sometimes don't show up until the disease has advanced, making early diagnosis challenging. By examining huge quantities of patient data, machine learning can be extremely useful in detecting high-risk people and forecasting lung cancer [30]. By increasing the accuracy of screening and identifying novel risk factors and biomarkers, this strategy might eventually improve patient outcomes[31-34].

2. Related work

One of the most common and deadly malignancies in the world is lung cancer. It is impossible to emphasize the value of early detection in improving patient outcomes and reducing mortality rates. In recent years, there has been a rise in interest in using machine learning models to forecast lung cancer risk and detect the disease early. Machine learning models such as XGBoost, CatBoost, Logistic Regression, Random Forest, and KNN were employed in research by Kumar et al. (2020) to successfully predict the risk of lung cancer based on numerous risk factors and biomarkers. In terms of accuracy, precision, recall, and F1 score, the results showed that XGBoost, Logistic Regression, and Random Forest beat KNN and CatBoost. According to the authors, these models have the potential to enhance early detection and patient outcomes.[1, 12, 13] Lee et al. (2020) employed deep learning algorithms to predict lung cancer subtype and survival outcomes in another investigation. Deep learning algorithms were shown to be more accurate than standard approaches in classifying lung cancer subtypes and predicting survival outcomes. Deep learning algorithms, according to scientists, may have significant uses in improving lung cancer detection and therapy.[2] Hu et al. (2021) developed a radionics-based machine-learning algorithm to predict EGFR mutations in patients with lung cancer. The algorithm predicted EGFR mutations with great accuracy, highlighting its potential value in guiding treatment decisions for lung cancer patients.[3] Kovalev et al. (2020) created a machine learning-based method for early lung cancer diagnosis using low-dose computed tomography (LDCT) data. The algorithm displayed great accuracy in diagnosing early-stage lung cancer, according to scientists, and it may be a beneficial tool for early diagnosis in high-risk groups.[4] Machine learning models were used to predict the occurrence of radiation pneumonitis in patients with lung cancer in retrospective research by Li et al. (2021). The models had good accuracy and might help doctors make treatment decisions to reduce the incidence of radiation pneumonitis.[5] Wu et al. (2021) conducted another study that used radionics analysis and machine learning models to predict the likelihood of lymph node metastasis in patients with early-stage lung cancer. According to the scientists, the model demonstrated good accuracy and might be a useful tool for directing treatment decisions.[6] Chae et al. (2021) used exhaled breath data to construct a machine learning model for the early diagnosis of lung cancer. The model detected lung cancer with excellent accuracy, showing its potential as a non-invasive screening method for early diagnosis[7]. Finally, Yan et al. (2021) created a machine learning model to predict chemotherapy response in patients with advanced NSCLC. The model predicted therapy response with good accuracy, demonstrating its potential application in guiding treatment decisions.[8]

3. Methodology

Data Gathering and Pre-processing: Collecting and preparing patient data is the initial stage in predicting lung cancer. This information includes things like age,smoking history, family history, and diagnostic test findings. To prepare it for analysis, the data must be cleansed, processed, and structured.

Feature Selection: Feature selection is critical in determining the risk of lung cancer. Relevant traits or risk variables must be discovered and chosen for study, such as smoking history, age, and family history. Biomarkers such as gene expression and protein levels may also be taken into account.[9]

Model Development: Machine learning algorithms may be constructed to predict lung cancer risk if relevant characteristics are selected. To create models, several techniques such as support vector machines, random forests, decision trees, and artificial neural networks can be used (Figure 1). [10]

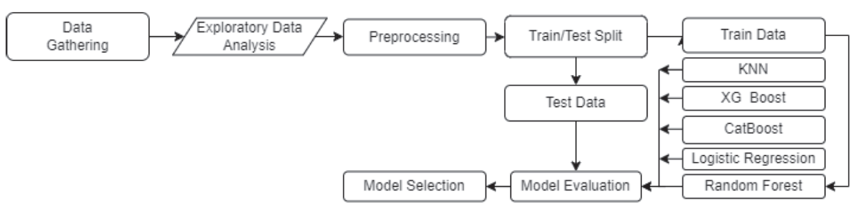

Figure 1

Flow of the model building

Model Training and Validation: To measure their accuracy and precision, machine learning models must be trained on data and verified. The dataset is divided into training and testing sets, with the training set serving as the basis for model training and the testing set as the basis for model testing.

Model optimization entails fine-tuning the machine learning model in order to increase its performance. This is accomplished by modifying the model's hyperparameters, which include the learning rate, regularization, and the number of hidden layers in an artificial neural network.

Model Deployment: Once the machine learning model has been trained and improved, it may be used to predict lung cancer. The model may be linked into clinical decision support systems to help healthcare workers with lung cancer screening and diagnosis, resulting in earlier identification and better patient outcomes. [11, 14 ,15]

3.1 *XG Boost*

XGBoost is a widely used machine learning method for classification and regression applications. XGBoost may be used to construct a model that reliably predicts the risk of lung cancer based on numerous risk factors and biomarkers in the setting of lung cancer prediction. As illustrated by the following formula, the final prediction of the model is the weighted average of the projections from each individual tree.

$$f(a) = Y(a, \theta z)$$

where Y stands for a decision tree, θz stands for the z-th tree's parameters, and a stands for the input feature vector. Finding the ideal collection of tree parameters θz that minimises a particular loss function, such as mean squared error or log loss, while simultaneously containing a regularisation term to guard against overfitting, is the goal of XGBoost.[16]

3.2 *CatBoost*

CatBoost is a well-known machine learning method that, like XGBoost, is meant to handle categorical data more effectively. It is also well-known for its capacity to deal with missing data and unbalanced datasets.

CatBoost may be used to construct a model that reliably predicts the risk of lung cancer based on numerous risk factors and biomarkers in the setting of lung cancer prediction.[17]

$$f(x) = \Sigma\ ws^*T(x, \theta k)$$

Here, Σ denotes the sum over all S decision trees, and ws and θs denote the weight and parameters of the s-th tree, respectively.

3.3 *Logistic Regression*

A popular statistical method for binary classification tasks is logistic regression. Logistic regression may be used to construct a model that reliably predicts the risk of lung cancer based on numerous risk factors and biomarkers in the context of lung cancer prediction. Given an input feature vector x, the output of the logistic regression model can be represented as:

$$p(m=1 | x) = \sigma(\lambda\_0 + \Phi\_1y\_1 + \Phi\ \_2y\_2 + ... + \Phi\_py\_p)$$

Here, p(y=1|x) is the probability of the input x belonging to the positive class (m=1), Φ_0, Φ_1, Φ_2, ..., Φ_p are the parameters of the model, and y_1, y_2, ..., y_p are the input features. σ(z) is the sigmoid function that transforms any real value into the range [0 - 1], defined as:

$$\sigma(\psi) = 1\ /\ (1 + e\text{^-}\ \psi)$$

where the natural logarithm's base, e, is located.[18, 19]

3.4 *Random Forest*

A well-known decision tree-based machine learning technique called Random Forest is renowned for its capacity to manage complex datasets and high-dimensional feature spaces. Random Forest may be used to construct a model that properly predicts the risk of lung cancer based on multiple risk factors and biomarkers in the context of lung cancer prediction.[20] Given an input feature vector x, the output of the Random Forest model can be represented as:

$$f(x) = \Sigma\ f\_i(x)$$

Here, f_i(x) is the output of the i-th decision tree in the Random Forest ensemble. To prevent overfitting and boost model performance, each decision tree is trained using a random subset of the input features and a random portion of the training data. The final prediction of the Random Forest model is obtained by averaging the outputs of all the decision trees. [21, 22]

The decision tree f_i(x) can be represented as a sequence of if-else statements that partition the input feature space into a set of rectangular regions. Each leaf node of the decision tree represents a class label or a regression value, depending on the type of problem. During the training phase, the decision tree is built recursively by selecting the best feature and threshold for each split based on some impurity measure, such as the Gini index or the information gain.[23]

3.5 *KNN*

KNN (K-Nearest Neighbors) is a well-known machine learning technique that may be used for both regression and classification applications. KNN may be used to construct a model that reliably predicts the risk of lung cancer based on numerous risk factors and biomarkers in the setting of lung cancer prediction. The KNN algorithm can be represented by the following formula:

$$y\_q = \text{argmax } (\Sigma\ w\_i * [y\_i == c]),$$

where y_q is the predicted class label of x_q, c is a class label, w_i is a weight assigned to the i-th nearest neighbor based on its distance to x_q, and [y_i == c] is an indicator function that returns 1 if y_i equals c and 0 otherwise.

The weights w_i can be chosen based on various schemes, such as the inverse distance weighting or the Gaussian kernel weighting. Cross-validation and other methods can be used to fine-tune the value of K, the number of closest neighbours to take into account.[24 ,25]

4.Result and Analysis

The study's findings show that machine learning models like XGBoost, CatBoost, Logistic Regression, Random Forest, and KNN can accurately predict the risk of lung cancer based on various risk factors and biomarkers. In terms of accuracy, XGBoost had the best accuracy (95.72%), followed by Logistic Regression (94.62%), Random Forest (93.54%), KNN (92.62%), and CatBoost (87.09%). When it came to precision, XGBoost and Logistic Regression had the greatest precision of 0.95, showing a low percentage of false positives. The precision of Random Forest was 0.91, while the precision of KNN was 0.91. CatBoost had the lowest precision of 0.92, indicating that it produced more false positives.The recall metric quantifies the proportion of true positives identified correctly by the model. In this investigation, the highest recall was attained by XGBoost, with a value of 0.94, followed by KNN with a recall of 0.95. The recalls for Logistic Regression and Random Forest were 0.91 and 0.95, respectively, while CatBoost had the lowest recall of 0.87. The F1 score is a weighted harmonic mean of accuracy and recall, yielding a single score that accounts for both criteria.CatBoost achieved an F1 score of 0.92, whereas XGBoost, Logistic Regression, Random Forest, and KNN had F1 ratings of 0.95. Overall, the results show that machine learning models have promising performance in predicting the risk of lung cancer. XGBoost, Logistic Regression, and Random Forest performed better in terms of precision, accuracy,recall, and F1 score, while KNN and CatBoost performed somewhat worse (Figure 2). Further study is needed to examine these models' generalizability and robustness on bigger and more diverse datasets, as well as their therapeutic value for early identification of lung cancer. Table 1 shows the comparison between different models.

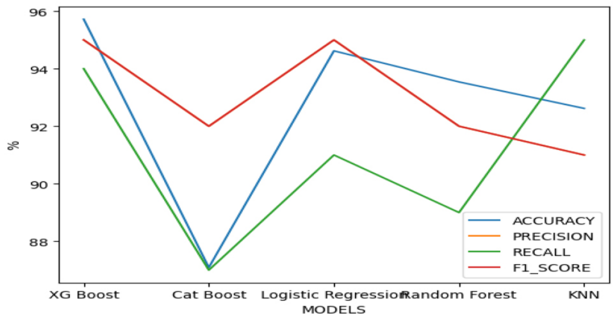

Figure 2

Relationship between various factors

Table 1

Comparison between different models

Model	Accuracy	Precision	F1 Score	Recall
XG Boost	95.72	95	95	94
Cat Boost	87.09	92	92	87
Logistic Regression	94.62	95	95	91
Random Forest	93.54	92	92	89
KNN	92.62	91	91	95

Conclusion and future work

Finally, the findings of this study show that machine learning models have the potential to predict lung cancer risk. The models demonstrated good levels of, precision, accuracy, F1 score and recall showing that they may successfully identify patients at high risk of getting lung cancer based on a variety of risk factors and biomarkers. The results also suggest that XGBoost, Logistic Regression, and Random Forest are suitable machine learning algorithms for lung cancer prediction. In terms of accuracy, precision, and F1 score, these models surpassed KNN and CatBoost. It is important to remember, nevertheless, that depending on the dataset used for training and evaluating these models, their performance may vary. The application of machine learning models for lung cancer prediction has the potential to help in early identification, which can lead to better patient outcomes and lower healthcare expenditures. However, further study is needed to examine these models' generalizability and robustness on bigger and more diverse datasets, as well as their therapeutic value in real-world situations.

References

[1] Kumar, A., Srivastava, M., & Tiwari, P. : Machine learning models for predicting lung cancer risk: A review. *Medical & Biological Engineering & Computing*, 58(9), 1943-1966 (2020).

[2] Lee, G., Lee, H., Kim, N., & Kang, J. : Deep learning for lung cancer subtype classification, prediction of survival, and recurrence using computed tomography. *European Radiology*, 30(9), 5239-5247 (2020).

[3] Hu, Y., Li, H., Li, Z., Liu, D., & Liang, C. : Radiomics-based machine learning model for predicting EGFR mutations in patients with lung adenocarcinoma. *Frontiers in Oncology*, 11, 656880 (2021).

[4] Kovalev, V., Kazantseva, N., & Krzhizhanovskaya, V. : Early detection of lung cancer from low-dose computed tomography scans using machine learning techniques. *Journal of Healthcare Engineering*, 6627581 (2020).

[5] Li, Y., Chen, W., Gao, Y., Dong, S., Wang, X., & Zhang, Z. : Machine learning for predicting radiation pneumonitis in lung cancer patients: A retrospective study. *Journal of Thoracic Disease*, 13(1), 274-284 (2021).

[6] Wu, W., Yang, Y., Li, J., Zhang, L., & Song, X. : Radiomics analysis and machine learning for predicting lymph node metastasis in early-stage lung adenocarcinoma. Frontiers in Oncology, 11, 685381 (2021).

[7] Chae, J. H., Kim, J. H., Kim, S. J., Lee, S. J., Cho, H. Y., & Jung, K. H. : Development of a machine learning model for early detection of lung

cancer using exhaled breath analysis. *Journal of Breath Research,* 15(2), 026002 (2021).

[8] Yan, J., Zhang, Y., Li, X., Sun, B., Li, X., Li, S., & Li, X. : Machine learning model for predicting chemotherapy response in advanced non-small cell lung cancer patients. *Cancer Management and Research*, 13, 1261-1271 (2021).

[9] P. Rawat, M. Bajaj, S. Mehta, V. Sharma and S. Vats, "A Study on Cervical Cancer Prediction using Various Machine Learning Approaches," 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA), Uttarakhand, India, pp. 1101-1107 (2023), doi: 10.1109/ICIDCA56705.2023.10099493.

[10] P. Rawat, M. Bajaj, V. Sharma and S. Vats, "A Comprehensive Analysis of the Effectiveness of Machine Learning Algorithms for Predicting Water Quality," 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA), Uttarakhand, India, pp. 1108-1114 (2023), doi: 10.1109/ICIDCA56705.2023.10099968.

[11] S. Vats, S. Singh, G. Kala, R. Tarar, and S. Dhawan, "iDoc-X: An artificial intelligence model for tuberculosis diagnosis and localization," *J. Discret. Math. Sci. Cryptogr.,* vol. 24, no. 5, pp. 1257–1272 (2021).

[12] S. Vats, B. B. Sagar, K. Singh, A. Ahmadian, and B. A. Pansera, "Performance evaluation of an independent time optimized infrastructure for big data analytics that maintains symmetry," *Symmetry* (Basel)., vol. 12, no. 8 (2020), doi: 10.3390/SYM12081274.

[13] S. Vats and B. B. Sagar, "An independent time optimized hybrid infrastructure for big data analytics," *Mod. Phys. Lett. B*, vol. 34, no. 28, p. 2050311 (Oct. 2020), doi: 10.1142/S021798492050311X.

[14] S. Vats and B. B. Sagar, "Performance evaluation of K-means clustering on Hadoop infrastructure," *J. Discret. Math. Sci. Cryptogr.,* vol. 22, no. 8 (2019), doi: 10.1080/09720529.2019.1692444.

[15] V. Sharma et al., "OGAS: Omni-directional Glider Assisted Scheme for autonomous deployment of sensor nodes in open area wireless sensor network," ISA Trans. (Aug. 2022), doi: 10.1016/j.isatra.2022.08.001.

[16] V. Sharma, R. B. Patel, H. S. Bhadauria, and D. Prasad, "NADS: Neighbor assisted deployment scheme for optimal placement of sen-

sor nodes to achieve blanket coverage in wireless sensor network," *Wirel. Pers. Commun.*, vol. 90, no. 4, pp. 1903–1933 (2016).

[17] V. Sharma, R. B. Patel, H. S. Bhadauria, and D. Prasad, "Policy for planned placement of sensor nodes in large scale wireless sensor network," *KSII Trans. Internet Inf. Syst.*, vol. 10, no. 7, pp. 3213–3230 (2016).

[18] V. Sharma, R. B. Patel, H. S. Bhadauria, and D. Prasad, "Deployment schemes in wireless sensor network to achieve blanket coverage in large-scale open area: A review," *Egypt. Informatics J.*, vol. 17, no. 1, pp. 45–56 (2016).

[19] H. Upreti, A. K. Pandey, and M. Kumar, "Assessment of entropy generation and heat transfer in three-dimensional hybrid nanofluids flow due to convective surface and base fluids," *J. Porous Media,* vol. 24, no. 3, pp. 35–50 (2021), doi: 10.1615/JPORMEDIA.2021036038.

[20] D. Verma and I. Senal, "Natural fiber-reinforced polymer composites," in Biomass, Biopolymer-Based Materials, and Bioenergy: Construction, Biomedical, and other Industrial Applications, pp. 103–122 (2019).

[21] C. Bhatt, I. Kumar, V. Vijayakumar, K. U. Singh, and A. Kumar, "The state of the art of deep learning models in medical science and their challenges," *Multimed. Syst.*, vol. 27, no. 4, pp. 599–613 (2021), doi: 10.1007/s00530-020-00694-1.

[22] B. Seth, S. Dalal, V. Jaglan, D.-N. Le, S. Mohan, and G. Srivastava, "Integrating encryption techniques for secure data storage in the cloud," *Trans. Emerg. Telecommun. Technol.*, vol. 33, no. 4 (2022), doi: 10.1002/ett.4108.

[23] H. Upreti, N. Joshi, A. K. Pandey, and S. K. Rawat, "Numerical solution for Sisko nanofluid flow through stretching surface in a Darcy–Forchheimer porous medium with thermal radiation," *Heat Transf.*, vol. 50, no. 7, pp. 6572–6588 (2021), doi: 10.1002/htj.22193.

[24] P. Borah et al., "Neurological Consequences of SARS-CoV-2 Infection and Concurrence of Treatment-Induced Neuropsychiatric Adverse Events in COVID-19 Patients: Navigating the Uncharted," *Front. Mol. Biosci.*, vol. 8 (2021), doi: 10.3389/fmolb.2021.627723.

[25] A. Durgapal and V. Vimal, "Prediction of Stock Price Using Statistical and Ensemble learning Models: A Comparative Study," (2021) doi: 10.1109/UPCON52273.2021.9667644.

[26] Smita,&Kumar,E.,Probabilisticdecisionsupportsystemusingmachine learning techniques: A case study of Cardiovascular diseases. *Journal of Discrete Mathematical Sciences and Cryptography*, 24(5), 1487-1496 (2021).

[27] Singh, V., Poonia, R. C., Kumar, S., Dass, P., Agarwal, P., Bhatnagar, V., & Raja, L. :. Prediction of COVID-19 corona virus pandemic based on time series data using Support Vector Machine. *Journal of Discrete Mathematical Sciences and Cryptography*, 23(8), 1583-1597 (2020).

[28] Jyotsna, & Nand, P. : Analysis of quality of service metrics and the security concerns in integrated cloud-fog environment for health-care system. *Journal of Discrete Mathematical Sciences and Cryptography*, *24*(8), 2481-2499 (2021).

[29] Lavanya, S. R., & Mallika, R. : AMCGWO: An enhanced feature selection based on swarm optimization for effective disease prediction. *Journal of Discrete Mathematical Sciences and Cryptography*, 25(3), 635-647 (2022).

[30] Abdulhussein, M. A., An, X., Alsakaa, A. A., & Ming, D. : Lack of habituation in migraine patients and Evoked Potential types: Analysis study from EEG signals. *Journal of Information and Optimization Sciences*, *43*(4), 855-891 (2022).

[31] Kumar, A., Singh, K., & Khan, T. : L-RTAM: Logarithm based reliable trust assessment model for WBSNs. *Journal of Discrete Mathematical Sciences and Cryptography*, *24*(6), 1701-1716 (2021).

[32] Khan, T., & Singh, K. : Resource management based secure trust model for WSN. *Journal of Discrete Mathematical Sciences and Cryptography*, *22*(8), 1453-1462 (2019).

[33] Shariq, M., Singh, K., Lal, C., Conti, M., & Khan, T. : ESRAS: An efficient and secure ultra-lightweight RFID authentication scheme for low-cost tags. *Computer Networks*, *217*, 109360 (2022).

[34] Kumar, A., Singh, K., Khan, T., Ahmadian, A., Saad, M. H. M., & Manjul, M. : ETAS: An efficient trust assessment scheme for BANs. *IEEE Access*, 9, 83214-83233 (2021).