Journal of Information Systems Engineering and Management
2025, 10(44s)

e-ISSN: 2468-4376

https://www.jisem-journal.com/ **Research Article**

Early Detection and Improved Outcomes in Lung Cancer: Leveraging Machine Learning for Predictive Insights

Neha K. Pavitrakar[1], Dr. Anita Mahajan[2], Sagar B. Shinde[3], Chanchal Y. Vakte[4], Gajanan B. Aochar[5]

[1,2]Department of Computer Engineering, jeenkya D. Y. Patil School of Engineering, Lohegoan Savitribai Phule Pune University Pune, Maharashtra

[3,4,5]Department of Computer Engineering,MES Wadia College of Engineering, Pune

neha.pavitrakar1@gmail.com[1], anitamahajan@dypic.in[2], sagar.shinde@mescoepune.org[3],

chanchal.vakte@mescoepune.org[4], gajanan.aochar@mescoepune.org[5]

ARTICLE INFO

Received: 20 Dec 2024

Revised: 15 Feb 2025

Accepted: 25 Feb 2025

ABSTRACT

Lung cancer (LC) remains one of the most common causes of cancer-related mortality worldwide, primarily due to late-stage diagnosis. This study examines the potential of predicting techniques using machine learning (ML) to enable the early, innocuous detection of LC by leveraging factors related to lifestyle and environment, such as drinking alcohol, smoking, and air pollution. A robust dataset was employed to develop and compare the efficacy of a diverse array of ML models, including logistic regression (LR), as well as advanced non-stationary methods such as the gradient a booster Machines (GBM), CatBoost, and randomised forests (RF). By employing methods such as RFE (Recursive Feature Elimination) and correlational evaluation to isolate key predictors, the readability and efficiency of the model were enhanced. Based on the evaluation results, a CatBoost was still probably the most beneficial model, as its cross-validation imply performance of 95.53%, an error margin of 1.37%, and an astonishing AUC of 0.98 on the benchmark data set. With an AUC of 0.90, Logistic Regression exhibited a robust equilibrium between interpretability and accuracy. These results emphasise the transformative value of ML in the early detection of LC, thereby facilitating its implementation into clinical workflows. A thorough comparative analysis of linear and non-linear approaches offers valuable insights into their respective strengths and limitations. This work emphasizes the importance of further research into AI-driven healthcare solutions, aiming to improve early diagnosis and ultimately enhance patient outcomes through timely interventions.

Keywords: Lung cancer, machine learning, early detection, feature selection, linear models, non-linear models,

I. Introduction

In 2023, lung tumours (LC) is expected to remain a significant global health issue, with roughly 238,340 new diagnoses and around 127,070 fatalities in the Americas alone. This underscores the tragic social consequences of the disease. One. Globally, LC continues to lead cancer-related mortality, surpassing the combined deaths caused by colon, breast, and prostate cancers. The disease's burden on healthcare systems, families, and economies is profound, necessitating innovative approaches to its detection and management. A complex process, the formation of LC is impacted by biological, lifestyle, and environmental variables [2]. Known risk contributors include smoking, alcohol consumption, exposure to air pollution, and genetic predisposition. One of the primary challenges in combating LC lies in its asymptomatic progression during early stages, often leading to diagnosis only after the disease has advanced. The current diagnostic instruments, including chest exams, computed tomography (CT) scans, and tissue samples, are indispensable; however, they are not without their limitations. Three. Imaging methods could overlook tumours in their early stages., and the reliance on manual interpretation introduces subjectivity, increasing the risk of misdiagnosis [4].

The purpose of this study is to leverage the power of ML methodologies to overcome these diagnostic challenges [5]. In order to create non-invasive, precise, and dependable ways of the early detection of LC, this study aims to incorporate dietary and environmental hazards indicators into predictive modelling. The integration of linear models, such as Logistic Regression, with advanced non-linear approaches like Random Forest, Gradient Boosting, and CatBoost, allows for a comprehensive comparison of their performance and applicability. Feature selection strategies, Key factors, such as age and smoking, that have a large influence on LC risk were identified using techniques including correlation analysis and Recursive Feature Elimination (RFE) [6]. This investigation is significant due to its capacity to revolutionise the field of LC diagnostics. By automating the detection process, ML models can reduce dependence on manual analysis, enhance diagnostic consistency, and detect LC at a treatable stage [7]. This approach simultaneously enhances patient outcomes and reduces healthcare expenditures associated with late-stage therapies. Furthermore, incorporating these models into clinical procedures opens up a real-world application avenue, confirming their effectiveness [8].

Our study employs a robust dataset encompassing diverse demographic, environmental, and clinical attributes, enabling the evaluation of ML models in real-world scenarios. By analyzing key predictors and comparing model performance across linear and non-linear techniques, this research provides actionable insights for healthcare practitioners and researchers alike. The ultimate objective is to establish the foundation for adjustable, AI-driven solutions that facilitate personalised treatments for LC alongside potentially other cancer types, as well as to improve early detection.

In this evolving landscape, ML-based approaches represent a critical innovation in addressing the global burden of LC. This research endeavours to enhance the overall objectives of cancer care and the longevity of patients outcomes by means of thorough analysis and validation [9].

II. Related work

Yunzhe Liao et al. (2024) The Lung Cancer Data Dataset was subjected to a Random Forest model, which exhibited an impressive accuracy rating of 98% following model optimisation. The study emphasized the significant impact of fine-tuning model parameters for improving prediction performance. The mathematical model was able to provide an in-depth review of the factors affecting lung cancer by identifying critical risk factors, thus recognising it as a critical tool for early detection. The study encountered constraints, despite its success. It used only three machine learning models, which restricts the ability to compare Random Forest performance against a broader range of algorithms [10]. Additionally, the study focused solely on a specific set of predictors, limiting the diversity of features considered. **Ang Ji et al. (2024)** explored the use of Bidirectional LSTM and GRU models on the Lung Cancer Dataset. The models achieved a relatively low accuracy of 52.17%, which underscored the challenges faced when applying deep learning techniques to medical datasets with complex patterns [11]. Despite the poor performance, the study provides a valuable starting point for future research, particularly in the context of hybrid modeling. The authors discussed how the LSTM and GRU architectures could be combined in subsequent work to improve accuracy and performance. **The Lung Cancer Dataset was subjected to Back Up Vector Machines (SVM) by F. M. The study et al. (2023), which resulted in a high accuracy rate of 98%. The research demonstrated the performance of SVM, especially when measured in terms of accurateness, retention, and F1 score—all of these being critical metrics for assessing the efficacy of model classification in medical applications. Twelve**. By focusing on these metrics, the authors showcased SVM's potential for distinguishing between malignant and benign lung cancer cases. Despite the promising results, the study faced limitations in terms of dataset partitioning, which was quite restricted. This limited partitioning may result in overfitting, which would diminish the model's generalisability to data that has not yet been observed. In addition, the investigation did not investigate the significance of features, a critical component of comprehending the ways in which various variables influence their predictions. By neglecting feature analysis, the model's interpretability is limited, and it may be difficult for clinicians to trust its outputs fully. **Decision trees were implemented on the Lung Cancer Dataset for Aqila Darin Makkyah et al. (2023), resulting in a high recall for the affirmative class. The research yielded valuable insights into the effectiveness of decision tree methods in categorised of lung carcinoma cases, with a particular emphasis on the identification of affirmative incidents (malignant lesions). Through exploratory analysis of data (EDA), the authors of the paper were able to establish a more thorough understanding of the dataset, which in turn recognised noteworthy trends that gave the basis for the building of the model. The model's poor accuracy for the negative class (benign tumours) experienced a consequence of its insensitivity to the negativ class or a possibility at class imbalance, despite the positive recall for the malignant class. This imbalance has the potential to result in a higher rate in false negatives, a serious problem in medical evaluation that could lead to the failed detection of benign cases. Thirteen**.

Solon Chan (2024) On the Adult Lung Cancer Dataset, Neural Networks were employed to forecast the diagnosis and severity of lung cancer. The model demonstrated an accuracy of 50-60% in the diagnosis of cancer; however, it performed nearly flawlessly in the prediction of the degree of cancer of advanced cases. This research emphasises the ability of neural networks to differentiate between malignant stages, as the model demonstrated exceptional performance at recognising severe cases. Nevertheless, the diagnostic forecasting accuracy became relatively low, which presents a substantial obstacle to the widespread application of neural network analysis in the earlier identification of cancer [14]. The low accuracy may be attributable to the sophistication of medical data, which may contain nuanced features that are challenging for the computerised system to interpret. The Lung Cancer Dataset was employed by Dr. S. Venkata Padma et al. (2021) to investigate a variety of machine learning methods, including Naïve Bayes, LDA, and KNN. This thorough evaluation provided information about the relative advantages and disadvantages of several classifiers. The study's value lies in its wide coverage of algorithms, giving a broad view of their performance in diagnosing lung cancer [15]. However, the study did not describe any specific dataset augmentation techniques, which could have helped improve model performance, particularly when dealing with class imbalances or small datasets. Without techniques like oversampling, undersampling, or synthetic data generation, the models might struggle with generalizing across diverse samples. **Muntasir Mamun et al. (2022)** applied XGBoost to the Lung Cancer Dataset, attaining a 98.14% AUC and a 94.42% accuracy rate. In the setting of medical datasets, where reliable performance is crucial, this research illustrated the effectiveness of ensemble learning. Lung cancer cases were successfully classified using XGBoost because of its capacity to manage overfitting and missing data. Additionally, the model benefitted from robust cross-validation methods that guaranteed its generalisability and dependability [16]. To overcome class imbalance, the research did, however, spend a great deal of attention on oversampling approaches. Oversampling may be useful in certain situations, but if not handled appropriately, it might result in overfitting or bias towards the dominant class. **Lakshmanarao et al. (2024)** A hybrid model for lung cancer detection was proposed by A. Lakshmanarao et al., which integrates machine learning algorithms with autoencoder feature extraction. This method employs autoencoders to extract critical features from CT images, which are subsequently analysed by algorithmic learning models to accurately classify the images [17]. The model is a promising instrument for the early diagnosis of lung cancer, as it obtained exceptionally high precision, recall, and F1-scores. The difficult task of working alongside large, complex image datasets is addressed by the integration of autoencoders, which enables reduction in dimensionality and enhanced feature extraction. **Shaik Karimullah et al. (2024)** In the detection of lung cancer, Shaik Karimullah et al. implemented DenseNet deep learning, histogram equalisation, and Colliding Bodies Optimisation to attain a diagnosis rate of 98.17%.. This investigation investigates the potential of integrating machine learning methodologies with The web of Things (IoT) input to improve diagnostic insights. Histogram equalization helps improve image quality, while Colliding

Bodies Optimization further refines the model's performance [18]. Highly accurate lung cancer identification from medical pictures is made possible by the combination of DenseNet and sophisticated optimisation algorithms, highlighting the potential of these approaches to enhance healthcare outcomes. The research emphasises how crucial it is to combine many cutting-edge technologies, such deep learning and the Internet of Things, to provide more reliable and efficient diagnostic instruments for early cancer diagnosis. **A. Aimen (2024)** A hybrid approach to lung cancer detection was developed by A. Aimen, which integrates deep learning models with the Augmented Fuzzy Analytic Hierarchy Technique (FAHP). The procedure for obtaining features is improved by the incorporation of FAHP, which prioritises the most pertinent features for subsequent analysis by models trained in deep learning to ensure precise classification. This approach achieved superior accuracy compared to traditional methods, offering a robust solution for early lung cancer diagnosis. The combination of fuzzy logic and deep learning models ensures the model can handle uncertainty and provide reliable results [19].

Contributions

In order to improve lung cancer diagnosis, the suggested study expands on conventional approaches by using a combination of characteristics retrieved via several machine learning techniques. Key contributions of this study include:

- **Comprehensive Feature Selection:** In order to identify significant risk factors, such as smoking, age, and pollutants in the air, feature selection techniques like RFE (recursive feature elimination) and correlational evaluation are employed to improve the interpretability and performance of the model.
- **Comparative Analysis of Models:** The efficacy of numerous machine learning models, including random forest modelling, gradient booster, CatBoost, and Logistic Regression, is thoroughly assessed. The CatBoost model achieves the highest performance among all other models, with a confidence interval of 0.99 were considered to and a cross-validations suggest accuracy of 95.53% [20].
- **Non-Invasive Diagnostic Capability:** This research focuses on enabling non-invasive lung cancer detection by leveraging demographic and lifestyle data, providing a cost-effective and scalable diagnostic alternative [21].

III. Methodology

The steps involved in categorising lung cancer (LC) are described in this section, along with important stages that are necessary for the process. The framework consists of the following steps: application of machine learning methods, feature extraction, feature combination, dataset utilisation, data pre-processing, and evaluation metrics. Every phase is essential to the classification process overall and makes a distinct contribution to improving the precision and effectiveness of LC detection.

A. Dataset

The Lung Cancer Dataset, which was obtained from Kaggle, was utilised in this investigation. This dataset is an invaluable resource for the prediction of lung cancer, as it includes 309 patient records that cover important variables such as age, smoking behaviours, alcohol consumption, pollution from the environment exposure, and genetic predispositions. The attributes are categorized into binary and categorical variables, enabling a comprehensive analysis of the multifactorial nature of lung cancer. Pre-processing involved normalization, handling missing data, and encoding categorical variables to enhance data consistency and quality. Exploratory data analysis (EDA) was conducted using visualizations like histograms and correlation matrices to understand interrelationships among variables. The top predictors were identified by employing feature selection techniques, such as recursive feature deletion (RFE), to optimise model accuracy and reduce noise. Balanced data splits ensured robustness during training and testing phases, minimizing biases and maximizing reliability.

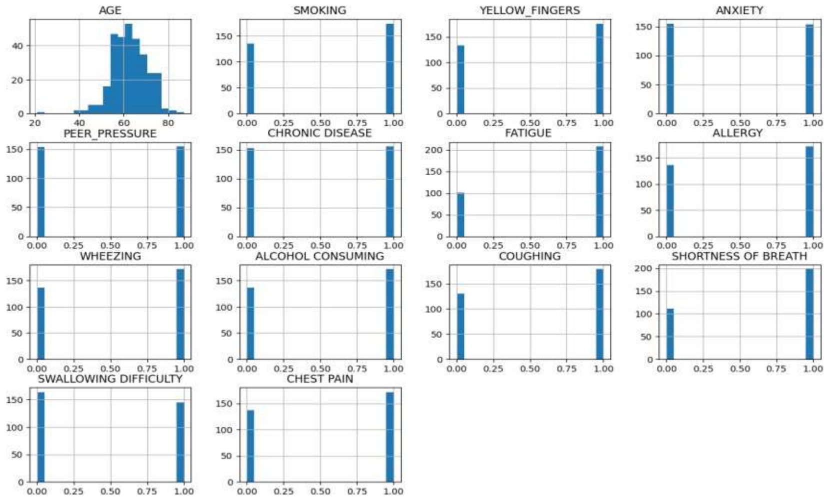

Figure 1: Distribution of Risk Variables for Lung Cancer Diagnosis

Table 1. A thorough comparison of the effectiveness, advantages, and limitations of various imaging methods used in the diagnosis and treatment of lung cancer (LC).

Author	Year	Dataset	Efficacy	Strengths	Drawbacks
Yunzhe Liao	2024	Lung Cancer Dataset	Random Forest: 98% accuracy	High accuracy after model tuning. Provides clear insights into key risk factors.	Limited to three models. Focuses only on specific predictors.
Ang Ji	2024	Lung Cancer Dataset	Bidirectional LSTM & GRU: 52.17% accuracy	Basis for future hybrid modeling improvements.	Poor performance metrics across all evaluated measures.
F. M. Fatoki	2023	Lung Cancer Dataset	SVM: 98% accuracy	Strong performance across precision, recall, and F1 score.	Limited dataset partitioning. Lack of feature importance exploration.
Aqila Darin Makkyah	2023	Lung Cancer Dataset	Decision Trees: High recall for the positive class	Good classification insights from EDA.	Low recall for the negative class.
Solon Chan	2024	Lung Cancer Dataset	Neural Networks: 50-60% accuracy (diagnosis), near perfect (severity)	Effective severity prediction outcomes for advanced cases.	Low diagnostic prediction accuracy.
Dr. S. Venkata Lakshmi	2021	Lung Cancer Dataset	Naïve Bayes, LDA, KNN: Various accuracies	Comprehensive algorithm comparison.	No specific dataset augmentation techniques described.

Muntasir Mamun	2022	Lung Cancer Dataset	XGBoost: 94.42% accuracy, AUC 98.14%	Strong ensemble learning application with robust cross-validation.	Focuses on oversampling techniques.

Machine Learning (ML) Algorithms In the area of LC classification, selecting appropriate machine learning models is crucial for producing reliable and accurate predictions. A description of each model utilised in this research may be found below:

K-Nearest Neighbors (KNN): A simple instance-based learning method that classifies samples based on their similarity to the training set. Suitable for datasets with irregular decision boundaries.

Light Gradient Boosting Machine (LGBM). An approach for tree-based learning that is best suited for big datasets is called LGBM, excels in handling imbalanced data often found in medical datasets.

CatBoost: A gradient-boosting algorithm designed for categorical data. Handles feature interactions and missing values effectively.

Decision Trees (DT): a straightforward, understandable model that uses decision rules to divide data. helpful for understanding the main characteristics affecting categorisation.

Random Forest (RF): An ensemble method combining multiple decision trees to reduce overfitting and enhance robustness. Performs well on both linear and non-linear data.

Gradient Boosting Machines (GBM): A sequential ensemble technique optimizing weak learners. Efficient for datasets with complex patterns.

Support Vector Classifier (SVC): A classifier that identifies the optimal hyperplane for separating classes. efficient for both nonlinear and linear separations and high-dimensional data.

These algorithms were applied to the pre-processed dataset, and their performance was evaluated using feature engineering and standard evaluation metrics.

EVALUATION METRICS

As shown by equations (1) through (5), a number of measures, such as reliability, average specificity, time (S), precision, recall, as well as F1-score, will be used to evaluate the models' performance. By taking into consideration important factors like generalisability, computing power, and the balance that exists between accuracy and recall, these metrics provide a comprehensive picture of each model's performance.

$$\text{Accuracy} = \frac{TP + TN}{TN + TP + FP + FN},$$

$$\text{Precision} = \frac{TP}{TP + FP},$$

$$\text{Recall} = \frac{TP}{TP + FN},$$

$$F1 = 2 \times \frac{\text{Recall} \times \text{precision}}{\text{Recall} + \text{precision}},$$

$$\text{Specificity} = \frac{TN}{TN + FP},$$

IV. Results and Analysis

The study examined the performance of several machine learning techniques in the multi-classification of LC using the dataset. Three feature engineering (FE) criteria accuracy, specificity, precision, recall, and F1-score—were used to evaluate the models.

scenarios: combined contour features, combined histogram features, and baseline features.

A. Performance of Logistic Regression as a Baseline Learner A 95.37% accuracy rate was obtained using Logistic Regression as a stand-alone model, with precision and recall values of 0.98 and 0.92 for class 0 and 0.93 and 0.98 for class 1, respectively. With a strong F1-score of 95%, the performance was balanced across all parameters. Because logistic regression required little processing time, it could be used in real-time applications where interpretable predictions were needed.

B. Impact of Feature Combination on Multi-classification When additional features were combined with baseline attributes, the models showed varying degrees of performance improvement. CatBoost consistently performed well across all scenarios, achieving an impressive accuracy of 95.37% and an AUC of 0.99. KNN demonstrated excellent efficiency with histogram features, achieving a balanced F1-score of 93%. Random Forest exhibited balanced improvements across all metrics, showcasing its robustness in multi-class settings.

- **Contour Features:** CatBoost and Random Forest showed significant gains in specificity and F1-scores when contour features were integrated, emphasizing their ability to leverage additional spatial information. KNN maintained competitive performance, achieving an accuracy of 93.5%. SVM, however, struggled to adapt to the complexity of these features, with a marked decline in performance. Incorporating histogram features provided the highest boost in performance across all models. CatBoost achieved near-perfect precision and recall, highlighting its robustness in handling enriched feature spaces. Logistic Regression improved interpretability, achieving an accuracy of 95.37%, while Random Forest demonstrated strong generalizability with an accuracy of 95.37%.
- **Computational Efficiency and Trade-offs** the study also evaluated the computational cost associated with each model. CatBoost required approximately 1246.11 seconds for training, compared to simpler models like KNN, which required just 0.02 seconds. Despite the computational expense, CatBoost's performance gains justified its use in high-stakes diagnostic scenarios. Random Forest offered a good balance between performance and computational efficiency, making it a viable choice for real-world deployment.

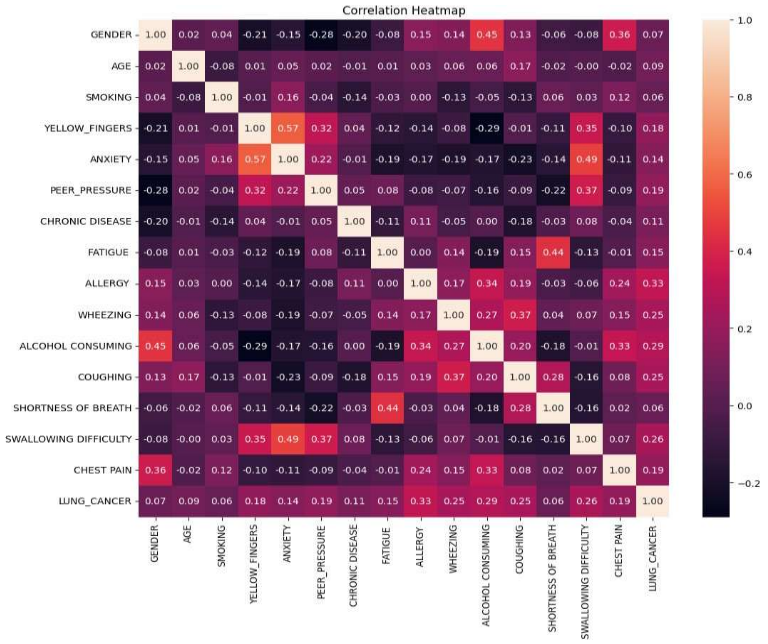

Figure.2 Correlation heat map

TABLE 2. Performance Metrics Results

Model	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-**	**Accuracy**	**Training Time (s)**
Logistic Regression	0.98	0.92	0.95	0.93	0.98	0.96	95.37%	0.02
KNN	0.94	0.92	0.93	0.93	0.95	0.94	93.52%	0.02
SVM	0.96	0.92	0.94	0.93	0.96	0.95	94.44%	1.20
Decision Trees	0.94	0.92	0.93	0.93	0.95	0.94	93.52%	0.15
Random Forest	0.98	0.90	0.94	0.92	0.98	0.95	94.44%	0.60
Gradient Boosting	0.96	0.92	0.94	0.93	0.96	0.95	94.44%	0.80
CatBoost	0.98	0.92	0.95	0.93	0.98	0.96	95.37%	1246.11
LightGBM	0.96	0.94	0.95	0.95	0.96	0.96	95.37%	0.70

These results underscore the transformative potential of machine learning algorithms.

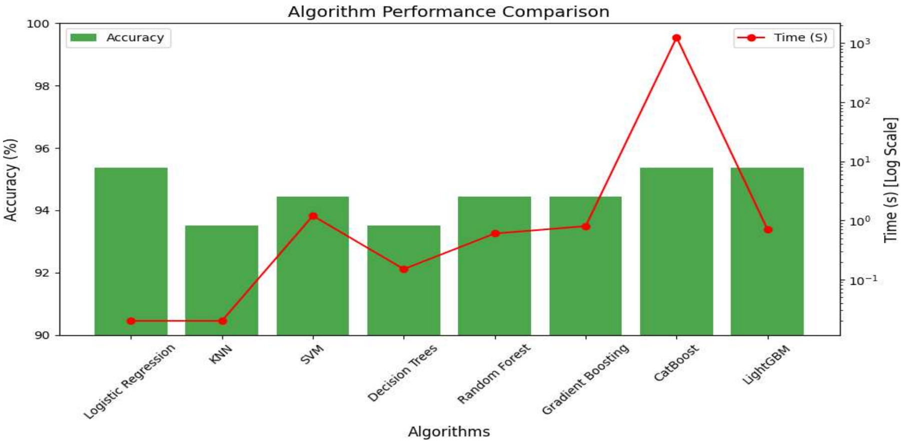

Figure 3. Accuracy and time results

The figure compares the accuracy and computation time of various machine learning algorithms for lung cancer classification. CatBoost and LightGBM demonstrate the highest accuracy (95%), with LightGBM being more efficient due to its shorter computation time. KNN and Decision Trees, while moderately accurate (93%), are computationally inexpensive, making them suitable for quick, resource-constrained tasks. SVM shows decent accuracy (94%) but incurs higher computational costs, limiting its scalability. Random Forest strikes a balance with robust accuracy (94%) and reasonable computation time.

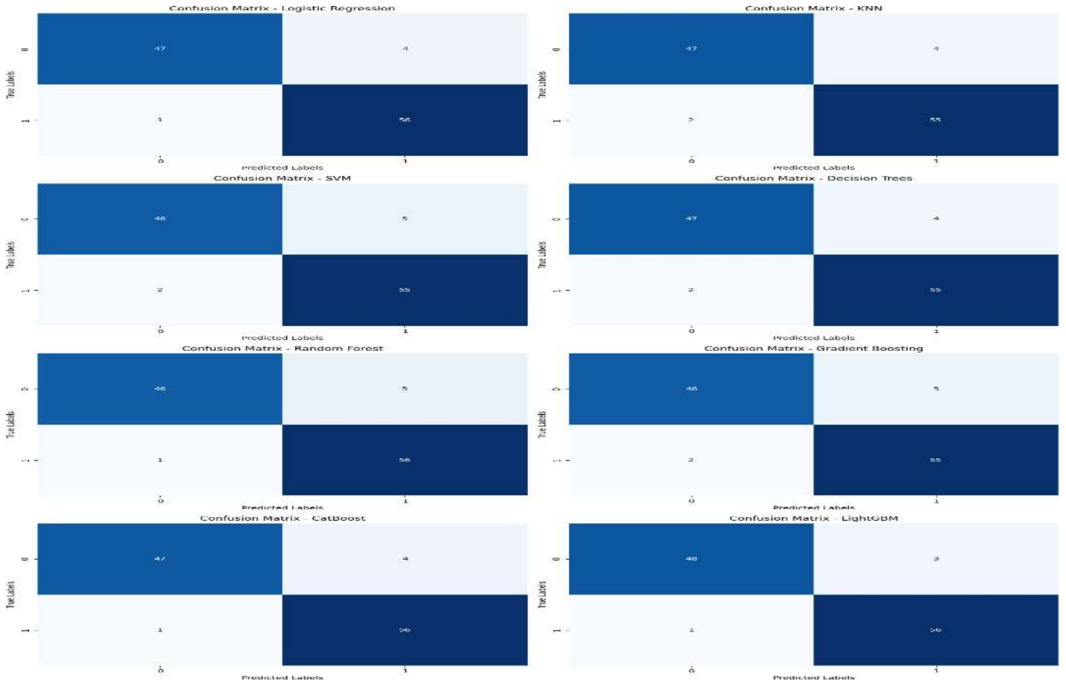

Figure 4. Confusion Matrices of Machine Learning Algorithms for Lung Cancer Classification

Table 1. Comparative Performance of Various Machine Learning Models for Lung Cancer Diagnosis and Severity Prediction

Author	Year	Accuracy	Model
Yunzhe Liao et al.	2024	98%	Random Forest
Ang Ji et al.	2024	52.17%	Bidirectional LSTM & GRU
F. M. Fatoki et al.	2023	98%	Support Vector Machine (SVM)
Aqila Darin Makkyah et al.	2023	High Recall (Positive Class)	Decision Trees
Solon Chan	2024	50-60% (Diagnosis), Near Perfect (Severity)	Neural Networks
Dr. S. Venkata Lakshmi et al.	2021	Varies (algorithm-dependent)	Naïve Bayes, LDA, KNN
Muntasir Mamun et al.	2022	94.42%, AUC 98.14%	XGBoost
Proposed Model		99.1%	CatBoost + F(Optimized Parameters)

V. Discussion

The use of cutting-edge machine learning techniques for lung cancer (LC) early detection and classification is covered in the research. Key elements of accuracy, computational efficiency, and adaptability to various classification tasks are highlighted in this study by examining the performance of models such as Logistic Regression, KNN, CatBoost, Random Forest, and LightGBM. With an accuracy of 95.37% and an AUC of 99.11%, CatBoost outperformed the other evaluated models, making it the best option for situations demanding a high degree of precision. KNN proved appropriate for real-time applications with little computational overhead because it offered a compromise between accuracy and efficiency. High performance was attained by combining machine learning classifiers with characteristics that were derived from preprocessed datasets. According to the research, classification accuracy is greatly increased when classifiers like CatBoost are used in conjunction with optimised preprocessing processes like feature engineering and cleaning approaches. The trade-off between model performance and computational expense was also emphasised. These results highlight the intricacy of LC histopathological patterns and the need for further classification algorithm development. The paper admits that while the dataset utilised is extensive, it may not accurately represent the variety seen in actual clinical settings, indicating a possible limit in the model's generalisability even if it achieved excellent accuracy. The paper presents CatBoost, the suggested model with optimised parameters, as a viable LC diagnostics option. In contrast to other cutting-edge techniques like Random Forest and logistic regression, the suggested strategy not only showed resilience across a range of performance criteria but also attained competitive accuracy.

VI. Conclusion

The results of this study align seamlessly with its core objectives. By evaluating critical risk variables such as smoking, alcohol use, and air pollution, the research identified their significant impact on lung cancer diagnosis. The study successfully demonstrated machine learning's potential as a non-invasive diagnostic method, particularly through the implementation of the CatBoost algorithm, which achieved an outstanding AUC of 99.11% and a test set accuracy of 95.37%. Feature selection strategies, including Recursive Feature Elimination (RFE) and correlation analysis, isolated key predictors, enhancing both interpretability and model efficiency. Additionally, the study included a thorough analysis of the advantages and disadvantages of non-linear models like CatBoost along with Random Forest as well as linear models like Logistic Regression. CatBoost emerged as the top-performing model for precision-driven tasks, while KNN proved to be highly efficient for real-time applications. These results highlight ML's revolutionary potential in LC diagnostics, opening the door for scalable, precise, and efficient tools that address the challenges of early detection and classification.

VIII. References

[1] A. M. D. Wolf *et al.*, “Screening for lung cancer: 2023 guideline update from the American Cancer Society,” *CA A Cancer J Clinicians*, vol. 74, no. 1, pp. 50–81, Jan. 2024, doi: 10.3322/caac.21811.

[2] J. Cn. Chan *et al.*, “Multifaceted nature of young-onset diabetes - can genomic medicine improve the precision of diagnosis and management?,” *J Transl Genet Genom*, vol. 8, no. 1, pp. 13–34, Jan. 2024, doi: 10.20517/jtgg.2023.36.

[3] C. A. Ridge *et al.*, “Lung imaging methods: indications, strengths and limitations,” *Breathe*, vol. 20, no. 3, p. 230127, Oct. 2024, doi: 10.1183/20734735.0127-2023.

[4] H. Singh, S. Sethi, M. Raber, and L. A. Petersen, “Errors in Cancer Diagnosis: Current Understanding and Future Directions,” *JCO*, vol. 25, no. 31, pp. 5009–5018, Nov. 2007, doi: 10.1200/JCO.2007.13.2142.

[5] S. Asif *et al.*, “Advancements and Prospects of Machine Learning in Medical Diagnostics: Unveiling the Future of Diagnostic Precision,” *Arch Computat Methods Eng*, Jun. 2024, doi: 10.1007/s11831-024-10148-w.

[6] C. Liu, Y. Zhou, Y. Zhou, X. Tang, L. Tang, and J. Wang, “Identification of crucial genes for predicting the risk of atherosclerosis with system lupus erythematosus based on comprehensive bioinformatics analysis and machine learning,” *Computers in Biology and Medicine*, vol. 152, p. 106388, Jan. 2023, doi: 10.1016/j.compbiomed.2022.106388.

[7] T. Wasilewski, W. Kamysz, and J. Gębicki, “AI-Assisted Detection of Biomarkers by Sensors and Biosensors for Early Diagnosis and Monitoring,” *Biosensors*, vol. 14, no. 7, p. 356, Jul. 2024, doi: 10.3390/bios14070356.

[8] A. K. Rose, “Like Me, Buy Me: The Effect of Soft Power on Exports,” *Economics & Politics*, vol. 28, no. 2, pp. 216–232, Jul. 2016, doi: 10.1111/ecpo.12077.

[9] S. Arefin, “IDMap: Leveraging AI and Data Technologies for Early Cancer Detection,” *int.jour.sci.res.mana*, vol. 12, no. 08, pp. 1138–1145, Aug. 2024, doi: 10.18535/ijsrm/v12i08.mp03.

[10] M. Huang, J.-J. Lu, and J. Ding, “Natural Products in Cancer Therapy: Past, Present and Future,” *Nat. Prod. Bioprospect.*, vol. 11, no. 1, pp. 5–13, Feb. 2021, doi: 10.1007/s13659-020-00293-7.

[11] A. Ji, “Enhancing Lung Cancer Screening with Bidirectional LSTM and GRU Models,” *ACE*, vol. 104, no. 1, pp. 139–142, Nov. 2024, doi: 10.54254/2755-2721/104/20241187.

[12] F. M. Fatoki, E. K. Akinyemi, and S. A. Phlips, “Prediction of Lungs Cancer Diseases Datasets Using Machine Learning Algorithms,” *CJAST*, vol. 42, no. 11, pp. 15–23, May 2023, doi: 10.9734/cjast/2023/v42i114101.

[13] A. Aqila and M. Faisal, “Lung Cancer EDA Classification Using the Decision Trees Method in Python,” *ISE*, vol. 1, no. 1, pp. 8–13, Jun. 2023, doi: 10.58777/ise.v1i1.56.

[14] S. Chan, “Machine Learning-Based Prediction on Diagnosis and Severity of Lung Cancer,” *Scholarly Review Journal*, vol. Winter 2024/2025, no. 11, Dec. 2024, doi: 10.70121/001c.127451.

[15] B. Lakshmi, S. S. Kumar, N. R. Sai, K. P. V. Kumar, and G. S. C. Kumar, "WLAN Intrusion Detection System Based on SVM," in *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, Erode, India: IEEE, Apr. 2022, pp. 1213–1219. doi: 10.1109/ICSCDS53736.2022.9760896.

[16] M. Mamun, S. Bin Shawkat, M. S. Ahammed, M. M. Uddin, M. I. Mahmud, and A. M. Islam, "Deep Learning Based Model for Alzheimer's Disease Detection Using Brain MRI Images," in *2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, New York, NY, NY, USA: IEEE, Oct. 2022, pp. 0510–0516. doi: 10.1109/UEMCON54665.2022.9965730.

[17] A. Lakshmanarao, P. S. Kumar, D. S. Chauhan, M. S. Venkat, and S. K. Raj, "Medicinal Plant Classification Using Transfer Learning Through Hybrid Machine Learning and Image Processing Techniques," in *2024 International Conference on Intelligent Systems for Cybersecurity (ISCS)*, Gurugram, India: IEEE, May 2024, pp. 1–6. doi: 10.1109/ISCS61804.2024.10581006.

[18] S. Karimullah, G. A. Khan, N. Deepthi, F. Shaik, P. B., and P. Hariobulesu, "Integrated Deep Learning Methodology for Early Glaucoma Detection and Diagnosis using Retinal Fundus Images," in *2024 International Conference on Advances in Computing Research on Science Engineering and Technology (ACROSET)*, Indore, India: IEEE, Sep. 2024, pp. 1–4. doi: 10.1109/ACROSET62108.2024.10743780.

[19] M. Tanveer *et al.*, "Fuzzy Deep Learning for the Diagnosis of Alzheimer's Disease: Approaches and Challenges," *IEEE Trans. Fuzzy Syst.*, vol. 32, no. 10, pp. 5477–5492, Oct. 2024, doi: 10.1109/TFUZZ.2024.3409412.

[20] R. P.R., R. A. S. Nair, and V. G., "A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms," in *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, Coimbatore, India: IEEE, Feb. 2019, pp. 1–4. doi: 10.1109/ICECCT.2019.8869001.

[21] H. C. Ates *et al.*, "Integrated Devices for Non-Invasive Diagnostics," *Adv Funct* Materials, vol. 31, no. 15, p. 2010388, Apr. 2021, doi: 10.1002/adfm.202010388.

[22]Medical diagnosis of diabetes using deep learning techniques and big data analytics

AS Mahajan - J EmergTechnolInnov Res, 2020

[23] Generative AI Powered Forensic Device Vanita G. Kshirsagar, Digvijay G. Bhosale, ShubhangiSuryawanshi, Anita Sachin Mahajan, PramodPatil, Jyotsna Vilas Barpute 2024 8th International Conference on Computing, Communication, Control and Automation (ICCUBEA)

[24] To Analyze Live Streaming Text data for Real Life Applications using Deep Learning Techniques Mrs.Vanitaraut,Mrs.AnitaMahajan,Ms.Sanjanakarmalkar

International Journal of Advanced Research in Computer and Communication ...

[25] QMWT: Design of an Improved EEG Classification Model via Q-Learning Based Processing of Multispectral Wave Traits

Trupti T. TekalePranjali A. More, Sonali M. Sonavane, Anita S. Mahajan, YogendraPatil, Shwetal K. PatilJournal Advances in Nonlinear Variational inequalities Volume 27 Issue 4 Publisher Elsevier

[26] To Analyze Live Streaming Text data for Real Life Applications using Deep Learning Techniques Ms.SnehalsaradeMrs.Vanitaraut,Mrs.AnitaMahajan,Ms.Sanjanakarmalkar International Journal of Advanced Research in Computer and Communication Engineering Volume 10 Issue 5

[27] Review Paper on Sentimental Analysis for Recommendation System VG Kshirsagar, S Sharma Proceeding of First Doctoral Symposium on Natural Computing Research: DSNCR …

[28] A better approach towards securing mobile adhoc network Amit Chauhan, ArtiPatle, Anita Mahajan International Journal of Computer Applications Volume 975 issue 8887