VER
Vascular & Endovascular Review

Lung Cancer Prediction Using Machine Learning: A Comparative Analysis Of Knn, Svm, Random Forest

Gajula Hema Sandhya[1], Pilli. Veeraswamy[2]

[1]PG STUDENT IN DEPT OF CSE IN SREE VAHINI INSTITUTE OF SCIENCE & TECHNOLOGY, TIRUVURU, ANDHRA PRADESH 521235.
[2]ASSISTANT PROFESSOR IN DEPT OF CSE IN SREE VAHINI INSTITUTE OF SCIENCE & TECHNOLOGY, TIRUVURU, ANDHRA PRADESH 521235.

ABSTRACT

Lung cancer is one of the leading causes of cancer-related deaths worldwide, and the likelihood of survival is greatly affected by how quickly and accurately the disease is detected. Although they work, traditional methods of diagnosing lung cancer often fail to identify the disease in its early stages. It is comforting to know that machine learning may improve lung cancer prognosis by sifting through complex patterns in medical data. The effectiveness of machine learning models is, however, dependent on the algorithms and optimisation techniques used. The purpose of this research is to examine and compare four machine learning methods—Random Forest, Logistic Regression, K-Nearest Neighbours (KNN), and Support Vector Machine—in order to forecast the occurrence of lung cancer. The Kaggle dataset was subjected to preprocessing, encoding, and feature selection processes in order to enhance model performance. The model parameters were fine-tuned using hyperparameter tuning in order to achieve an even higher level of accuracy. In order to assess the models, important performance metrics including as accuracy, precision, recall, and F1-score were used. While other models showed varying degrees of performance, the results show that the Logistic Regression technique performed best with a 90% accuracy rate. The results show that machine learning has potential for lung cancer prediction, and that model assortment and parameter optimisation are important. To improve predicted accuracy, future studies may investigate deep learning methods and use more patient data. In the end, using machine learning to diagnose lung cancer might result in earlier detection, better long-term effects, and a dramatic reduction in mortality rates.

KEYWORDS: Lung cancer, Machine learning, Random Forest, Logistic Regression, Hyperparameter tuning

How to Cite: Gajula Hema Sandhya, Pilli. Veeraswamy, (2025) Lung Cancer Prediction Using Machine Learning: A Comparative Analysis Of Knn, Svm, Random Forest, Vascular and Endovascular Review, Vol.8, No.5s, 419-427.

INTRODUCTION

The lungs are vital respiratory organs that play a key role in gas exchange, namely in exhaling carbon dioxide and inhaling oxygen, which is necessary for cellular metabolism [1]. Smoking, air pollution, and carcinogen exposure are increasing deaths from lung cancer, a malignant growth of the lung tissue, which is one of the most deadly illnesses globally. Chronic cough, haemoptysis, chest discomfort, and exhaustion are symptoms that often don't appear until the illness has progressed further [2]. The ability for electronic equipment to learn from data automatically is known as machine learning (ML) [3]. ML is a subset of artificial intelligence. From medical diagnostics to personalised recommendation systems, it has found revolutionary uses across fields. The analysis of high-dimensional datasets and the extraction of non-trivial patterns for predictive modelling are two areas where ML has shown particularly encouraging results in the healthcare industry [4].

Machine learning methods using K-Nearest Neighbours (KNN), Support Vector Machines (SVM), Random Forest Classifier (RFC), and Logistic Regression (LR) for disease classification applications, such as cancer prediction were used. The trade-offs between interpretability, computational efficiency, and classification performance are unique to each approach.

1.KNN is a straightforward and flexible method that classifies new data points according to the common vote of the adjacent neighbors in feature space [4] [5].
2.SVM looks for the best hyperplane to separate classes, demonstrating high accuracy in jobs involving binary categorization, including identifying instances that are malignant and those that are not [6].
3.A variety of decision trees are built using Random Forest, an ensemble learning approach, to increase resilience and reduce overfitting [7].
4.A sigmoid function is employed in logistic regression to represent the probability of a binary outcome and is favored for its interpretability in clinical settings [8].

In spite of these models' successes, there are still obstacles to overcome, such as getting good annotated datasets, tweaking hyperparameters, dealing with overfitting, and integrating models into clinical processes [8]. In order to evaluate the predictive performance of KNN, SVM, RFC, and LR, this research compares them using publically accessible lung cancer datasets. The performance is measured using metrics like exactness, correctness, precision, recall, and F1-score.

The long-term goal is to evaluate the practicality and efficacy of machine learning methods for detecting lung cancer at an early stage in order to enhance data-driven decision-making and patient outcomes.

The Lungs' Anatomy and Function

The thoracic cavity contains the lungs, which are spongy, cone-shaped organs. Because of the heart's restriction of space, the left lung has two lobes while the right has three [1]. They perform gas exchange approximately 12–20 times per minute, a process crucial to sustaining life. Protective mechanisms such as nasal hairs, mucus lining the airways, and the sweeping motion of cilia work collectively to filter airborne pollutants [1].

Lung Cancer: Overview and Classification

Lung cancer arises from epithelial cells and is separatedinto two keyclasses: small cell lung cancer (SCLC) and non- small cell lung cancer (NSCLC). NSCLC, which comprises big cell carcinoma, squamous cell carcinoma, and adenocarcinoma, makes up around 85% of cases. More aggressive, SCLC grows and spreads quickly, and is usually presented as restricted or vast [2][9].
Major risk factors include tobacco use (accounting for ~90% of cases), secondhand smoke, occupational exposure (e.g., asbestos, arsenic), air pollution, genetic predisposition, and emerging risks like vaping [2][10]. Notably, early-stage lung cancer often lacks symptoms, making early detection critical to improving survival, which remains around 19% overall [2].

Predictive Analytics in Medical Diagnosis

Using both historical and current data, predictive analytics makes predictions about the future outcomes using statistical models, ML, and AI [11]. In oncology, predictive modelling can identify patients at high risk, guide diagnostic testing, and inform personalized treatment strategies.

Machine Learning Paradigms

ML algorithms are categorized into four main paradigms:
1.Supervised Learning: Uses labelled data to train classifiers (e.g., SVM, decision trees, KNN, LR) for tasks like cancer diagnosis [12].
2.Unsupervised Learning: Uses methods like dimensionality reduction and clustering to find patterns in unlabeled data [13].
3.Semi-Supervised Learning: To increase learning efficiency, a small labeled dataset is combined with a larger unlabelled sample [14].
4.Reinforcement Learning: Involves learning optimal actions through feedback mechanisms, commonly used in robotics and adaptive systems [12].

review of Related Works

Attempt[14] to compare the efficacy of several traditional machine learning techniques in detecting lung cancer from clinical data, including Support Vector Machines (SVMs), Decision Trees, and Random Forests. A dataset including patient demographics, clinical features, and histopathology information was used by the researchers. They improved the algorithms' performance by using a systematic approach to feature selection. In order to ensure the reliability of the results, the study included extensive testing and verification. With a sensitivity level of 90% and an accuracy of 92%, SVM clearly performed better than the other algorithms. Thorough data preparation may substantially affect the outcomes of machine learning in healthcare, as the study shown, and it is crucial to choose the correct features to improve model performance.

[15] In order to identify lung cancer in radiological images, such as CT scans and chest X-rays, [25] investigated the use of Convolutional Neural Networks (CNNs). As part of the study, a CNN was trained to recognise intricate patterns associated with lung cancer by training a deep learning model that makes use of a large dataset of labelled visual data. By comparing the CNN's performance to that of more traditional diagnostic methods, the researchers proved that deep learning may be useful for medical imaging. The CNN algorithm significantly surpassed the conventional approaches, with an impressive 97% accuracy rate. With the ability to automate image processing, deep learning approaches may reduce radiologists' burden and improve diagnosis accuracy, according to the authors. The results of this study show that deep learning may revolutionise the identification of lung cancer, marking a significant step forward in the use of AI in medical imaging.

In their study, Liu and colleagues compared traditional machine learning algorithms for lung cancer detection in CT scans with CNNs and other deep learning models [16], [17]. An extensive evaluation of the model's performance was carried out by the researchers using a dataset that included various instances of lung cancer. In order to maximise the model's efficiency and accuracy, many CNN architectures were evaluated to determine the optimal configuration for this specific job. With an accuracy rate of 95%, the results showed that CNN models routinely outperformed more conventional methods. The study highlighted the advantages of deep learning over conventional approaches for extracting complex characteristics from imaging data. Deep learning has the potential to significantly enhance radiology's capacity to detect lung cancer, according to the researchers.

Random Forests, K-Nearest Neighbours (KNN), and Logistic Regression were among the machine learning methods that Zhang and colleagues thoroughly assessed using a dataset of lung cancer patients [18], [19]. Finding the most efficient way to use clinical and imaging data for early detection of lung cancer was the primary goal of the research. To make sure their findings were solid and to avoid overfitting, the researchers used cross-validation techniques. Random Forests demonstrated a strong mix of recall and precision, as the research discovered that it obtained an accuracy rate of 94%. Based on their capacity to manage complicated datasets and prevent overfitting, ensemble processes like Random Forests have shown to be very effective in medical diagnostics. As a result, these procedures have practical applications in diagnosing lung cancer.

The methodological merits of the research summarised in Table 1—which include radiomics-based techniques, high accuracy, effective use of CNNs, EHR integration, and transfer learning—were highlighted in relation to lung cancer prediction. On the other hand, it draws attention to typical flaws, such as small datasets, reliance on poor data quality, difficulty in understanding results, and difficulties in applying findings to other areas.

Table 1. Strengths and Weaknesses of the current systems

Study	Strengths	Weaknesses
Ausawalaithong et al. (2019)	High accuracy (93.5%) with CNNs on X-rays	Limited dataset, black-box interpretability [20]
Yeh et al. (2020)	Uses EHR for early risk prediction (90.2%)	Dependent on EHR quality and completeness [21]
Islam et al. (2019)	Effective use of transfer learning (95.1%)	May not generalize well to new domains [22]
Li et al. (2020)	Combines radiomics and ML (92.5%)	Limited by radiomics data extraction [23]
Wang et al. (2019)	CNN on CT scans with strong results (94.2%)	Dataset size limits scalability [24]

These findings underscore the growing success of ML, particularly deep learning, in identifying lung cancer from complex clinical and imaging data. However, challenges such as interpretability, generalizability, and integration into clinical workflows remain open research questions.

METHODOLOGY

Introduction

This chapter presents the methodology for creating an artificial intelligence-based (ML) lung cancer prediction system. Emphasis is placed on the design of the dataset pipeline, preprocessing procedures, classifier architecture, performance evaluation, and tool selection. Ensuring validity and reliability, the section lays the groundwork for replicability and future extension.

Proposed System

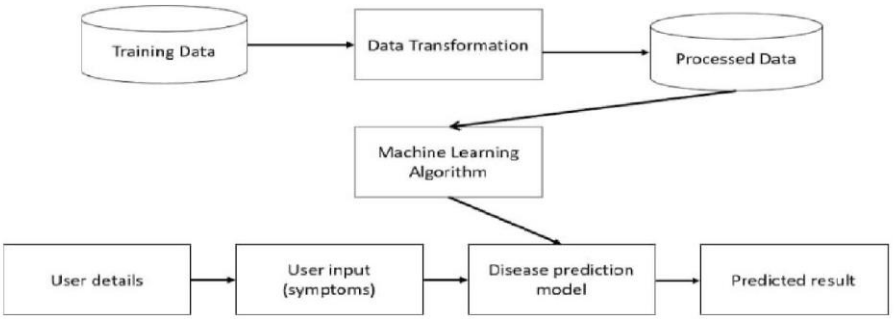


Fig. 1. System Architecture of Lung Disease Prediction

A comparative study is conducted using four supervised classifiers: Support Vector Machine (SVM), Random Forest Classifier (RFC), K-Nearest Neighbours (KNN), and Logistic Regression (LR), to forecast lung cancer risk. Input features include age, smoking habit, air pollution exposure, genetic predisposition, symptoms, and biomarker indicators. Models are trained to classify patients into risk categories: Low, Medium, or High.

Prior to model training, data undergoes comprehensive preprocessing: cleaning, imputing, normalization, feature selection, and class balancing. Hyperparameter tuning (e.g., SVM kernel, number of trees for RFC, optimal k in KNN) is performed via grid search and k-fold cross-checking. Measures of performance include precision, accuracy, recall, and F1-score.

Discussion of Dataset

Dataset Description and Collection

The dataset originates from a publicly available Kaggle repository called "Lung Cancer Risk & Prediction Dataset,"[29] which contains approximately 1,000 records and 25 attributes, including demographics, lifestyle, environmental exposure, symptoms, and a three-level risk target variable (Low, Medium, High).

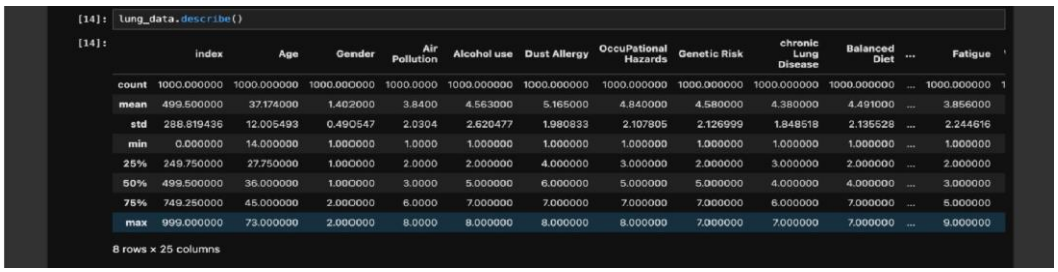
Features utilized include age, gender, air pollution index, smoking and passive smoking, occupational hazards, genetic risk,

chronic respiratory disease history, and clinical symptoms (chest pain, fatigue, weight loss, etc.). The target variable indicates lung cancer risk level. Dataset labels and features were extracted in a manner consistent with other published studies.

Table 2: Description of Dataset Features for Lung Cancer Prediction

S/N	Feature	Description
1	Index	Unique ID of records
2	Age	Patient's age
3	Gender	Patient's gender (1 = Male, 2 = Female)
4	Air Pollution	Exposure level to air pollutants (scale of 1–10)
5	Alcohol Use	Frequency of alcohol consumption (scale of 1–10)
6	Dust Allergy	Sensitivity to dust allergens (scale of 1–10)
7	Occupational Hazards	Exposure to hazardous work conditions (scale of 1–10)
8	Genetic Risk	Family history of lung cancer (scale of 1–10)
9	Chronic Lung Disease	Presence of chronic lung diseases (scale of 1–10)
10	Balanced Diet	Quality of diet (scale of 1–10)
11	Overweightness	Obesity level (scale of 1–10)
12	Smoking	Smoking frequency (scale of 1–10)
13	Passive Smoker	Exposure to secondhand smoke (scale of 1–10)
14	Chest Ache	Intensity of chest pain (scale of 1–10)
15	Coughing of Blood	Frequency of coughing up blood (scale of 1–10)
16	Exhaustion	Level of fatigue (scale of 1–10)
17	Weight Loss	Extent of weight loss (scale of 1–10)
18	Shortness of Breath	Severity of breathlessness (scale of 1–10)
19	Gasping	Intensity of wheezing (scale of 1–10)
20	Swallowing Difficulty	Difficulty in swallowing (scale of 1–10)
21	Clubbing of Finger Nails	Changes in nail appearance (scale of 1–10)
22	Frequent Cold	Frequency of catching colds (scale of 1–10)
23	Dry Cough	Severity of dry cough (scale of 1–10)
24	Snoring	Frequency of snoring (scale of 1–10)
25	Level	Target variable (Low, Medium, High – indicating lung cancer risk)

Figure 3 shows the statistical summary of a lung cancer dataset, containing the lowest, quartiles, maximum, count, mean, and standard deviation values for every aspect. Figure 4 indicates that the dataset has no missing values, as all features show a count of zero null entries.



```
[14]: lung_data.describe()
```

[14]:

	index	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chronic Lung Disease	Balanced Diet	...	Fatigue
count	1000.000000	1000.000000	1000.000000	1000.0000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	...	1000.000000
mean	499.500000	37.174000	1.402000	3.8400	4.563000	5.165000	4.840000	4.580000	4.380000	4.491000	...	3.856000
std	288.819436	12.005493	0.490547	2.0304	2.620477	1.980833	2.107805	2.126999	1.848518	2.135528	...	2.244616
min	0.000000	14.000000	1.000000	1.0000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	...	1.000000
25%	249.750000	27.750000	1.000000	2.0000	2.000000	4.000000	3.000000	2.000000	3.000000	2.000000	...	2.000000
50%	499.500000	36.000000	1.000000	3.0000	5.000000	6.000000	5.000000	5.000000	4.000000	4.000000	...	3.000000
75%	749.250000	45.000000	2.000000	6.0000	7.000000	7.000000	7.000000	7.000000	6.000000	7.000000	...	5.000000
max	999.000000	73.000000	2.000000	8.0000	8.000000	8.000000	8.000000	7.000000	7.000000	7.000000	...	9.000000

8 rows × 25 columns

Fig. 3. Describing the dataset

```
[10]: lung_data.isnull().sum()

[10]: index                      0
      Age                        0
      Gender                     0
      Air Pollution              0
      Alcohol use                0
      Dust Allergy               0
      OccuPational Hazards       0
      Genetic Risk               0
      chronic Lung Disease       0
      Balanced Diet              0
      Obesity                    0
      Smoking                    0
      Passive Smoker             0
      Chest Pain                 0
      Coughing of Blood          0
      Fatigue                    0
      Weight Loss                0
      Shortness of Breath        0
      Wheezing                   0
      Swallowing Difficulty      0
      Clubbing of Finger Nails   0
      Frequent Cold              0
      Dry Cough                  0
      Snoring                    0
      Level                      0
      dtype: int64
```

Data Pre-processing

- Data Cleaning: Missing or invalid entries were handled via imputation or removal as appropriate.
- Standardization: Feature scaling was performed using scikit-learn's StandardScaler, ensuring zero mean and unit variance.
- Feature Selection: Key predictors were identified using scikit-learn's SelectKBest with chi-squared scoring to reduce dimensionality and enhance interpretability.
- In order to discourse class Imbalance: The Synthetic Minority Over-Sampling Technique (SMOTE) remained applied via the imbalanced-learn package to balance class distribution and prevent bias toward majority classes.
- Train-Test Split: Using scikit-learn's train_test_split function, the cleaned dataset was separated into subgroups for training (80%) and testing (20%).

Model Selection and Training

Each algorithm employed is briefly described below and trained on the preprocessed dataset:
Support Vector Machine (SVM): Constructs a hyperplane to maximize class separation margin. Kernel and regularization parameters were tuned.
Logistic Regression (LR): Predicts probabilities via the sigmoid function. Implemented using the 'lbfgs' solver with regularization parameter C tuning.
Using the Manhattan or Euclidean distance metric, K-Nearest Neighbors (KNN) is a non-parametric classifier it forecasts outcomes by using the majority label among k neighbors. Grid search was used to get the ideal k.
Random ForestClassifier (RFC): A bootstrap-aggregated ensemble of decision trees. In tune among the options were min_samples_split, max_depth, and multiple estimators (n_estimators).

Hyperparameter Tuning

Each machine learning algorithm's hyperparameters were adjusted to maximize its performance in lung cancer prediction. The specific hyperparameters tuned for each model included:
1.Support Vector Machine (SVM): In order to maximize the model's ability to distinguish between lung cancer risk levels, the regulationconstraint (C) and the kernel type (example., linear, polynomial, RBF) were changed.
2.K-Nearest Neighbors (KNN): To increase the precision of categorizing whereas balancing bias and variance, the number of neighbors (K) and the distance metric (such as Manhattan or Euclidean) were adjusted.
3.Random Forest Classifier: To increase model performance, decrease overfitting, and improve generalization in lung cancer classification, the number of decision trees (n_estimators), maximum tree depth, and minimum number of samples needed for node splitting were tuned.
4.Logistic Regression: To guarantee that the model performed effectively when applied to unknown data, the regularization parameter (C) was adjusted to prevent overfitting.
Hyperparameter tuning's objective was to achieve the highest possible accuracy by tailoring each model to the unique characteristics of lung cancer risk factors. This approach ensured a well-optimized model setup, maximizing predictive performance in lung cancer classification

System Design

Figure 2 outlines the sequential workflow of a disease prediction system, starting from entering patient details to validating data, extracting features, and matching values. It then proceeds to classify the data, predict the disease, and finally display the results.

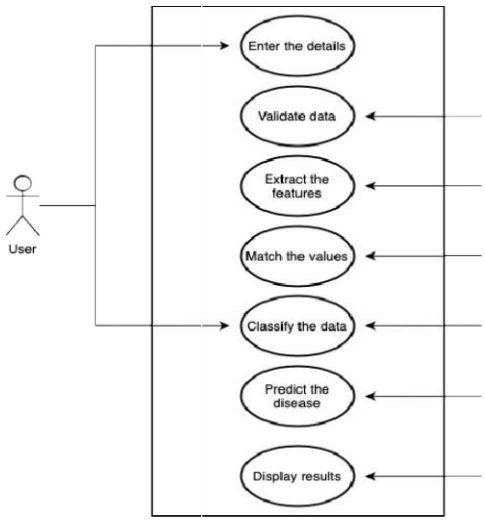


Figure 2. Illustrates the Use Case Diagram, showing user interactions, data input, analytics processing, and output classification.

RESULTS AND FINDINGS

A. Result of Model Training

This project trained four machine learning models: Random Forest Classifier, K-Nearest-Neighbors, Support Vector Machine, and Logistic Regression Classifier. An 80-20 dataset split was used to train the logistic regression model. The training data allowed the modelto identify patterns and categorize new occurrencesefficiently. Its prediction skills were measuredby means of a range of performance pointers. The KNN model was trained using 80% of the dataset, with the remaining 20% set aside for testing. The classifier used 3 nearest neighbors and the Minkowski distance (p=2).

Since KNN is a non-parametric model, instead of learning, commits the training data to memory explicit outlines. Predictions were made by assigning new data points to their closest neighbor's predominant class. The SVM model was trained by means of an 80-20 train-examination split with Bagging to improve generalization. It utilized an RBF kernel with C=7 and trained 75 SVM classifiers about various data subsets. The One-vs-Rest strategy handled multi-class classification, and the model's performance was evaluated by means of a confusion matrix and accuracy score on the test set.

The Random Forest model was trained on a preprocessed dataset, where categorical variables were encoded, and highly correlated features (Pearson > 0.9) were removed. An 80-20 train-test split was used with stratification to maintain class balance. To reduce overfitting, the model was fine-tuned by:

1. Limiting tree depth and number of trees
2. Increasing the minimum samples required for splits and leaf nodes
3. Using fewer features per tree and reducing sample size per tree
4. Applying class weighting to handle class imbalance

A. Performance Evaluation

Table 2: Classification Report Logistic Regression

Class	Precision	Recall	F-1 Score	Support
1	0.92	0.84	0.88	67
2	0.81	0.88	0.84	58
3	0.95	0.96	0.95	95
Overall Accuracy: 0.90	Macro Avg: 0.89	Weighted Avg: 0.90		

Logistic Regression had high accuracy and was particularly effective in predicting class 3. However, it showed slightly lower recall for class 1, which means some high-risk patients may have been misclassified.

K-Nearest Neighbors (KNN): The KNN model attained an accuracy of 0.86, performing best for class 1 with a recall of 0.97, meaning it correctly identified most patients in this category. However, it struggled with class 2, where recall dropped to 0.69, leading to a higher number of false negatives. The weighted F1-score of 0.86 recommends that while the model is fairly balanced, its classification of class 2 could be improved. Table 3 presents precision, recall, F1-score, and support for each class, showing an overall model accuracy of 86%.

Table 3: Classification Report KNN

Class	Precision	Recall	F-1 Score	Support
1	0.83	0.97	0.90	67
2	0.89	0.61	0.78	58
3	0.88	0.91	0.89	75
Overall Accuracy: 0.86	Macro Avg: 0.87	Weighted Avg: 0.86		

KNN was particularly strong in identifying class 1 but struggled with class 2. The high recall for class 1 suggests that KNN can effectively detect early-stage lung cancer cases but may require tuning for better performance across all classes.

Support Vector Machine (SVM): SVM demonstrated an accuracy of 0.88, showing balanced performance across all classes. It performed exceptionally well for class 3, achieving a recall of 1.00, meaning it correctly identified all patients in this category.

However, its performance for class 2 was slightly weaker, with a recall of 0.69. The weighted F1-score of 0.88 indicates a well-rounded model. Table 4 presents precision, recall, F1-score, and support for each class, showing an overall model accuracy of 88%.

Table 4: Classification Report SVM

Class	Precision	Recall	F-1 Score	Support
1	0.94	0.91	0.92	67
2	1.00	0.69	0.82	58
3	0.79	1.00	0.88	75
Overall Accuracy:0.88	Macro Avg: 0.91	Weighted Avg: 0.88		

SVM had a perfect recall for class 3, making it ideal for identifying confirmed lung cancer cases. However, its lower recall for class 2 suggests potential improvements in fine-tuning the model parameters.

Random Forest Classifier: The Random Forest classifier achieved an accuracy of 0.83. It performed best for class 0, with a recall of 1.00, meaning all patients in this category were correctly identified. However, its recall for class 2 was 0.53, indicating a high number of false negatives. The weighted F1-score of 0.82 suggests that while the model is strong in certain areas, it struggles with class 2. Table 5 presents precision, recall, F1-score, and support for each class, showing an overall model accuracy of 83%.

Table 5: Classification Report Random Forest

Class	Precision	Recall	F-1 Score	Support
1	0.84	1.00	0.91	73
2	0.76	0.97	0.85	61
3	1.00	0.53	0.69	66
Overall Accuracy: 0.83	Macro Avg: 0.87	Weighted Avg: 0.82		

Class 2 had lesser recall for Random Forest, indicating it failed to accurately identify a substantial percentage of patients, in contrast to class 0, where it performed quite well in terms of prediction. This indicates that it could benefit from feature selection or more data balance methods to enhance its performance.

In conclusion, Logistic Regression outperformed all other models with respect to accuracy (90%), indicating great performance across the board for lung cancer risk levels. Its excellent recall and F1-score made it a dependable option for detecting lung cancer patients, and it was especially good at predicting class 3 cases.

The detection of advanced lung cancer patients (class 3) was an area where SVM excelled, with a recall of 1.00. A decrease in recall for class 2 was an indication of misclassification, nevertheless.

Class 1 lung cancer patients were best identified by KNN, which had a recall of 0.97. Class 2 had a decline in performance and an increase in false negatives as a result.

With a recall of 1.00, Random Forest proved to be highly predictive for class 0. Class 2 was its weak spot, with a recall of only 0.53; this indicates that it needs further tweaking or feature selection to perform better.

Due to its balanced recall, high accuracy, and precision, Logistic Regression emerged as the best-performing model overall. Both SVM and KNN demonstrated excellent prediction capabilities, especially for distinct stages of lung cancer. Although it is robust, Random Forest's classification performance might be improved using hyperparameter tweaking and data balance approaches.

SUMMARY AND CONCLUSION

Summary

This study compared and contrasted four machine learning algorithms—Support Vector Machines (SVM), Random Forest Classifier, K-Nearest Neighbours (KNN), and Logistic Regression—in terms of their ability to predict the risk of lung cancer.

Data acquisition, preprocessing, feature selection, model training, hyperparameter adjustment, and performance evaluation using accuracy, precision, recall, and F1-score metrics were all part of the systematic methodology followed in the research, which made use of a Kaggle dataset.

Among the models evaluated, Logistic Regression stood out as the best practical algorithm for lung cancer prediction due to its exceptional accuracy and well-rounded performance. We learnt a lot about the algorithms' practical usefulness for early lung cancer diagnosis from their individual strengths and shortcomings. This research focusses on the potential of machine learning methods to improve diagnostic accuracy and clinical decision-making in the context of lung cancer patients.

LIMITATIONS

This research has limitations, despite the optimistic findings. To begin with, the results may not be applicable to a broader population since the dataset came from Kaggle and may not necessarily reflect all incidences of lung cancer across all demographics and locations. Second, the characteristics included in the research were small, so there's a chance that crucial clinical or genetic variables were overlooked, which might have an impact on the accuracy and resilience of the models. Further validation and greater feature integration are needed for real-world implementation, according to these limits.

FUTURE WORK

Research into more advanced approaches, such as deep learning using Artificial Neural Networks (ANNs), might lead to improved lung cancer prediction in the future by better capturing patterns in complicated medical data. Improving the generalisability and accuracy of models might be achieved by expanding datasets to include more varied populations as well as new genetic and clinical variables. Furthermore, by combining clinical decision support systems with predictive models, it is possible to achieve real-time diagnosis. This would allow for earlier intervention and ultimately lead to better patient outcomes.

CONCLUSION

Using publicly available medical data, this study demonstrates how machine learning might aid in the early detection of lung cancer. In terms of lung cancer prediction, the best-performing model was Logistic Regression, demonstrating its appropriateness for classification tasks. Although there are certain limits, the results show how AI might revolutionise healthcare and build the groundwork for better diagnostic tools in the future. Machine learning models may greatly improve lung cancer diagnosis and patient treatment with further study and improvement.

.

REFERENCES

1. Cleveland Clinic, “Lungs: Location, Anatomy, Function & Complications,” Cleveland Clinic — Health Library. [Online]. Available: https://my.clevelandclinic.org/health/body/8960-lungs. (Verified)
2. American Cancer Society, “Key Statistics for Lung Cancer,” 2022. [Online]. Available: https://www.cancer.org/cancer/types/lung-cancer/about/key-statistics.html. (Verified)
3. A. L. Samuel, “Some Studies in Machine Learning Using the Game of Checkers,” IBM Journal of Research and Development, vol. 3, no. 3, pp. 210–229,1959.
4. E. E. Onuiri, O. Ukandu, and K. Umeaka, “International Journal of Research Publication and Reviews Machine Learning Models for Lung Cancer Subtype Classification: A Systematic Review,” International Journal of Research Publication and Reviews, vol. 5, no. 9, pp. 1299–1308, 2024.
5. T. M. Cover and P. E. Hart, “Nearest Neighbor Pattern Classification,” IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27, Jan. 1967, doi: 10.1109/TIT.1967.1053964.
6. C. Cortes and V. Vapnik, “Support-Vector Networks,” Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
7. L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf. (Verified)
8. D. W. Hosmer and S. Lemeshow, Applied Logistic Regression, 2nd ed. New York: Wiley, 2000.
9. S. Garg, P. Pundir, G. Rathee, P. K. Gupta, S. Garg, and S. Ahlawat, “On Continuous Integration / Continuous Delivery for Automated Deployment of Machine Learning Models using MLOps,” arXiv, preprint arXiv:2202.03541, Feb. 2022. [Online]. Available: https://arxiv.org/abs/2202.03541. (URL added)
10. W. D. Travis, E. Brambilla, A. P. Burke, A. Marx, and A. G. Nicholson, “The 2015 World Health Organization Classification of Lung Tumors,” Journal of Thoracic Oncology, vol. 10, no. 9, pp. 1243–1260, Sep. 2015.
11. National Cancer Institute, “Lung Cancer Prevention,” PDQ® Cancer Information Summaries. [Online]. Available: https://www.cancer.gov/types/lung/hp/lung-prevention-pdq. Accessed: Aug. 9, 2025. (Verified)
12. G. Shmueli and O. R. Koppius, “Predictive Analytics in Information Systems Research,” MIS Quarterly, vol. 35, no. 3, pp. 553–572,2011.
13. H. Sutton, “Peter Morgan Sutton,” BMJ, vol. 348, no. mar31 11, pp. g2466-g2466, Mar. 2014, doi: 10.1136/bmj.g2466.
14. A. Kumar, S. Singh, and P. Arora, “Comparative Analysis of Machine Learning Algorithms for Lung Cancer Detection,” Procedia Computer Science, vol. 132, pp. 556–563,2018.
15. G. Cai et al., “Medical AI for Early Detection of Lung Cancer: A Survey,” arXiv, preprint arXiv:2410.14769, Oct. 2024. [Online]. Available: https://arxiv.org/abs/2410.14769. (URL added)
16. A. Esteva et al., “Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks,” Nature, vol. 542, no. 7639, pp. 115–118,Feb. 2017, doi: 10.1038/nature21056.
17. Y. Zhang, L. Jiang, and W. Li, “Evaluation of Machine Learning Methods for Lung Cancer Detection Using ClinicalData,” IEEE Access, vol. 7, pp. 109960–109968, 2019.

18. A. B. Goldberg, “SSL with Realistic Tuning,” University of Wisconsin–Madison, Computer Sciences, Tech. Rep., 2009.
19. A. Chaudhari, A. Singh, S. Gajbhiye, and P. Agrawal, “Lung Cancer Detection using Deep Learning,” arXiv, preprint arXiv:2501.07197, Jan. 2025. [Online]. Available: https://arxiv.org/abs/2501.07197. (URL added)
20. W. Ausawalaithong, S. Marukatat, A. Thirach, and T. Wilaiprasitporn, “Automatic Lung Cancer Prediction from Chest X-ray Images Using Deep Learning Approach,” arXiv, preprint arXiv:1808.10858, Aug. 2018. [Online]. Available:https://arxiv.org/abs/1808.10858. (URL added)
21. M. C. H. Yeh et al., “Artificial Intelligence–Based Prediction of Lung Cancer Risk Using Nonimaging Electronic Medical Records: Deep Learning Approach,” Journal of Medical Internet Research, vol. 23, no. 8, e26256, Aug. 2021, doi:10.2196/26256.
22. M. I. Islam et al., “VER-Net: a hybrid transfer learning model for lung cancer detection using CT scan images,” BMC Medical Imaging, vol. 24, Art. no. 98, May 2024.
23. J. Li, Z. Li, L. Wei, and X. Zhang, “Machine Learning in Lung Cancer Radiomics,” Machine Intelligence Research, vol. 20, no. 6, pp. 753–782, 2023, doi: 10.1007/s11633-022-1364-x.
24. J. Wang et al., “Lung Cancer Detection Using Co-learning from Chest CT Images and Clinical Demographics,” arXiv, preprint arXiv:1902.08236, Feb. 2019. [Online]. Available: https://arxiv.org/abs/1902.08236. (URL added)
25. A. Kumar, S. Singh, and P. Arora, “Comparative Analysis of Machine Learning Algorithms for Lung Cancer Detection,” Procedia Computer Science, vol. 132, pp. 556–563, 2018. (Same as [14])
26. T. M. Cover and P. E. Hart, “Nearest Neighbor Pattern Classification,” IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27, Jan. 1967, doi: 10.1109/TIT.1967.1053964. (Same as [5])
27. T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. Springer, 2009.
28. A. B. Goldberg, “SSL with Realistic Tuning,” University of Wisconsin–Madison, Computer Sciences, Tech. Rep., 2009. (Same as [18])
29. "Lung Cancer Risk & Prediction Dataset," Kaggle, Accessed: Sep. 18, 2025. [Online]. Available: https://www.kaggle.com/datasets/ankushpanday1/lung-cancer-risk-and-prediction-dataset