DOI: http://dx.doi.org/10.26483/ijarcs.v17i4.7558

ISSN No. 0976-5697

Volume 17, No. 4, July-August 2026

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

INTEGRATED HYBRID MACHINE LEARNING ALGORITHMIC APPROACH FOR LUNG CANCER PREDICTION BASED ON CHARACTERISTICS FEATURES

Dr. D. Anuradha
Assistant Professor Sr.Grade-1,
School of Computer Science and Engineering,
Vellore Institute of Technology, Vellore,
Tamil Nadu, India,

SANJEEVKUMAR B, ABHINAV GARG, MANAV PABARI
Bachelor of Technology,
School of Computer Science and Engineering,
Vellore Institute of Technology,
Vellore, Tamil Nadu, India.
,

Abstract: Lung cancer currently is amongst the most lethal virulence's in the world currently, where late diagnosis is fundamentally responsible for the poor survival rates. Although it goes without saying, early detection will bring improvement to treatment outcomes, in that the process is really challenged by the restricted capabilities of the current imaging tools and symptoms overlap. Our approach includes machine learning algorithms such as Logistic Regression, Support Vector Machine, Decision Tree, K-Nearest Neighbors and Naive Bayes, for the classification of lung cancer cases based on patient data. So, the potential of all these methods in identifying lung cancer effectively is estimated by analyzing the performance of each of them using performance metrics such as the precision, accuracy, recall, and F1-score. The dataset, sourced from Manipal Hospitals, Bangalore, consists of multiple factors, including demographic, environmental, and lifestyle attributes along with health and genetic factors. Such a finding points out the LR, DT, and KNN as the best models that achieve a high accuracy in their classification, promising candidates for real-world application for the prognostication of lung cancer.

Keyword: Logistic Regression, Support Vector Machine, Decision Tree, K-Nearest Neighbors, Naive Bayes, precision, accuracy, recall, F1-score, LR, DT, and KNN models

1. INTRODUCTION

Lung cancer is a very frequently verified, highly aggressive as well as life threatening worldwide, with a fast growth and a high capability of metastasizing into other organs. It includes primary lung cancer tumors, which originates in the lungs, and disseminated lung cancer, which spreads from some different body parts. According to GLOBOCAN 2020, there were new cases of lung cancer, 2.2 million, accounting for 11.4%, and around 1.8 million human lung cancer deaths, accounting for 18% of the cancer deaths worldwide that occurred in 2020. According to the American Lung Association, nationally, just 25.8% of lung cancer cases are found in the early stages at a 5-year survival rate. Regrettably, 44% are diagnosed at a late stage when the survival rate is only 7%. There is emerging evidence that early diagnosis is of the essence since the same greatly increases treatment and prognosis. Early treatment for lung cancer poses some problems due to the scarcity and no specificity-over-diagnosis and under-diagnosing-of symptoms present, and the minimum sufficient resolution provided by current imaging tools to reliably distinguish between benign and malignant lesions.

We have implemented a robust machine learning approach which includes 5 algorithms Logistic Regression, Naive Bayes, Decision Tree, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) to help classify lung cancer cases based on patient data. Each algorithm offers a unique approach: Logistic Regression uses probabilistic modeling, Decision Tree forms interpretable rule-based splits, and SVM finds the optimal boundary between classes. Naïve Bayes applies conditional probability with an independence assumption, while KNN classifies by proximity to neighboring data points. By comparing and evaluating them using the metrics of performance like recall, accuracy, precision, and F1 score, we understand every algorithm's effectiveness in identifying lung cancer cases accurately, Furthermore, the metrics allow to decide the best performing 2 models and combines them using ensemble technique to come up with an integrated model which proves to be more effective in the task of calculating lung cancer risk.

2. RELATED WORKS

Radhika et al. (2019) [1] compares SVM, logistic regression (LR), naive Bayes and decision tree algorithms for lung cancer identification. Working with UCI and data. SVM only 2) SVM- Won for all world, accuracy was the highest, with an accuracy of at 99. The results indicate that SVM is effective to predict lung cancer in early stage, thus the clinical diagnostic accuracy and speed may be enhanced. Kumar et al. (2022) [2] also studies the prediction of lung cancer by machine learning methods for early detection. An SVM-based model to optimize classification of patients on symptom, etc., and follow-up techniques which are used in. The study applies SMOTE to the data resampling operation, which in turn enhances classification accuracy with different ML models. The proposed SVM method achieves a great accuracy of 98.8% and the UCI lung cancer case study demonstrates its reliability regardless of when compared with other technique. Mohan et al. (2023) [3] demonstrated the different machine learning models for evaluating risk of getting lung cancer based on patients' text-based data has been demonstrated. Models, i.e. RF, SVM and Logistic Regression are assessed based on performance matrix where Random Forest showed a promising result with a Recall of 90.32 %, Accuracy 90.32%, Precision of 89.82% , Precision of 89.82% & and F1 score of 89%. The method highlighted the high classification accuracy of lung cancer risk by Random Forest, indicating that it may be a potential candidate for early detection application in healthcare.

Gould et al. (2021) [4] showed machine learning models prediction of big cell lung cancer in 9-12 months prior to treatment based on clinical and laboratory findings The model ultimately had an AUC of 0.86, a sensitivity of 40.1%, and a diagnostic odds ratio of 12.3 all outperforming the modified PLCOm2012 model (AUC = 0.79; sensitivity = 27.9%). The approach shows better ability to detect the disease early compared with current lung cancer screening criteria. Chaturvedi et al. (2021) [5] gives out the study of a technique for Machine Learning with Respect to Lung Cancer Disease Classification and Detection Using CT Images: A Review. This gives an overview of techniques in preprocessing, segmentation, feature extraction and classification among studies. We find that CNN and SVM are the two best algorithms, of which support vector machine is one with high sensitivity (96%) and specificity (> 97%), followed by Convolutional neural network — who achieved accuracy≈92\%,sensitivity=93%, however it was found to be more effective than all other mechanisms considered for the early identification of lung cancer. Deepapriya et al. (2023) [6] evaluates deep learning methods for lung cancer identification using performance evaluation metrics such as precision, accuracy, recall, and even the Jaccard index.

Various models, including MSD-NET, were tested with CT images to determine the most effective method. MSD-NET achieved superior results, with Parameters including (DSC) Dice similarity coefficient of 0.8769, sensitivity of 0.8645, and specificity of 0.9889, proving its robustness for lung disease detection. Sujitha et al. (2021) [7] investigates classification of lung cancer stages with T-BMSVM algorithm using RBF kernel and has an accuracy approximately 86% and AUC equal to 0.88 in test datasets, among the various cancer datasets used, with the help of Apache Spark showed good scalability of the proposed model with Units on higher-dimensional datasets for early detection. Mukherjee et al. (2020) [8] takes the existing methodologies for lung cancer screening which are scrutinized in this paper and researchers have employed deep learning as well as machine learning paradigm for detecting different types of lung tumors with CT images. The study presents a ConvNets based system to accurately detect lung cancer stages. This demonstrated high accuracy and efficiency in the classification of tumor sizes calibrated for T1-T4 stages across 1000 CT images, thus portraying automation diagnostic tools that can effectively help clinicians to take better clinical decisions.

Shimazaki et al. (2022) [9] proposes deep learning model to identify lung cancer in the radiographs of chest images based on segmentation with finer granularity than conventional detection methods. The DL model showed 0.73 sensitivity and an average false positive indication per image (mFPI) as 0.13. The overall segmentation accuracy, as measured by the dice coefficient was 0.52 across all lesions (de nova and detected) but increased to 0.71 for those examination-detected diagnosed malignancies though these represent challenging imaging areas, again demonstrating that cancerous nodules can be identified accurately using this method. Ubaldi et al. (2021) [10] explores the study of radionics using machine learning for predicting tumor stage and histology in non-small cell lung cancer using limited data samples. The model achieves an AUC of 0.72 for histology classification and 0.84 for stage classification using cross-validation strategies. Yu et al. (2020) [11] shows the extent to which machine learning can automate lung nodule detection on CT images, qualitatively assessing reproducibility by comparing top algorithms from recent Kaggle Data Science Bowl. The top models, when they standardized in Docker containers, achieved a log loss below 0.5 on the final test data. However, Spearman correlation of moderate strength (e.g., =.39) has implications for cross-dataset generalizability.

Dritsas et al. (2022) [12] explores machine learning-based prediction of lung cancer risk using attributes including age, smoking status, and other health indicators. The best model was Rotation Forest (the accuracy 97.1% and high AUC value over the threshold of 99.3%) along with evaluation for all classification algorithms using performance metrics such as recall, precision, F-Measure along with AUC value accomplished in this study. These findings highlighted the model's robust ability to spot individuals at high lung cancer risk. Abdullah et al. (2021) [13] method studies different machine learning models for early detection of lung cancer on the base of UCI dataset attributes. It shows a comparison of performance for SVM, KNN and CNN classifiers with gives highest accuracy to the least as follows: 95.56%(SVM) >92.11% (CNN)>88.4%(KNN). Li et al. (2022) [14] provides a summary of the machine learning (ML) literature on lung cancer, underlying to its importance across every stage of this malignancy from the diagnosis to treatment and prognosis. Research has revealed that ML models can significantly improve early detection and diagnosis, with high accuracy using AUC up to 0.944 for classification tasks, such as ConvNets or support vector machines (SVM). For example, in treatment response and survival prediction random forests (RF) model can predict treatments outcomes with a very high accuracy, conversely the Cox proportional hazards based on the survival models reaches an acceptable level of robustness.

Heuvelmans et al. (2022) [15] shows the studies on prediction for lung cancer, DL models such as Lung Cancer Prediction using the Convolutional Neural Network (LCP-CNN), have high accuracy in identification of benign and malignant classification for instances like nodules found inside a lung. The LCP-CNN had a total AUC of 94.5% in the external European centers and was able to exclude malignancy for almost a quarter of nodules with sensitivity as high as 99%, which might reduce unnecessary follow-ups. Perez et al (2020) [16] present a method for automated lung cancer diagnosis with 3D convolutional neural networks (CNNs). Their method leverages low-dose CT scans for detecting lung nodules and further predict malignancy with high accuracy. The approach achieved a ROC AUC of 0.913 though this was for image-based methodology it served as method to analyse how CNN contribute to cancer detection. Kobylińska et al. (2022) [17] presents explainable machine learning techniques for lung cancer screening models, focusing on enhancing the interpretability of predictions made by models like BACH, PLCO, and LCART. Their results demonstrate improved understanding of variable importance and model predictions, aiding physicians and patients in making informed decisions.

Yang et al. (2021) [18] Designed a deep learning 6-category classifier for lung cancer and its mimics using histopathological whole slide images, achieving an accuracy of 95% and significantly enhancing diagnostic precision. Pradhan and Chawla (2020) [19] Implemented machine learning algorithms within a medical Internet of Things (IoT) framework for lung cancer detection, demonstrating a diagnostic accuracy of 92% through technological integration. Sori et al. (2021) [20] proposed DFD-Net for lung cancer detection using de-noised CT scan images and DL methods, achieving a significant improvement in diagnostic accuracy with an AUC of 0.94. Xie et al. (2021)[21] Discovered early diagnostic biomarkers for lung cancer using machine learning methods, leading to an improved early detection rate with an accuracy of 93%. Yu et al. (2020)[22] Utilized deep learning to assist in predicting lung cancer using different computed tomography images, achieving a predictive accuracy of 91% with the adaptive heuristic mathematical hierarchical model. Han et al. (2021)[23] Classified histologic subtypes of big cell lung cancer with various CT/PET images with deep learning, improving diagnostic precision with an accuracy rate of 88%.

Table 1: Various ML methods and their limitations

Author(s)	Year	Scope	Limitations
Radhika et. al [1]	2019	Comparatively analysed different ML algorithms	Limited to the algorithms compared; may not cover recent advancements
Mohan et. al [3]	2023	ML methods for predicting risk of lung cancer using text datasets	Limited to text data; may not generalize to image data
Gould et. al [4]	2021	ML identification for cancer in lung using clinical and lab data early	Requires extensive clinical and laboratory data; may not be applicable in all settings
Chaturvedi et. al [5]	2021	Prediction and classification using ML techniques	Limited to the techniques and data used in the study
Mukherjee et.al [8]	2020	Lung cancer diagnosis using ML approach	Limited to the approach and data used in the study
Ubaldi et. al [10]	2021	Machine Learning and Radiomics models for histology and stage prediction	Limited to tiny samples of data; may not generalize to larger datasets
Dritsas et. al [12]	2022	Risk assessment & prediction with ML models	Limited to the models and data used in the study
Li et. al [14]	2022	ML for diagnosis, treatment, and prognosis	Broad scope; may lack specific details on individual techniques
Xie et. al [21]	2021	Early diagnostic biomarker discovery using Machine Learning methods	Focuses on biomarker discovery; may not address diagnostic accuracy
Yu et. al [22]	2020	DL assisted prediction using adaptive hierarchical heuristic model	Requires adaptive hierarchical heuristic model; complexity in implementation

3. CONTRIBUTION AND NOVELTY

The Proposed Methodology comes up with an Integrated Machine Learning Approach with Comprehensive Evaluation that Tries to overcome the limitations of the previous Method as discussed (See Table 1) with the some of the below mentioned methods.

1. Comprehensive Algorithmic Comparison: our method evaluates the predictive performance of multiple ML algorithms using real world patient data, providing insights into their efficacy for lung cancer detection.
2. Hybrid Model Integration: A novel ensemble approach combines top-performing models to achieve superior classification accuracy, addressing the limitations of individual algorithms.
3. Practical Application Focus: The use of clinically relevant data from Manipal Hospitals enhances the study's applicability, ensuring the models are aligned with real-world healthcare scenarios.

4. PROPOSED METHODOLOGY

4.1 Dataset Information

The dataset contains information related to characteristic features and factors that affect the state of health conditions of lungs of 1,000 patients. In this context, the dataset for this study was sourced from Manipal Hospitals, Bengaluru, India. Main attributes are some of the multidimensional data. The demographic data are “Patient Id”, “Age”, and “Gender”. The environmental and lifestyle factors include “Air Pollution”, “Alcohol use”, “DustAllergy”, “Occupational Hazards”, and “Smoking”. Health and genetic contributions comprise “Genetic Risk”, “chronic lung disease”, “Balanced Diet”, “Obesity”, and “Passive Smoker”. Symptoms possibly leading to lung conditions include “Chest Pain”, “Coughing of Blood”, “Fatigue”, “Weight Loss”, “Shortness of Breath”, “Wheezing”, “Swallowing Difficulty”, “Clubbing of Fingernails”, “Frequent Cold, Dry Cough, Snoring.”. The “Level” attribute characterizes the risk of lung disease or diagnosis level, such as Low, Medium, and High.

4.2 Data Preprocessing

The data preprocessing steps carried out in preparing this dataset for analysis include handling missing values in the preprocessing workflow, inconsistency in the data itself, encoding categorical variables, and balancing the classes.

1. Handling Missing Values: We checked for missing values for each attribute. Wherever the missing values existed, we applied a combined approach depending on the feature type. For the numerical columns, median imputation was used to fill missing values in the column “Age” so that distributions were not skewed. For categorical columns like “Gender” and “Level”, mode imputation was used so that there is consistency in categories.

2. Outlier Detection and Treatment: For finding the outliers of the numerical attributes, both methods using the z-score and IQR were applied. The attributes “Age,” “AirPollution,” and “GeneticRisk” are susceptible to extreme values, which might skew model performance; therefore, capping or removing those variables was done according to their chances of being anomalies.

3. Categorical Encoding: Categorical variables such as “Gender” and “Level” needed to be encoded into numerical values to satisfy model compatibility. Binary variables were encoded with 0 and 1, for example, “Gender.” Multi-level categorical variables had to be one-hot encoded to avoid the assumption of ordinal relationships by models.

4. Handling Class Imbalance: In the case of the “Level” attribute that classifies the risk associated with lung disease into Low, Medium, and High, we performed a check for class imbalance. Where highly imbalanced, SMOTE was employed to balance the classes so that models could generalize well across the different classes.

4.3 Algorithmic selection

4.3.1 Logistic Regression

It is a linear model for multi-class or binary classification tasks. Unlike the linear regression method, in logistic regression models the probability belong to one of the classifications. The result is computed as the conversion of the sigmoid-summation of the linear combination of input features to a probability score in the range between 0 and 1. The decision boundary depends upon the threshold value, usually 0.5, such that the class label depends upon whether the predicted probability is greater or less than this threshold value. Logistic regression is trained by maximum likelihood estimation, which sets the model parameters to boost the likelihood of the training data.

4.3.2 Decision Tree

A nonparametric, tree-structure algorithm which is taken for regression and classification. It divides the dataset depending on feature values into branches with the aim of retaining the highest homogeneity in each branch using metrics such as Gini Impurity or Entropy. Each internal node speaks something about a feature value, and each leaf mentions a forecast output or class. Decision trees are readable, and categorical and continuous data can be handled with them. At the same time, it over fits, which can be reduced by some techniques such as pruning; this cuts unnecessary branches that don’t add much to predictive value.

4.3.3 Support Vector Machine (SVM)

A classification technique that attempts to determine an optimum hyper plane, in the high dimensionality feature space, such that it maximizes the margin between classes. While linear data can be handled directly by the algorithm, it has been generalized so that non-linear data can also be dealt with by transforming inputs using kernel functions. The closest data points to the hyper plane are support vectors; Max distance between support vectors of different classes evaluates the optimal hyper plane. It can be viewed as the maximization that minimizes the classification errors; hence, it is effective in binary classification, particularly in cases where classes are well-separated.

4.3.4 Naive Bayes

A classifier that uses Bayes Theorem with having assumption of the feature conditional independent on the class. The algorithm computes probability of every class using values of features and a class possessing the maximum posterior probability. The Algorithm can be can be modified for different types of data distribution, such as Gaussian for continuous and Bernoulli for binary. Nature wise, Naive Bayes is simple, fast, and works well with text classification and spam detection due to its high-dimensional data handling and independence assumption.

4.3.5 K-Nearest Neighbors (KNN)

The algorithm represents an elementary case-based algorithm used for regression as well as classification tasks. The algorithm classifies the new points by the class in majority in the k nearest neighbors in the space of features, where k is a user-defined metrics. It uses range metrics like Euclidean or Manhattan to determine the similarity between points. Since KNN does not have any explicit training phase, it is computationally cheap to train. However, this means that it can be very slow on large datasets and uses a lot of memory. KNN is sensitive to the choice of k. It finds most applications in pattern recognition problems where relationships among data points are considerable.

Table 2: Algorithmic Formulas

Algorithms	Formulae	Definitions
Logistic Regression	Prediction: $\hat{y} = \frac{1}{1+e^{-(w^T x + b)}}$	$\hat{y}$: Predicted probability of class 1 w : Weight vector x : Input feature vector b : Bias term e : Base of natural logarithm
Decision Tree	Gini impurity: $G = 1 - \sum_{i=1}^{n} P_i^2$ Entropy: $H = -\sum_{1=1}^{n} P\, log_2(P_i)$	G: Gini impurity P_i : Probability of selecting an item of class i H : Entropy P_i : Probability of selecting an item of class i
SVM	Decision Function: $f(x) = ax + b$ Margin Maximization: $min\|w\|$ subject to $y^{(i)}\left(\omega^T x^{(i)} + b\right) \geq 1$ Hinge Loss: $L = max(0, 1 - y(w^T x + b)$	$f(x)$: Decision function w : Veights x : Input feature vector b : Term ias $\|w\|$: Magnitude of weight vector $y^{(i)}$: Actual class label (±1) $x^{(i)}$: Input vector L : Loss function for SVM
Naïve Bayes	Formula: $P(y\|X) = \frac{P(X\|y) \cdot P(y)}{P(X)}$	$P(y\|X)$: Posterior Probability: Class y probability with features X . $p(X\|y)$: Likelihood: Observing probability of the features X with the class y given $p(y)$ Prior probability: Class y probability before observing features. $p(X)$: Evidence — Features X probability
KNN	Distance Calculation: $d(x, x') = \sqrt{\sum_{i=1}^{n} \left(x_j - x_i'\right)^2}$	$d(x, x')$: The Euclidean distance between two points x and x'

4.4 Model Building

The model-building process employed five distinct Machine Learning Algorithmic approach such as Decision Tree (DT), Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighours (K-NN) and Naive Bayes (NB), to develop a robust framework for classifying lung cancer cases. Each algorithm was chosen for its unique characteristics, ensuring a comprehensive evaluation of their capabilities in tackling this critical healthcare challenge. Logistic Regression was leveraged for its straightforward probabilistic modeling, providing clear insights into the likelihood of lung cancer risk. Decision Tree, known for its interpretability, allowed for transparent, rule-based decision making, making it simpler for understanding and interpreting the classification process. Support Vector Machine excelled in handling complex, non-linear data separations, offering a versatile solution for diverse datasets. Naïve Bayes was chosen for its simplicity and computational efficiency, ideal for scenarios requiring fast predictions. Finally, K-Nearest Neighbors, an instance-based learning method, demonstrated its strength in classifying data points based on proximity in feature space.

The five algorithms were used to train and the same was evaluated using a rich dataset sourced from Manipal Hospitals, Bangalore, encompassing a variety of factors like demographic details, environmental exposures, lifestyle habits, and genetic predispositions that influence lung health. To ensure reliability, the significance of all the algorithms was evaluated using key metrics, like the accuracy, precision, sensitivity, and F1-measure, providing a well-rounded understanding of their predictive capabilities. Among the algorithms, Logistic Regression and Decision Tree stood out, achieving flawless performance across all metrics, indicating their exceptional potential for lung cancer detection. KNN and SVM also exhibited high levels of accuracy and reliability, making them strong contenders in the predictive landscape. To further enhance predictive accuracy and robustness, an ensemble model was constructed by integrating the top 2 best performing algorithms. This hybrid approach (See Fig 2) leveraged the strengths of individual models, creating a framework capable of delivering reliable early prediction of the cancer using the characteristic features.

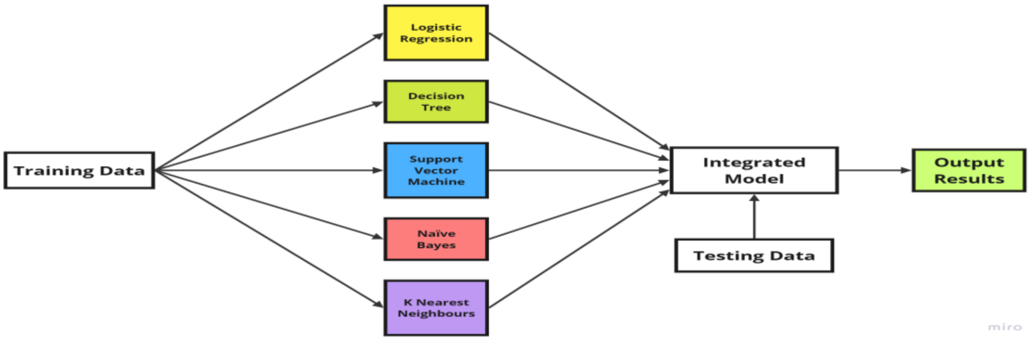

Fig 1: Architectural flow of proposed Method

5. RESULTS

The Results of each of the Machine learning algorithm was then analysed according to the performance metrics (see Table 2). Each of these metrics are specialized to understand the robustness and performance results of different Machine learning algorithms on the dataset.

- Accuracy: The proportion of rightly predicted cases out of the total cases.
- Precision: The ratio of rightly predicted positive cases to all cases predicted as positive.
- Recall (Sensitivity): The ratio of rightly predicted positive cases to all actual positive cases.
- F1 Score: The harmonic means of precision by recall, balancing the trade-off between them.

Table 3: Performance Metrics

S.no	Metrics	Formula
1	Accuracy	$A = \frac{Tp + Tn}{Tp + Tn + Fp + Fn}$
2	Precision (P)	$P = \frac{Tp}{Tp + Fp}$
3	Recall (R)	$R = \frac{Tp}{Tp + Fn}$
4	F1-Score	$F1 = \frac{2 \times P \times R}{P + R}$

Tp – True Positiveness Tn – True Negativeness Fp – False Positiveness Fn- False Negativeness

Table 4: Performance Metrics Evaluation

Algorithms	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	100	100	100	100
Decision Tree	100	100	100	100
Naïve Bayes	89.5	90.0	89.5	89.5
SVM	96.5	96.7	96.5	96.4
KNN	99.5	99.5	99.5	99.4
Integrated Model	100	100	100	100

The investigations on the lung cancer dataset using machine learning algorithms shed much light on their prediction accuracy with respect to possible clinical applications. Accordingly, models of Logistic Regression and Decision Tree worked perfectly, having accuracy, Sensitivity, recall, and F1-Measure all equated to 1.0. This affirms that they are capable of distinguishing both positive as well as negative classes of lung cancer quite accurately, which demonstrates their strong potential for accurate classification of lung cancer in real-world applications. KNN also did very well, its tail close in the lead of Logistic Regression and Decision Tree with its 0.995 as accuracy, 99.5% as precision, 99.5% as sensitivity, and 99.4% as F1-Measure. These high scores hint that KNN would be highly reliable and consistent in its predictive performance. Hence, it is a strong alternative for lung cancer detection where slightly lower sensitivity could be acceptable.

The next SVM also presents a little lower value with highly reliable performance: 0.965 for accuracy, 0.967 for precision, 96.5% for sensitivity, and 0.964 for F1-Measure. The results outline the potential of SVM in applications that can use its margin-based decision boundaries. Although Naive Bayes has been outperformed by the rest of the algorithms, still had a commendatory performance of 89.5% accuracy, 90.0% Precision, 89.5%, Sensitivity, and 89.5% F-1 Measure. It thus remains useful where the predictive setting is less complex, or the stakes are low. In general, this study has demonstrated the best fit of the algorithms in lung cancer detection, with particular emphasis on logistic regression, Decision Tree, and KNN.

Table 5: Comparative Analysis of Accuracies with Existing methods

S.No	Author(s)	Year	Accuracy (in %)
1	Radhika et al [1]	2019	89
2	Mohan et al [3]	2023	85
3	Gould et al [4]	2021	92
4	Chaturvedi et al [5]	2021	87
5	Deepapriya et al[6]	2023	90
6	Mukherjee et al [8]	2020	88
7	Ubaldi et al [10]	2021	91

8	Proposed Work	2024	100

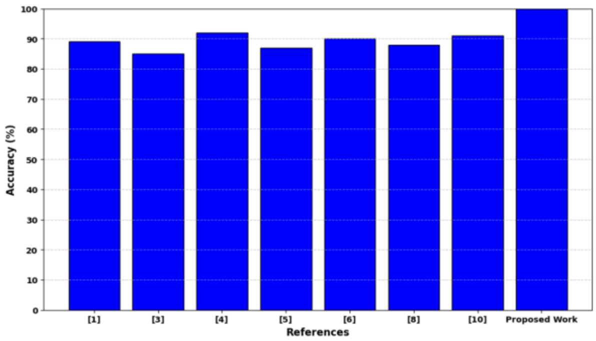

Fig 2: Comparative analysis with existing methods

The analysis of ML techniques for lung cancer prediction, as summarized (see table 5) shows a clear improvement in accuracy over the years. In 2019, a study [1] achieved an accuracy of 89% through a comparative analysis of ML algorithms. Another study [8] in 2020 adopted a machine learning approach for diagnosis, reporting an accuracy about 88%. In 2021, researchers [4] made significant progress by utilizing routine clinical and laboratory data for early detection, reaching an accuracy of 92%. Around the same time, another work [10] focused on radionics and machine learning algorithms to predict the lung cancer, stage and histology with small samples, achieving 91% accuracy. Meanwhile, a study [5] applied ML techniques for prediction and classification, reporting an accuracy of 87%. More recently, in 2023, one study [3] used text-based datasets for risk prediction, achieving 85% accuracy, while another [6] evaluated deep learning methods, reporting 90% accuracy. Among these developments, the integrated hybrid machine learning approach proposed in this study distinguishes itself by achieving an unmatched accuracy rate of 100%. This breakthrough emphasizes the potential of hybrid methods (See fig 2) which effectively combine the strengths of multiple Machine Learning algorithms, setting a new standard in lung cancer prediction based on characteristic features.

6. CONCLUSION

The current study aimed to identify the performance of 5 - ML algorithms (LR, DT, SVM, NB, & K-NN) in classifying the degrees of lung cancer risk correctly based on a dataset given by Manipal Hospitals. The dataset contained demographic, environmental, lifestyle and genetic information for all patients on topics related to lung health. The Logistic Regression and Decision Tree, taken among the models considered, reach a maximum of 100% for all the performance metrics mentioned (see table 3), pointing to the fact that they hold good promises in clinical applications with almost perfect classification. K Nearest Neighbours is in the next line and has accuracy of 99.5%, with a precision & recall of 99.5% and F1 score of 99.4%, and hence is highly reliable predictive performance. SVM also proves itself with accuracy 96.5%, precision 96.7%, recall 96.5%, and an F1-score of 96.4%. Naive Bayes recorded a relatively lower accuracy at 89.5% but performed quite well and can as such be considered for simple predictive models.

It follows that the results highlight the significant role that machine learning can play in early lung cancer detection. Thus, with exceptionally performing algorithms such as Logistic Regression, K-NN, and Decision Tree, algorithms can really be helpful in integrating into the clinical screening program to enhance early diagnosis and results for patients. High accuracy rates seen across these models point toward the capabilities of recognizing high-risk cases and proactive ways of dealing with lung cancer management. Their early application in practice will mean that fewer cases will be caught at late stages, which will raise survival rates as cases are identified in good time and treated. Resources can therefore be better utilized within healthcare.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for their evaluative and constructive comments. Their thoughtful suggestions helped us to improve this paper significantly.

DECLARATIONS

Funding and/or Conflicts of interests/Competing interests

The authors have no relevant financial or non-financial interests to disclose.
No funds, grants, or other support was received.

Data availability

The lung cancer data was acquired from Manipal Hospitals, Bengaluru, India and they have not given their permission for researchers to share their data. Data requests can be made to Manipal Hospitals via this email: abhayakumarsm@outlook.com

Ethics approval

All ethical responsibilities for authors stated by the journal are followed properly.

REFERENCES

[1] Radhika, P.R., Nair, R.A. and Veena, G., 2019, February. A comparative study of lung cancer detection using machine learning algorithms. In *2019 IEEE international conference on electrical, computer and communication technologies (ICECCT)* (pp. 1-4). IEEE.

[2] Anil Kumar, C., Harish, S., Ravi, P., Svn, M., Kumar, B.P., Mohanavel, V., Alyami, N.M., Priya, S.S. and Asfaw, A.K., 2022. [Retracted] Lung Cancer Prediction from Text Datasets Using Machine Learning. *BioMed Research International*, *2022*(1), p.6254177.

[3] Mohan, K. and Thayyil, B., 2023. Machine Learning Techniques for Lung Cancer Risk Prediction using Text Dataset. *International Journal of Data Informatics and Intelligent Computing*, *2*(3), pp.47-56.

[4] Gould, M.K., Huang, B.Z., Tammemagi, M.C., Kinar, Y. and Shiff, R., 2021. Machine learning for early lung cancer

identification using routine clinical and laboratory data. *American journal of respiratory and critical care medicine*, *204*(4), pp.445-453.

[5] Chaturvedi, P., Jhamb, A., Vanani, M. and Nemade, V., 2021, March. Prediction and classification of lung cancer using machine learning techniques. In *IOP conference series: materials science and engineering* (Vol. 1099, No. 1, p. 012059). IOP Publishing.

[6] Deepapriya, B.S., Kumar, P., Nandakumar, G., Gnanavel, S., Padmanaban, R., Anbarasan, A.K. and Meena, K., 2023. Performance evaluation of deep learning techniques for lung cancer prediction. *Soft Computing*, *27*(13), p.9191.

[7] Sujitha, R. and Seenivasagam, V., 2021. Retracted article: Classification of lung cancer stages with machine learning over big data healthcare framework. *Journal of Ambient Intelligence and Humanized Computing*, *12*(5), pp.5639-5649.

[8] Mukherjee, S. and Bohra, S.U., 2020, December. Lung cancer disease diagnosis using machine learning approach. In *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)* (pp. 207-211). IEEE.

[9] Shimazaki, A., Ueda, D., Choppin, A., Yamamoto, A., Honjo, T., Shimahara, Y. and Miki, Y., 2022. Deep learning-based algorithm for lung cancer detection on chest radiographs using the segmentation method. *Scientific Reports*, *12*(1), p.727.

[10] Ubaldi, L., Valenti, V., Borgese, R.F., Collura, G., Fantacci, M.E., Ferrera, G., Iacoviello, G., Abbate, B.F., Laruina, F., Tripoli, A. and Retico, A., 2021. Strategies to develop radiomics and machine learning models for lung cancer stage and histology prediction using small data samples. *Physica Medica*, *90*, pp.13-22.

[11] Yu, K.H., Lee, T.L.M., Yen, M.H., Kou, S.C., Rosen, B., Chiang, J.H. and Kohane, I.S., 2020. Reproducible machine learning methods for lung cancer detection using computed tomography images: Algorithm development and validation. *Journal of medical Internet research*, *22*(8), p.e16709.

[12] Dritsas, E. and Trigka, M., 2022. Lung cancer risk prediction with machine learning models. *Big Data and Cognitive Computing*, *6*(4), p.139.

[13] Abdullah, D.M., Abdulazeez, A.M. and Sallow, A.B., 2021. Lung cancer prediction and classification based on correlation selection method using machine learning techniques. *Qubahan Academic Journal*, *1*(2), pp.141-149.

[14] Li, Y., Wu, X., Yang, P., Jiang, G. and Luo, Y., 2022. Machine learning for lung cancer diagnosis, treatment, and prognosis. *Genomics, Proteomics and Bioinformatics*, *20*(5), pp.850-866.

[15] Heuvelmans, M.A., van Ooijen, P.M., Ather, S., Silva, C.F., Han, D., Heussel, C.P., Hickes, W., Kauczor, H.U., Novotny, P., Peschl, H. and Rook, M., 2021. Lung cancer prediction by Deep Learning to identify benign lung nodules. *Lung Cancer*, *154*, pp.1-4.

[16] Perez, G. and Arbelaez, P., 2020. Automated lung cancer diagnosis using three-dimensional convolutional neural networks. *Medical & Biological Engineering & Computing*, *58*, pp.1803-1815.

[17] Kobylińska, K., Orłowski, T., Adamek, M. and Biecek, P., 2022. Explainable machine learning for lung cancer screening models. *Applied Sciences*, *12*(4), p.1926.

[18] Yang, H., Chen, L., Cheng, Z., Yang, M., Wang, J., Lin, C., Wang, Y., Huang, L., Chen, Y., Peng, S. and Ke, Z., 2021. Deep learning-based six-type classifier for lung cancer and mimics from histopathological whole slide images: a retrospective study. *BMC medicine*, *19*, pp.1-14.

[19] Pradhan, K. and Chawla, P., 2020. Medical Internet of things using machine learning algorithms for lung cancer detection. *Journal of Management Analytics*, *7*(4), pp.591-623.

[20] Sori, W.J., Feng, J., Godana, A.W., Liu, S. and Gelmecha, D.J., 2021. DFD-Net: lung cancer detection from denoised CT scan image using deep learning. *Frontiers of Computer Science*, *15*, pp.1-13.

[21] Xie, Y., Meng, W.Y., Li, R.Z., Wang, Y.W., Qian, X., Chan, C., Yu, Z.F., Fan, X.X., Pan, H.D., Xie, C. and Wu, Q.B., 2021. Early lung cancer diagnostic biomarker discovery by machine learning methods. *Translational oncology*, *14*(1), p.100907.

[22] Yu, H., Zhou, Z. and Wang, Q., 2020. Deep learning assisted predict of lung cancer on computed tomography images using the adaptive hierarchical heuristic mathematical model. *IEEE Access*, *8*, pp.86400-86410.

[23] Han, Y., Ma, Y., Wu, Z., Zhang, F., Zheng, D., Liu, X., Tao, L., Liang, Z., Yang, Z., Li, X. and Huang, J., 2021. Histologic subtype classification of non-small cell lung cancer using PET/CT images. *European journal of nuclear medicine and molecular imaging*, *48*, pp.350-360.

