Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 01s (2025)

A Comparative Study on Prediction Efficiency in Lung Cancer Detection Using Machine Learning Models

Dr. D. Kumaresan[1], Dr. B. Santhosh Kumar[2*]

[1]Department of Computer and Information Science, Faculty of Science, Annamalai University, Annamalai Nagar, Tamil Nadu, India.
[2*]Assistant Professor, Department of Computer Applications, Periyar Arts College, Cuddalore, Tamil Nadu, India.

Email: aucsedks@yahoo.co.in
*Corresponding Email: santhoshcdm@gmail.com

Article History:
Received: *15-10-2024*
Revised: *10-11-2024*
Accepted: *30-01-2025*

Abstract:

Machine learning can be utilized to analyze lung cancer data and make predictions using models trained on datasets. This analysis and prediction can help clinicians and patients by reducing and improving early detection of lung cancer. Some methods for using machine learning to analyze lung cancer datasets namely Logistic Regression, Multilayer Perceptron, SMO, J48, Random Forest, and REP Tree. Machine learning algorithms employ computational techniques to extract information directly from the dataset. These algorithms find a suitable variable for the prediction of lung cancer using different machine-learning approaches and performance matrices. This paper considers a lung cancer prediction dataset with 25 parameters. Numerical illustrations are provided to prove the proposed results with accuracy parameters.

Keywords: Lung cancer detection, machine learning, prediction, and accuracy parameters.

1. Introduction

Improving survival rates is largely dependent on early detection of lung cancer, which continues to be the primary cause of death globally related to lung cancer. Accurate lung cancer prediction and classification are essential to improving clinical decision-making and treatment outcomes. ML approaches as a valuable platform for medical diagnostics, offering the data analysis and prediction of large-scale clinical, imaging, and genomic data, identifying patterns, and predicting disease progression. A key challenge in applying ML to lung cancer detection lies in variable selection. Identifying the most relevant features – from clinical to radiological, can significantly improve the accuracy and effective predictive models. Redundant variables can lead to overfitting, hinder model generalization, and increase computational complexity. Therefore, effective variable selection is crucial for optimizing model performance. Examines different machine learning approaches to lung

cancer detection, with a focus on how different variable selection techniques. By examining the impact of feature selection on prediction accuracy, we aim to provide valuable insights into best practices for optimizing model performance in lung cancer diagnosis.

Accurate identification of this disease significantly increases cancer survival rates. A powerful lung cancer detection system should precisely locate tumors, often using computed tomography (CT). The phases of lung cancer identification and detection using different techniques discussed (Ignatious and Joseph, 2015). Numerous studies have emphasized how crucial feature selection is to raise the precision of machine learning models used to detect lung cancer. Selected the most informative radiomic features by combining SVM and RFE. The findings showed that decreasing the dimensionality of the input dataset reduced overfitting and increased SVM's classification accuracy for lung cancer (Zhang et al., 2019).

Integrated clinical (Yang, 2020) and radiomic features in developing ML models for early-stage lung cancer prediction. Using Lasso regression for feature selection, the researchers identified the most relevant variables that significantly contributed to improving the model's diagnostic accuracy. Their findings suggested that combining clinical and radiomic features, followed by efficient feature selection, could boost the robustness of ML-based lung cancer detection systems, leading to better prediction performance and reduced computational costs. PCA which is used to fine the important features and reported improved model accuracy and faster convergence rates in neural networks. Their study highlighted that feature selection not only improves prediction performance but also reduces model complexity and training time, making deep learning models more efficient in clinical practice (Ali et al., 2021).

Examined the use of radiomic features in lung cancer prediction and the importance of selecting the most relevant variables. They compared several feature selection techniques, including Mutual Information (MI), and showed an appropriate selection of radiomic features using Random Forest (RF) and SVM. Their results showed that optimized variable selection is key to improving the diagnostic ability of models based on medical imaging data (Parmar, 2018). Liu et al., 2018 investigated integrating genomic and imaging data in machine-learning models for lung cancer detection. Variable selection, combining genetic markers with radiomic features. The study demonstrated that feature selection methods tailored to multi-modal datasets could lead to substantial improvements in model performance. Elastic Net effectively minimized the number of irrelevant variables, resulting in a more robust prediction model with higher accuracy.

The data analysis and prediction for weather datasets and Play Golf class variables and to performance parameters and its conditions to playing golf or not using different machine learning approaches using J48, RT, DS, LMT, HT, REP, and RF. The performance results were calculated using measure accuracy using different test statistics. Out of seven machine learning approaches, the Random tree algorithm returns the best performance (Rajesh and Karthikeyan, 2017). Detect lung cancer within various lung images as input and classify different cancers using ML and DL methods (Bhuvaneswari and Therese, 2015).

Lung cancer detection using tumors from X-ray, CT, and MRI images. The detection process was completed using image processing techniques and methods. Mean filter and median filters are common pre-processing techniques for various stages. The accuracy parameters are used to prove the proposed research namely SNR, MSE, and PSNR utilized (Asuntha and Srinivasan, 2020). Data mining and machine learning approaches are used to

retrieve actionable results through teaching and testing using cross-validations. The data analysis and prediction related to the corresponding datasets using different DM and ML algorithms used to retrieve the actionable results (Kannan and Naveen, 2020; Rajesh et al., 2019; Rajesh and Karthikeyan, 2019; Rajesh et al., 2019)

2. Backgrounds and Methodologies

Logistic Regression

Binary outcomes, such as true or false, can be predicted using statistical modeling techniques like logistic regression. Because it's easy to understand and straightforward, it's a popular option in machine learning (Kohavi and Sahami, 1996).

Step 1. **Follow the primary steps of data preparation.**
Step 2. **Based on the dataset and problems to perform the model.**
Step 3. **Training and testing the data until convergence.**
Step 4. **Prediction based on new features.**
Step 5. **Evaluation using a confusion matrix with different performance metrics.**

2.1 Multilayer Perception

This core architecture of deep learning is applied to a number of tasks, such as regression and classification. Three different kinds of layers are commonly found in an MLP's architecture.

Input layer
Step 1. **Hidden Layers**
Step 2. **Output layer**

2.2 SMO

SVM training is done through Sequential Minimal Optimization. Classification and regression are the two distinct processes for which machine learning techniques are frequently employed. Optimizing quadratic programming is the goal of this machine learning technique.

Step 1. **Initialization using SVM.**
Step 2. **Select Two Different Lagrange Multipliers**
Step 3. **Optimization to the pair**
Step 4. **Update corresponding model performance**
Step 5. **Checking the corresponding Convergence:**
Step 6. **Repeat the** process and complete through 2 to 5 and consolidate.

2.3 J48

J48, adopted by using the C4.5 algorithm, is a problem-solving tool that fully classifies the results when applying decision tree approaches. Since J48 solves problems using both numerical and categorical features, it is an appropriate method that is regularly used worldwide.

Step 1. **Select a suitable attribute**
Step 2. **Different nodes are found then splitting the nodes**
Step 3. **Adopt the recursion approach**
Step 4. **Pruning using J48 with accuracy**
Step 5. **Finding the missing values**
Step 6. **Method of post-pruning with fully constructed**
Step 7. **Predict the leaf node based on different conditions**

2.4 Random Forest

RF is the most widely used decision tree approach in ensemble learning that uses bagging to determine classification and regression tasks.
Step 1. **Bootstrapped the sampling approaches**
Step 2. **Feature selection works with a random approach**
Step 3. **Construct the DT based on various conditions**
Step 4. **The voting system followed by regression tasks**

2.5 REP Tree

The Reduced Error Pruning (REP) Tree is a popular DT method for machine learning problems involving classification. In order to prevent overfitting, the DT is constructed using a reduced-error pruning technique.
Step 1. **Construction is based on the decision**
Step 2. **Follow the Splitting criteria recursively**
Step 3. **Reduced Error Pruning using eliminating branches**
Step 4. **The final DT is constructed with prediction**

2.6 Common Evaluation Metrics in Machine Learning

Various performance metrics are widely used around the world to determine the accurate performance of regression models.
Kappa (Cohen's Kappa): Measures the agreement between two observations on a categorical dataset. κappa = (Y(A) - Y(E)) / (1 - Y(E))
where: Y(A) called as observed agreement, and Y(E) called as expected agreement.
MAE: Utilized for calculating the absolute mean difference between the actual and predicted observations. (Akusok, 2020).

$$\text{Mean Absolute Error} = \Sigma|y_i - \hat{y}_i| * 1/n$$

where: n: how many iterations, y_i: actual value, ŷ_i: predicted value
RMSE: The average square root of the differences between the actual and predicted observations (Hosseini, 2019).

$$\text{Root Mean Squared Error} = \sqrt{(1/n * \Sigma(y_i - \hat{y}_i)^2)}$$

RAE: Find the difference between the relative error for actual and predicted values of observations. (Chi, 2020).

$$\text{Relative Absolute Error} = (\Sigma|y_i - \hat{y}_i|) / (\Sigma|y_i - \bar{y}|)$$

RRSE: Similar approaches are based on the method of RAE but use RMSE instead of MAE.

$$\text{Root Relative Squared Error} = \sqrt{(\Sigma (y_i - \hat{y}_i)^2)} / \sqrt{(\Sigma (y_i - \bar{y})^2)}$$

TPR or Recall: Find the positive results with a correctly predicted.

True Positive Rate = True_positive / (True_positive + False_negative)

FPR: Find the actual negative results with incorrectly predicted as positive.

False Positive Rate = False_positive / (False_positive + True_negative)

Precision: Find the positive predicted instances

Precision = True_positive / (True_positive + False_positive)

Recall: Similar results reflected in TPR.

F-Measure: The average values based on Harmonic_Mean between precision and the recall.

F_Measure = 2 * (precision*recall) / (precision+recall)

Area Under the Curve (AUC)

ROC AUC: Determine the area under the receiver operating characteristic curve using the model performance to analyze both positive and negative approaches. Better performance in this case was mentioned by the strongly positive AUC. Use PRC AUC to determine the area under the behaviors of precision values and recall values reflected in the curve when the distribution is unbalanced. Larger values indicate that PRC-AUC performs better when precision is more important than recall.

3.0 Experimental Outcomes

Kaggle website data repository which is used to download the related dataset. This research is based on a lung cancer prediction dataset that contains 25 features or parameters with different categories listed in Table 1 (Kaggle 2018).

Table 1. Sample Dataset

Patient_Id	Age	Gender	Air_Pollution	Alcohol_use	Dust_Allergy	Occupational_Hazards	Genetic_Risk	Chronic_Lung_Disease	Balanced_Diet	Obesity	Smoking	Passive_Smoker	Chest_Pain	Coughing_Blood	Fatigue	Weight_Loss	Shortness_Breath	Wheezing	Swallowing_Difficulty	Clubbing_Fingernails	Frequent_Cold	Dry_Cough	Snoring	Level
P1	33	1	2	4	5	4	3	2	2	4	3	2	2	4	3	4	2	2	3	1	2	3	4	L
P10	17	1	3	1	5	3	4	2	2	2	2	4	2	3	1	3	7	8	6	2	1	7	2	M
P100	35	1	4	5	6	5	5	4	6	7	2	3	4	8	8	7	9	2	1	4	6	7	2	H
P1000	37	1	7	7	7	7	6	7	7	7	7	7	7	8	4	2	3	1	4	5	6	7	5	H
P101	46	1	6	8	7	7	7	6	7	7	8	7	7	9	3	2	4	1	4	2	4	2	3	H
P102	35	1	4	5	6	5	5	4	6	7	2	3	4	8	8	7	9	2	1	4	6	7	2	H
P103	52	2	2	4	5	4	3	2	2	4	3	2	2	4	3	4	2	2	3	1	2	3	4	L
P104	28	2	3	1	4	3	2	3	4	3	1	4	3	1	3	2	2	4	2	2	3	4	3	L
P105	35	2	4	5	6	5	6	5	5	5	6	6	6	5	1	4	3	2	4	6	2	4	1	M

***** L-Low, M-Medium, H-High

Table 2. Machine Learning Models with Correctly and Incorrectly Classification (%)

ML Approaches	Correctly Classified	Incorrectly Classified
Logistic	99	1

Multilayer Perceptron	100	0
SMO	98	2
J48	100	0
Random Forest	100	0
REP Tree	99	1

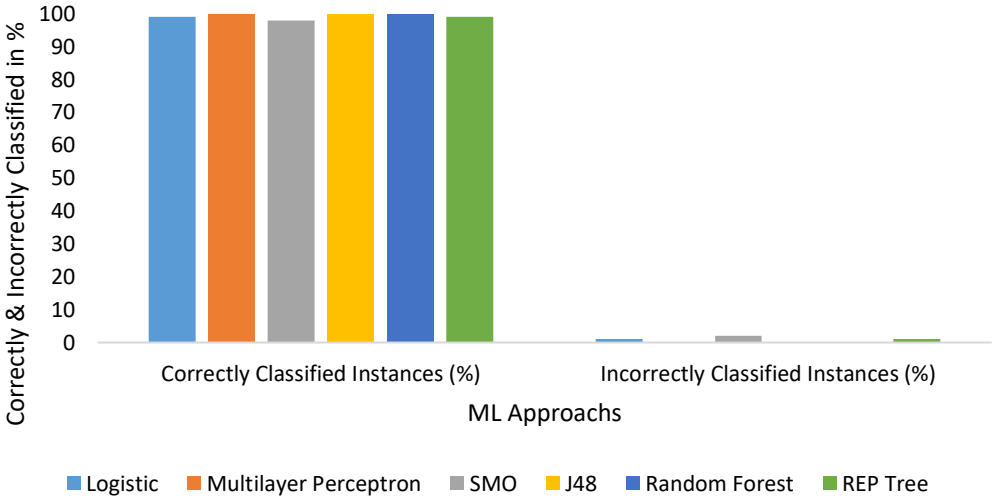

Figure 1. Comparison of Correctly & Incorrectly Classification

Table 3. Kappa statistic

ML Models	**Kappa statistic**
Logistic_Regression	0.9900
Multilayer Perceptron	1.0000
SMO	0.9821
J48	1.0000
LMT	1.0000
Random Forest	1.0000
REP Tree	0.9987

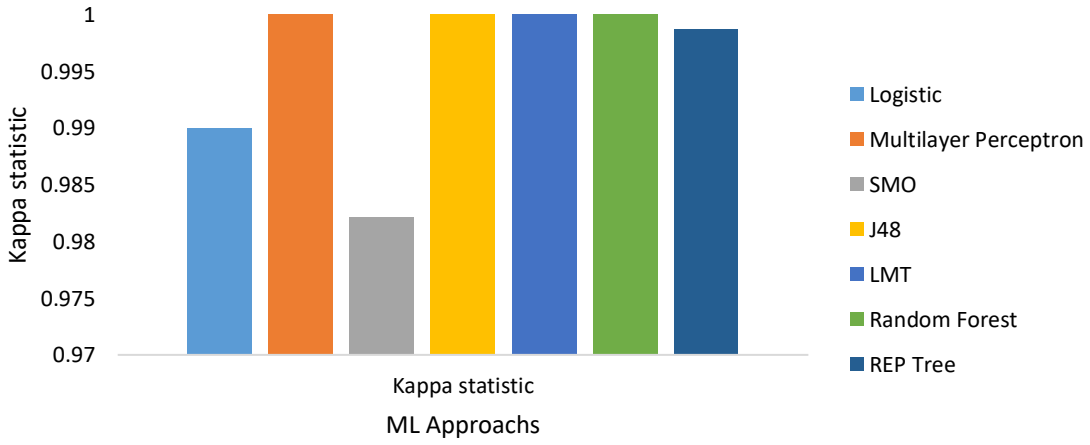

Figure 2. Kappa Statistics for Each Model

Table 4. Machine Learning Models with MAE and RMSE

ML Models	MAE	RMSE
Logistic_Regression	0.0010	0.0003
Multilayer Perceptron	0.0018	0.0035
SMO	0.2222	0.2722
J48	0.0000	0.0000
Random_Forest	0.0001	0.0013
REP_Tree	0.0120	0.0024

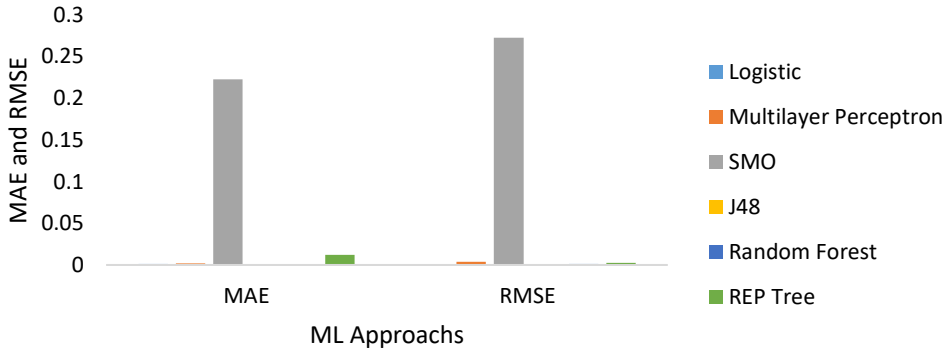

Figure 3. Presents the MAE and RMSE Values for Each Model

Table 5. RAE (%) and RRSE (%)

ML Models	RAE	RRSE
Logistic_Regression	0.0019	0.0670
Multilayer Perceptron	0.4174	0.7331
SMO	50.1438	57.8179
J48	0.0000	0.0000
Random Forest	0.0196	0.2687
REP Tree	0.0015	0.0378

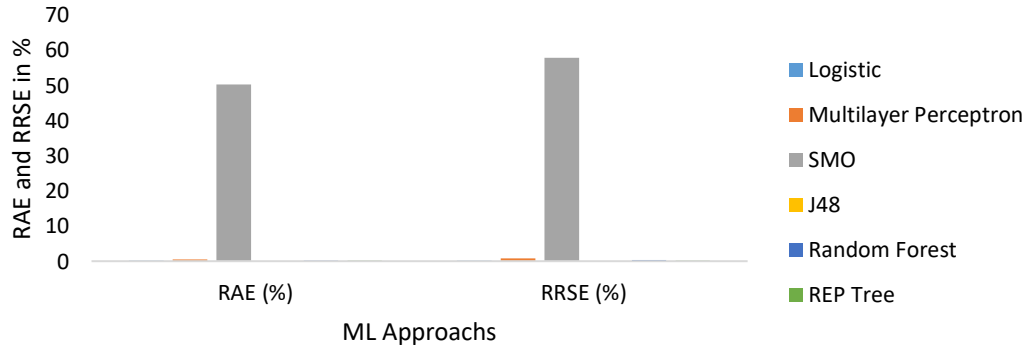

Figure 4. Visualizes the RAE and RRSE Values for Each Model

Table 6. Time Spent on model creation (Seconds)

ML Models	Time
Logistic_Regression	0.6300
Multilayer_Perceptron	4.6600
SMO	0.4100
J48	0.1400
Random_Forest	0.4900
REP_Tree	0.0700

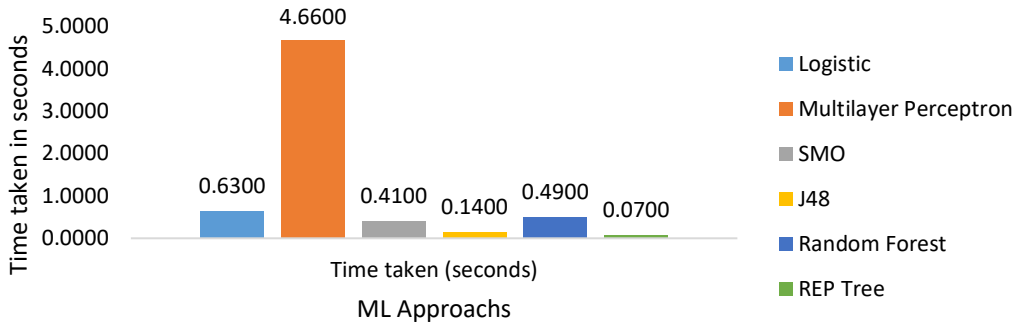

Figure 5. Depicts the Time Taken to Build Each Model

Table 7. Common Evaluation Metrics in Machine Learning

ML Approaches	Logistic	Multilayer Perceptron	SMO	J48	Random Forest	REP Tree
TP Rate	0.9000	1.0000	0.8000	1.0000	1.0000	0.9000
FP Rate	0.1000	0.0000	0.2000	0.0000	0.0000	0.1000
Precision	0.9000	1.0000	0.8000	1.0000	1.0000	0.9000
Recall	0.9000	1.0000	0.8000	1.0000	1.0000	0.9000
F-Measure	0.9000	1.0000	0.8000	1.0000	1.0000	0.9000
MCC	0.9000	1.0000	0.8000	1.0000	1.0000	0.9000
ROC Area	0.9000	1.0000	0.8000	1.0000	1.0000	0.9000
PRC Area	0.9000	1.0000	0.8000	1.0000	1.0000	0.9000

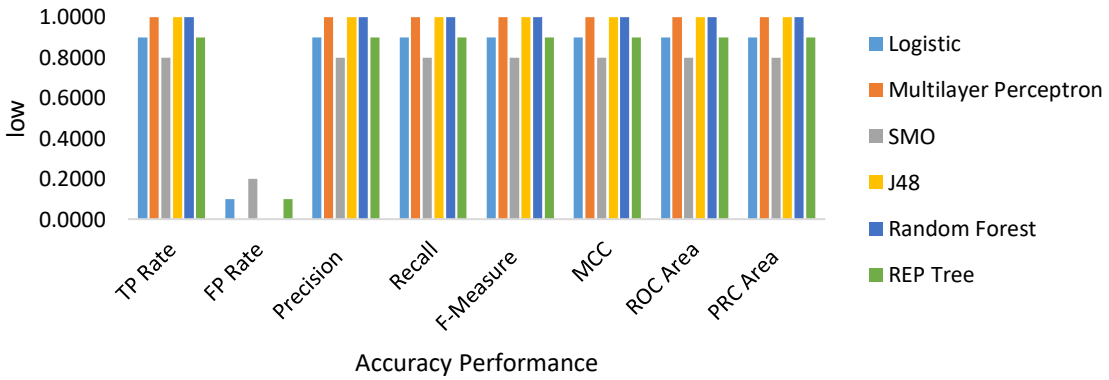

Figure 6. Common Evaluation Metrics in Machine Learning

3. Result and Discussion

The analysis of various ML models for lung cancer identification using the dataset with 25 parameters was conducted. The models evaluated include Logistic Regression, Multilayer Perceptron, SMO, J48, Random Forest, and REP Tree. The analysis covered key performance metrics such as instances of correct and incorrect classification, Kappa statistics, MAE, RMSE, and RAE, among others.

As shown in Table 2 and Figure 1, the maximum classification accuracy (100%) was achieved by Multilayer Perceptron, J48, and Random Forest. Logistic Regression and REP Tree both achieved 99%, while SMO had 98%. The Kappa statistic was used to evaluate the inter-rater agreement. Multilayer Perceptron, J48, Random Forest, and LMT achieved a perfect score of 1.000, indicating almost perfect agreement, while Logistic Regression and REP Tree were close with values of 0.9900 and 0.9987, respectively (Table 3 and Figure 2).
In terms of error metrics, the J48 model outperformed the others with an MAE and RMSE returns of 0, for using Random Forest with an MAE=0.0001 and RMSE=0.0013. SMO showed the highest MAE and RMSE values, indicating more significant prediction errors (Table 4 and Figure 3). The lowest relative errors, in both RAE and RRSE, were achieved by J48, followed by Random Forest. In contrast, SMO exhibited the highest relative errors, indicating a lower prediction performance (Table 5 and Figure 4).

Regarding training time, REP Tree was the fastest, taking only 0.07 seconds, followed by J48 with 0.14 seconds. The Multilayer Perceptron took the longest time (4.66 seconds) to train, but this was offset by its perfect performance in classification accuracy (Table 6 and Figure 5). All models, except SMO, achieved a TP rate, precision, recall, and F-measure of 1.000, with the SMO model trailing slightly behind at 0.800 for all metrics (Table 7 and Figure 6).

This analysis highlights that J48 and Random Forest are the most efficient models for lung cancer prediction, delivering the highest accuracy and lowest error metrics with minimal training time. The Multilayer Perceptron also performed exceptionally well, but with a longer training duration. SMO, while still performing reasonably, had the lowest accuracy and highest error metrics, indicating it may not be the best choice for this dataset.

4. Conclusion and Further Research

The study demonstrates using different ML models for detecting and classifying the lung cancer using 25 parameters. Among the models evaluated, J48, Random Forest, and Multilayer Perceptron exhibited perfect classification performance with 100% accuracy. These models also scored highly in terms of Kappa statistic, low error metrics (MAE, RMSE, RAE, and RRSE), and efficient training time, particularly J48 and Random Forest, which balanced high accuracy with minimal computational cost. The SMO model, while performing reasonably, showed lower accuracy and higher error rates compared to the other methods. The superior performance of Random Forest and J48 demonstrates the robustness and reliability of ensemble and decision-tree-based models for lung cancer prediction.

In conclusion, decision-tree-based models such as J48 and ensemble methods like Random Forest are recommended for lung cancer detection tasks due to their high accuracy, low error metrics, and efficient training times. These models provide clinicians with valuable

tools for early detection and diagnosis, ultimately contributing to improved treatment outcomes for patients.

Further Research

This study includes Multilayer Perceptron, further exploration of deep learning approaches with different optimization techniques like CNNs and RNNs can be conducted to assess their efficacy, particularly with larger and more complex datasets such as imaging data from CT or MRI scans. To transition these machine learning models into practical clinical use, further validation is necessary through real-time applications and clinical trials. Investigating the performance of these models on real-world patient data will provide insights into their applicability in clinical environments. By addressing these areas, future research can enhance the accuracy, scalability, and clinical relevance of ML algorithms in lung cancer detection, paving the way for more sophisticated and personalized diagnostic tools.

References

[1] Akusok, A. What is Mean Absolute Error (MAE)? Retrieved from https://machinelearningmastery.com/mean-absolute-error-mae-for-machine-learning/ (2020)
[2] Ali, M., et al. The impact of feature selection on neural networks for lung cancer prediction. Journal of Biomedical Informatics (2021) 114: 103684.
[3] Asuntha, A., Srinivasan, A. Deep learning for lung Cancer detection and classification. Multimedia Tools and Applications (2020) 79(11): 7731-7762.
[4] Bhuvaneswari, P., Therese, A. B. Detection of cancer in the lung with K-NN classification using genetic algorithm. Procedia Materials Science (2015) 10: 433-440.
[5] Chi, W. Relative Absolute Error (RAE) – Definition and Examples. Medium. https://medium.com/@wchi/relative-absolute-error-rae-definition-and-examples-e37a24c1b566 (2020)
[6] Hosseini, S. M., Hosseini, S. M., & Mehrabian, M. R. Root mean square error (RMSE): A comprehensive review. International Journal of Applied Mathematics and Statistics, (2019): 59(1): 42–49.
[7] Ignatious, S, Joseph, R. Computer aided lung cancer detection system. In 2015 Global Conference on Communication Technologies (2015) 555-558.
[8] Kaggle, https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link/data (2018)
[9] Kannan, V., Naveen, V. J. Detection of lung cancer using image segmentation. International Journal of Electrical Engineering & Technology, (2020) 2(11): 7-16.
[10] Kohavi, R., Sahami, M. Error-based pruning of decision trees. In International Conference on Machine Learning (1996): pp. 278-286.
[11] Liu, Y., et al. Integrating genomic and imaging data using elastic net for lung cancer detection. IEEE Transactions on Medical Imaging (2022) 41(5): 1352-1363.
[12] Parmar, C., et al. Radiomics feature selection in lung cancer: an empirical study. Medical Physics (2018) 45(12): 5472-5481.
[13] Rajesh, P., Karthikeyan, M. A comparative study of data mining algorithms for decision tree approaches using the Weka tool. Advances in Natural and Applied Sciences (2017) 11(9): 230-243.

[14] Rajesh, P., Karthikeyan, M. Data mining approaches to predict the factors that affect agriculture growth using stochastic models. International Journal of Computer Sciences and Engineering (2019) 7(4): 18-23.
[15] Rajesh, P., Karthikeyan, M., Arulpavai, R. Data mining approaches to predict the factors that affect the groundwater level using a stochastic model. In AIP Conference Proceedings (2019) 2177(1): 1
[16] Rajesh, P., Karthikeyan, M., Santhosh Kumar, B., Mohamed Parvees, M. Y. Comparative study of decision tree approaches in data mining using chronic disease indicators (CDI) data. Journal of Computational and Theoretical Nanoscience (2019) 16(4): 1472-1477.
[17] Yang, X., et al. Clinical and radiomic feature selection for lung cancer risk prediction: a Lasso-based approach. Medical Physics (2020) 47(8): 3757-3768.
[18] Zhang, L., et al. Feature selection and machine learning-based lung cancer classification using radiomics. Computers in Biology and Medicine (2019) 107: 41-46.