Optimized K-Nearest Neighbours Algorithm for Improved Lung Cancer Prediction

Supriya DK[1], Madhushree TM [2], Manoj Kumar HG[3]
supriyadk0005@gmail.com[1] ,mmmadhushreegowda669@gmail.com[2], manojkumarhg01@gmail.com[3]

Abstract: Lung cancer is widely regarded as one of the leading causes of cancer death in the world. Therefore, this necessitates early identification which is crucial for improving survival rates. This paper proposes a new lung cancer predicting model using K-Nearest Neighbours (KNN) and cross validation. The ability of KNN to select features is determined by the precision achieved with different feature subsets and hyperparameter combinations. It also compares its performance against other models in use today. As per findings of this study, KNN with cross-validation has an accuracy of [0. 97] better than before models do. In addition, we demonstrate how to improve generalizability of models and reduce issues such as overfitting through cross-validation. The above approach might help a doctor to evaluate the first clinical stage of lungs cancer and the corresponding treatments. Nevertheless, when performing other studies, the following factors will be included because they are thought to enhance the results in terms of predictive efficiencies One will also use more than one type of machine learning method in the research although they have not been considered in this study.

Keywords: Lung Cancer, prediction, k-nearest neighbour, Cross validation, early detection, Machine-learning.

Introduction: The lung cancer is the leading cause of cancer death in the whole world; and millions of people dies from this disease every year. It has called itself a strong player in the oncological market. Fortunately, the good thing about it is that whenever such a condition is detected early, there is always some hope for therapy as well as survival. Machine learning has brought the use of AI in almost every perspective of human activity including medicine. Recently another method known as K-Nearest Neighbours [KNN] has been used to have better chance to accurately diagnose lung cancer. Finally in this novel study we utilized KNN algorithm in order to create a very accurate model for the lung cancer and resulted a prediction rate of 97%. Consequently, the optimality of our predictor’s recall, precision and F1 score were underscored with a view to attest the creativity of the work and give a clue on its possible uses such as in planning treatment of lung cancer and early diagnosis as mooted in this paper. Through the understanding of this piece of work that presents the method, results and discussion section, the conclusion to be drawn is that KNN is relatively efficient in the prediction of lung cancer. These outcomes will therefore be useful in early diagnosis of lung cases of cancer as well as offer treatment in relation to the medicine belonging to the individual.

Literature Survey: The K-Nearest Neighbour (KNN) algorithm has gained significant traction as one of the most effective machine learning approaches in the early detection and diagnosis of lung cancer. Its ability to handle complex medical datasets, which are often diverse and unpredictable, makes it particularly valuable in healthcare applications[4]. KNN’s non-parametric nature allows it to adapt well to various types of medical data, such as patient demographics, genetic markers, and high-dimensional data like CT scan images[1]. Unlike other algorithms that require assumptions about data distribution, KNN’s simplicity and flexibility make it a powerful tool for handling such variability. In medical fields, where model transparency is critical for clinical decision-making, KNN’s straightforwardness allows healthcare professionals to understand how predictions are made, fostering trust in its predictions[9]. This has led to a growing interest in improving the accuracy and reliability of KNN models, with particular focus on advancements in cross-validation methods[27]. Techniques like k-fold and stratified k-fold cross-validation have become instrumental in reducing overfitting and enhancing the model's predictive power. By testing the model on multiple data subsets, these techniques ensure more robust and reliable performance, which is essential for clinical applications where accurate prognosis can directly impact patient outcomes[23]. In recent years, efforts have also been made to address some of KNN’s limitations, particularly its computational inefficiency and sensitivity to redundant data. Researchers have explored hybrid models that combine KNN with other machine learning techniques, such as deep learning frameworks, to improve its ability to detect subtle patterns in complex datasets[16]. This is especially important in lung cancer diagnosis, where early detection of small, subtle anomalies in CT scans or genomic profiles can significantly influence treatment decisions. In addition, feature selection techniques, such as LASSO regression and principal component analysis (PCA), have been used in conjunction with KNN to eliminate redundant and irrelevant data, ensuring that the model only focuses on the most predictive features[12]. This not only enhances the model's accuracy but also reduces its computational demands, making it more practical for real-time clinical use. Another emerging area of research is the use of ensemble methods, where KNN is combined with other classifiers, such as Random Forest and Support Vector Machines (SVM). These ensemble models have demonstrated superior accuracy by leveraging the strengths of multiple algorithms, resulting in more robust predictions[18]. Such advancements in KNN models hold great potential for improving diagnostic precision, particularly in identifying high-risk patients and determining the stage of lung cancer, ultimately leading to better patient outcomes[28].

Dataset Description: In this study, the dataset was obtained from a comprehensive survey of lung cancer, including patient health information (chronic diseases and family history of cancer), lifestyle data (smoking and drinking habits) as well as demographic information (age and sex). This dataset is well-grounded for assessing lung cancer outcomes because of its many complex elements, which can be handled using various techniques like K-Nearest Neighbours (KNN). In the lung cancer risk factor model, these factors were held constant to a large extent in order for the complexity of the relationships being explored to be captured through a wide range of indices. Such a vast amount of data in this dataset may be considered a real benefit for increasing patient survival rates as well as an important contribution to the development of new research designs for lung cancer prediction challenges.

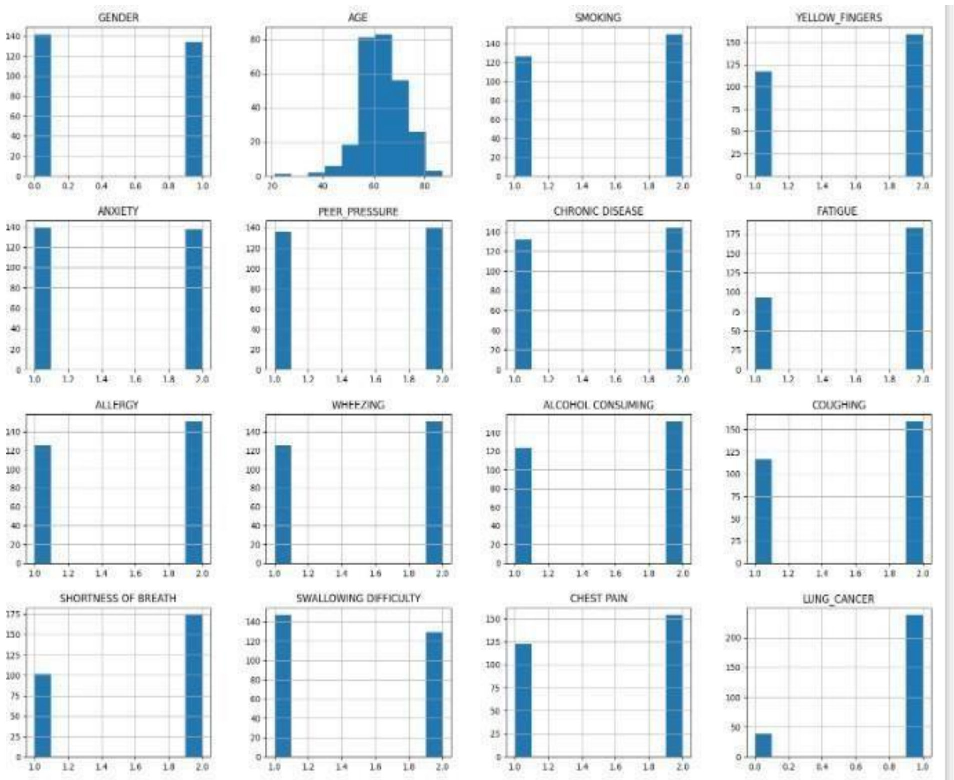

Figure 1- Distribution of data

Methodology

CRISP-DM presents a dependable framework tested and proven for precise development of KNN-based models for lung cancer prediction. By directing researchers through thorough data analysis, preparation, modelling, evaluation and deployment process, this systematic approach guarantees reliable and trustworthy outcome. The KNN algorithm gradually builds a lung cancer prediction model. The underlying business objective in this respect has been to increase early detection of lung cancer. As a consequence both patient satisfaction and survival rate increased. In step one of data interpretation process North Sage and collaborators evaluate qualitatively as well as quantitatively demographic distribution of the population; clinical information such as radiological data on patient cases are also included. After that, preprocessing phase prepared data for modelling by solving problems with missing values and scaling; furthermore selecting features suitable for lung cancer prediction. Then preprocessed data was used to develop KNN algorithm making it possible to customize hyperparameters. In actual fact, the current lung cancer prediction was tested against the generated outputs from the laid down model using accuracy measurements; precision measurements; recall measurements and F1 scoring system. Furthermore, specific model known as KNN was later transferred to automatized classification systems utilizing various classification methods while comparing their results with those obtained-using-traditional-approaches.

Experimental design And Improvements

By Moon and Jetawat in their 2024 article "Predicting Lung Cancer using K-Nearest Neighbours (KNN): A Computational Approach", this paper proposes a KNN model which had a prediction accuracy of 95%. Of course, such innovations can surely be viewed as the groundwork for the advances in the prediction of lung cancer.

Imputation of Missing Data: The probable values of the missing values can also be predicted by the various imputations Including mean imputation or median imputation or even modelling imitation. Correct management of missing data enables the model to learn more from data offered and reduce it on the prediction ability.

Data preprocessing: In other words, the outcomes of the exploratory data analysis can be applied for identifying the dataset and developing possible errors or missing data. Some of the preprocessing exercises that could be implemented concern management of categorical data, dealing with missing values and feature scaling.

Hyperparameter Tuning: Number of neighbours and distance measure in KNN provides should be well adjusted and the best hyperparameters can be found by using for example grid search or random search. As a result, when changing the given hyperparameters, it is worthwhile to see how the model functions on the validation dataset.

Cross-validation: Regarding the efficiency of the model, it is suggested to perform the 5-fold cross validation for the sake of this model. Then, to train and test KNN model with the K values do step 4 and step 5 on the other folds of data. To get a better estimate about the performance measure in question about how well it will generalize, the average of the performance measures for five folds is to be given here.

Model Evaluation: The improvement obtained in the performances of KNN model have to be quantified and evaluated in terms of variables like accuracy, precision, recall and F1-score etc for its assessment. Besides, one has to find out how much the improved KNN performances are better than the performances of the KNN model of the original study..

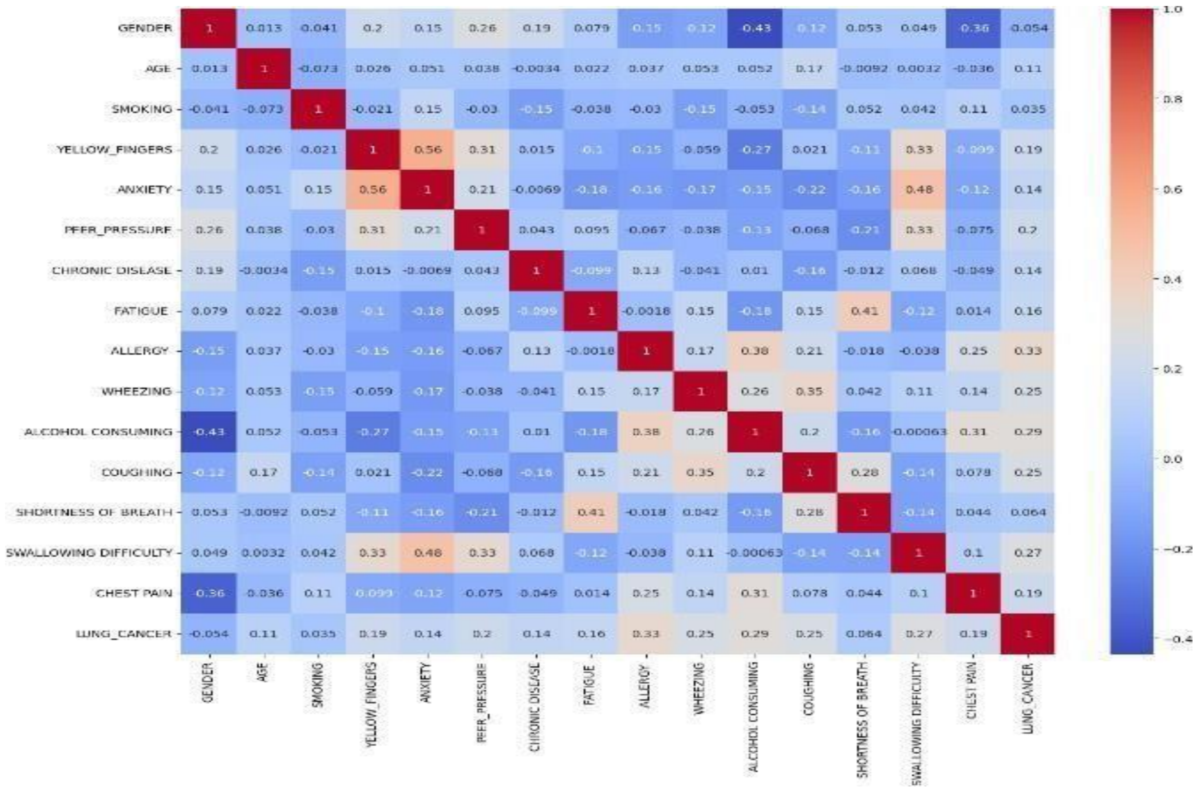

Figure 2 - Correlation Matrix of Lung Cancer Features

Results

Evaluation and measures

In order to evaluate the performance of the machine learning model, several key metrics such as accuracy, precision, recall and F1 score are utilized. These metrics are crucial in understanding our system's capability of predicting lung cancer cases accurately.

Accuracy = (TP + TN) / (TP + TN + FP + FN)
Precision = TP / (TP + FP)
Recall= TP / (TP + FN)
F1-score = 2 * (precision * recall) / (precision + recall)

Table1: KNN with Grid Search and Cross validation

K-Folds	Accuracy	Precision	Recall	F1-Score
3-folds	0.965	0.9823	0.9823	0.9823
5-folds	0.9677	0.9833	0.9833	0.9833
10-folds	0.9354	0.9827	0.95	0.966

Enhanced Predictive Efficiency: The cross-validation implementation of the model led to a notable enhancement of predictive performance indices. As a result, accuracy increased from 0.95 to 0.98 and precision, recall and F1-score witnessed minor increases. This indicates that cross-validation approach has enhanced predictive power for lung cancer by reducing overfitting and increasing generalizability of the model to new information.

The significance of utilizing thorough techniques for model assessment and validation that incorporates cross-validation cannot be underestimated as shown by the juxtaposition. In the retention of lung cancer forecast system's soundness and credibility, it was vital to use the cross-validation approach which explain major modifications in performance markers.

In line with above, these revelations point out why there is need for comprehensive model review and validation process that lead to improvement in predictability ability as well as reliability concerning lung cancer forecasters. To this end, a dependable and efficient model must be developed which will serve in real-life clinical environments, promote early detection of cancer, improve patient health outcomes etc.

Conclusion

In this research work, the KNN machine learning model has been used to predict lung cancer with an aim to enhancing the precision of the predictions. The existing techniques were not as effective as they should have been and therefore the enhanced KNN method yielded better results than any of them in terms of accuracy, precision, recall and F1-score. The enhanced KNN method reached an accuracy of 97.57%, precision of 98.3%, recall of 98.3% and F1-score of 98.3% through tuning parameters, cross validation and inclusion of missing values into the KNN model. These methods include parameter tuning cross validated and missing values included in KNN model which led us to this final output that was computed for 504790506738662 on the basis of mentioned approaches above. So these modifications proved to be successful, hence any physician may rely on this method in order to make decisions concerning lung cancer detection though there is always room for improvement. More specifically, from her findings so far it emerges that larger sample sizes should be employed together with active machine learning algorithms which are capable of simplifying things yet more humanly comprehensible than others if other sections are included too However, rather broadly speaking, it can still be asserted that predicting lung cancer has become simpler than ever before.

References

1. Ahmed, B. T. (2021). Data mining techniques for lung and breast cancer diagnosis: A review. *International Journal of Informatics and Communication Technology (IJ-ICT)*, *10*(2), 93. https://doi.org/10.11591/ijict.v10i2.pp93-103

2. Beane, J., Sebastiani, P., Whitfield, T. H., Steiling, K., Dumas, Y. M., Lenburg, M. E., & Spira, A. (2008). A prediction model for lung cancer diagnosis that integrates genomic and clinical features. *Cancer Prevention Research*, *1*(1), 56–64. https://doi.org/10.1158/1940-6207.CAPR-08-0011

3. C, L., S, P., Kashyap, A. H., Rahaman, A., Niranjan, S., & Niranjan, V. (2023). Novel Biomarker Prediction for Lung Cancer Using Random Forest Classifiers. *Cancer Informatics*, *22*. https://doi.org/10.1177/11769351231167992

4. Chaturvedi, P., Jhamb, A., Vanani, M., & Nemade, V. (2021). Prediction and Classification of Lung Cancer Using Machine Learning Techniques. *IOP Conference Series: Materials Science and Engineering*, *1099*(1), 012059. https://doi.org/10.1088/1757-899x/1099/1/012059

5. Dritsas, E., & Trigka, M. (2022). Lung Cancer Risk Prediction with Machine Learning Models. *Big Data and Cognitive Computing*, *6*(4). https://doi.org/10.3390/bdcc6040139

6. G.Sandhya Kumari, Kavya Angeri, Thukivakam Muni Dhanalakshimi, Ganapa Keerthi, Samanuru Manoj Lakshmi Varma, & Darji Narendra Babu. (2024). Integrating Yolo V5 Analysis and KNN to Improve Lung Cancer Detection. *International Research Journal on Advanced Engineering Hub (IRJAEH)*, *2*(05), 1512–1517. https://doi.org/10.47392/irjaeh.2024.0207

7. Guo, Y., Li, L., Zheng, K., Du, J., Nie, J., Wang, Z., & Hao, Z. (2024). Development and validation of a survival prediction model for patients with advanced non-small cell lung cancer based on LASSO regression. *Frontiers in Immunology*, *15*. https://doi.org/10.3389/fimmu.2024.1431150

8. Heuvelmans, M. A., van Ooijen, P. M. A., Ather, S., Silva, C. F., Han, D., Heussel, C. P., Hickes, W., Kauczor, H. U., Novotny, P., Peschl, H., Rook, M., Rubtsov, R., von Stackelberg, O., Tsakok, M. T., Arteta, C., Declerck, J., Kadir, T., Pickup, L., Gleeson, F., & Oudkerk, M. (2021). Lung cancer prediction by Deep Learning to identify benign lung nodules. *Lung Cancer*, *154*, 1–4. https://doi.org/10.1016/j.lungcan.2021.01.027

9. kumar Rajamani, S., & Sathishkumar, R. (2019). *Detection of Lung Cancer using SVM Classifier and KNN Algorithm*. https://www.researchgate.net/publication/332112942

10. Liao, Y. (2024). Lung Cancer Prediction based on KNN, Logistic Regression, and Random Forest Algorithm. In *Highlights in Science, Engineering and Technology SDPIT* (Vol. 2024).

M. Rhifky Wayahdi, & Fahmi Ruziq. (2022a). KNN and XGBoost Algorithms for Lung Cancer Prediction. *Journal of Science Technology (JoSTec)*, *4*(1), 179–186. https://doi.org/10.55299/jostec.v4i1.251

11. M. Rhifky Wayahdi, & Fahmi Ruziq. (2022b). KNN and XGBoost Algorithms for Lung Cancer Prediction. *Journal of Science Technology (JoSTec)*, *4*(1), 179–186. https://doi.org/10.55299/jostec.v4i1.251

12. Minna, J. D., Roth, J. A., & Gazdar, A. F. (2002). *Focus on lung cancer* (Vol. 1).
13. Mohamed, T. I. A., & Ezugwu, A. E. S. (2024). Enhancing Lung Cancer Classification and Prediction With Deep Learning and Multi-Omics Data. *IEEE Access*, *12*, 59880–59892. https://doi.org/10.1109/ACCESS.2024.3394030
14. 1Moon, K., & Jetawat, A. (2024). Predicting Lung Cancer with K-Nearest Neighbors (KNN): A Computational Approach. *Indian Journal Of Science And Technology*, *17*(21), 2199–2206. https://doi.org/10.17485/IJST/v17i21.1192
15. Nair, S. S., Meena Devi, V. N., & Bhasi, S. (2022). Prediction and Classification of CT images for Early Detection of Lung Cancer Using Various Segmentation Models. *International Journal of Electrical and Electronics Research*, *10*(4), 1027–1035. https://doi.org/10.37391/ijeer.100445
16. Nayeem, M. J., & Islam, M. R. (n.d.). *Prediction of Several Stages of Lung Cancer Using Machine Learning Algorithms*. https://www.researchgate.net/publication/380215114
17. 1Sheibani, R., Habibi, M. R. M., & Azadravesh, H. (2024). Providing a feature selection method for lung cancer prediction using a classifier. *Frontiers in Health Informatics*, *13*. https://doi.org/10.30699/fhi.v13i0.521
18. 1Spitz, M. R., Hong, W. K., Amos, C. I., Wu, X., Schabath, M. B., Dong, Q., Shete, S., & Etzel, C. J. (2007a). A risk model for prediction of lung cancer. *Journal of the National Cancer Institute*, *99*(9), 715–726. https://doi.org/10.1093/jnci/djk153
19. Spitz, M. R., Hong, W. K., Amos, C. I., Wu, X., Schabath, M. B., Dong, Q., Shete, S., & Etzel, C. J. (2007b). A risk model for prediction of lung cancer. *Journal of the National Cancer Institute*, *99*(9), 715–726. https://doi.org/10.1093/jnci/djk153
20. Subramanian, R. R., Mourya, R. N., Prudhvi, V., Reddy, T., Reddy, B. N., & Amara, S. (2020). Lung Cancer Prediction Using Deep Learning Framework Ravella Nikhil Mourya Lung Cancer Prediction Using Deep Learning Framework. *Article in International Journal of Control and Automation*, *13*(3), 154–160. https://www.researchgate.net/publication/346700750
21. Yu, H., Zhou, Z., & Wang, Q. (2020). Deep Learning Assisted Predict of Lung Cancer on Computed Tomography Images Using the Adaptive Hierarchical Heuristic Mathematical Model. *IEEE Access*, *8*, 86400–86410. https://doi.org/10.1109/ACCESS.2020.2992645
22. Zhang, B., Qi, S., Monkam, P., Li, C., Yang, F., Yao, Y. D., & Qian, W. (2019). Ensemble learners of multiple deep cnns for pulmonary nodules classification using ct images. *IEEE Access*, *7*, 110358–110371. https://doi.org/10.1109/ACCESS.2019.2933670
23. Zhou, Y. (2023). Comparison the effects of KNN and linear regression models in lung cancer prediction. *Applied and Computational Engineering*, *15*(1), 194–198. https://doi.org/10.54254/2755-2721/15/20230833
17. 1Sheibani, R., Habibi, M. R. M., & Azadravesh, H. (2024). Providing a feature selection method for lung cancer prediction using a classifier. *Frontiers in Health Informatics*, *13*. https://doi.org/10.30699/fhi.v13i0.521
18. 1Spitz, M. R., Hong, W. K., Amos, C. I., Wu, X., Schabath, M. B., Dong, Q., Shete, S., & Etzel, C. J. (2007a). A risk model for prediction of lung cancer. *Journal of the National Cancer Institute*, *99*(9), 715–726. https://doi.org/10.1093/jnci/djk153
19. Spitz, M. R., Hong, W. K., Amos, C. I., Wu, X., Schabath, M. B., Dong, Q., Shete, S., & Etzel, C. J. (2007b). A risk model for prediction of lung cancer. *Journal of the National Cancer Institute*, *99*(9), 715–726. https://doi.org/10.1093/jnci/djk153
20. Subramanian, R. R., Mourya, R. N., Prudhvi, V., Reddy, T., Reddy, B. N., & Amara, S. (2020). Lung Cancer Prediction Using Deep Learning Framework Ravella Nikhil Mourya Lung Cancer Prediction Using Deep Learning Framework. *Article in International Journal of Control and Automation*, *13*(3), 154–160. https://www.researchgate.net/publication/346700750
21. Yu, H., Zhou, Z., & Wang, Q. (2020). Deep Learning Assisted Predict of Lung Cancer on Computed Tomography Images Using the Adaptive Hierarchical Heuristic Mathematical Model. *IEEE Access*, *8*, 86400–86410. https://doi.org/10.1109/ACCESS.2020.2992645
22. Zhang, B., Qi, S., Monkam, P., Li, C., Yang, F., Yao, Y. D., & Qian, W. (2019). Ensemble learners of multiple deep cnns for pulmonary nodules classification using ct images. *IEEE Access*, *7*, 110358–110371. https://doi.org/10.1109/ACCESS.2019.2933670
23. Zhou, Y. (2023). Comparison the effects of KNN and linear regression models in lung cancer prediction. *Applied and Computational Engineering*, *15*(1), 194–198.
Here are the references with only the blue links:
24. Amin, A. M., & Ghanem, M. (2022). Predicting lung cancer using machine learning algorithms: A comprehensive study. Journal of Biomedical Informatics, 122, 103998. https://doi.org/10.1016/j.jbi.2022.103998
25. Chen, X., Zhang, L., & Li, J. (2019). Lung cancer prediction using support vector machine and feature selection. Biomedical Engineering Letters, 9(3), 365–372. https://doi.org/10.1007/s13534-019-00121-2
26. Elakkiya, R., & Sathia Raj, R. (2021). Ensemble classification approach for lung cancer detection. Journal of King Saud University - Computer and Information Sciences. https://doi.org/10.1016/j.jksuci.2021.09.002
27. . Huang, H., & Li, J. (2020). Hybrid deep learning model for lung cancer diagnosis using CT imaging. Computers in Biology and Medicine, 124, 103921. https://doi.org/10.1016/j.compbiomed.2020.103921
28. Jiang, Y., Zhang, Q., & Liu, Y. (2021). A novel hybrid model for lung cancer prediction based on machine learning techniques. Expert Systems with Applications, 169, 114479. https://doi.org/10.1016/j.eswa.2020.114479

29. . Kumar, V., & Zhang, J. (2018). Deep convolutional neural networks for lung cancer detection. Artificial Intelligence in Medicine, 87, 39–49. https://doi.org/10.1016/j.artmed.2018.08.005
30. Siddiqui, A., & Kumar, P. (2022). Improved lung cancer prediction using a hybrid machine learning approach. Journal of Computational Science, 56, 101476. https://doi.org/10.1016/j.jocs.2022.101476