International Journal of Innovative Research and Scientific Studies, 8(1) 2025, pages: 2671-2679

IJIRSS
International Journal of Innovative Research & Scientific Studies

ISSN: 2617-6548

URL: www.ijirss.com

Non-image lung cancer prediction utilizing KNN model promoting health consciousness

Ting Tin Tin[1*], Chow Mun Chun[2], Lim Yong Zhe[3], Hii Puong Tih[4], Chan Man Tze[5]

[1]*Faculty of Data Science and Information Technology, INTI International University, 71800 Nilai, Negeri Sembilan, Malaysia.*
[2,3,4,5]*Faculty of Computing and Information Technology, Tunku Abdul Rahman University of Management and Technology, Malaysia.*

Corresponding author: Ting Tin Tin (*Email: tintin.ting@newinti.edu.my*)

Abstract

One of the most prevalent types of cancer worldwide, lung cancer, has a high mortality rate. Imaging sets are commonly used for lung cancer diagnosis. However, imaging sets have limitations in terms of accuracy that can cause false negative cases, which leads to delayed treatment due to late diagnosis, ultimately reducing patients' survival rates. Societies with low income might need a more economical way to predict lung cancer since the imaging sets require a significant amount of money. This study uses a secondary dataset that contains non-image data such as demographics, lifestyle, and symptoms to create a model to detect lung cancer. The performance of several machine learning models, including SVM, Decision Tree, ANN, Logistic Regression, Random Forest, XGBoost, AdaBoost, Gradient Boost, Light Gradient Boosting, KNN, and Naive Bayes, is compared after the dataset has been preprocessed and divided into training and testing data. It is found that lung cancer is more likely to be diagnosed in females and those with any allergies, alcohol consumption, or difficulty swallowing. Next, it is shown that the KNN model is the best model, with an accuracy of 96.39%, a precision score of 100%, and an F1-score of 97.81%, despite having the lowest recall score among other models. A successful prediction model eases the burden on low-income families to predict the possibility of disease occurrence without spending money on X-rays, thus increasing health consciousness.

Keywords: Data analysis, Healthcare, Lung cancer detection, Machine learning, Non-image data.

DOI: 10.53894/ijirss.v8i1.5040
Funding: This study received no specific financial support.
History: Received: 3 January 2025 / **Revised**: 7 February 2025 / **Accepted:** 16 February 2025 / **Published:** 28 February 2025

Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Publisher: Innovative Research Publishing

1. Introduction

The second-leading cause of death worldwide is cancer with the most common cancer diagnosed being lung cancer, breast cancer, and prostate cancer. Among these cancers, lung cancer has the highest mortality rate with a total of 1.762 million deaths caused by lung cancer, with men having the highest rate than women [1]. Although smoking can lead to lung

cancer, a study states that smoking is not the sole cause of lung cancer but other risk factors like radon exposure, secondhand smoke, occupational exposure, and infections can lead to lung cancer for non-smokers Corrales, et al. [2]. Hatuwal and Thapa [3] stated that tests like imaging sets are used to carry out a diagnosis for lung cancer and its subtypes, it also states that diagnosing cancer types is time-consuming, with a potential of misdiagnosis, which can lead to incorrect treatment and thus endangering the patient's life. Therefore, a suitable Machine Learning model must be developed so that it can help in early detection of lung cancer [3]. Meanwhile, most authors in previous works used CT or histopathological images of lungs with machine learning techniques like Support Machine Vector, CNN, and Naive Bayes for lung cancer detection. This research paper has considered using non-image data like the patient's demographic and symptoms instead of using lung images with various machine learning algorithms used in previous works or rarely mentioned in other related works.

Because of its high mortality rate, lung cancer is known as one of the most terrifying types of cancer. According to this statistic, there are about 1.79 million deaths in the year 2020 caused by lung cancer [4]. Therefore, early lung cancer detection is important as it can increase patients' survival rates when patients receive early treatment. However, an article states that false negative cases, caused by radiological errors or MDT decisions, can lead to delayed diagnosis, causing the patient's survival rate to decrease and increasing the cost of the treatment for a late-stage tumor [5]. Therefore, it is imperative to create a machine learning model that can detect lung cancer early and decrease the rate of false negative cases occurring. What's more, Patra [6] study states that little research has been done on the prediction of the onset of lung cancer at an early stage by using classification algorithms to prevent a high mortality rate through effective control measures [6]. Furthermore, it is crucial for clinical cancer research to develop improved predictive models using multivariate data and advanced diagnostic technologies. A study using machine learning techniques and sizable data sets with previous patients diagnosed can train a prediction model which can predict lung cancer diagnosis [7]. Lastly, non-invasive imaging techniques like CT scans and PET scans are commonly used for early lung cancer detection. However, Alghamdi, et al. [8] states that these techniques have limitations in terms of accuracy and specificity[8]. Hence, there is a need to develop more advanced imaging techniques or combine existing techniques with artificial intelligence (AI) algorithms to improve the early detection and diagnosis of lung cancer. The objective of this article is to perform pre-processing on the dataset for testing and training the models, to build a statistical model that can be used in lung cancer detection by using Python and to examine the performance of classification algorithms for lung cancer detection based on accuracy.

2. Literature Review

Lung cancer is a type of cancer that starts in the lungs and the speed of spreading it to other parts of the body are depended on the type of lung cancer. Early detection of lung cancer is crucial for successful treatment, as the survival rates decrease significantly in the later stages of the disease. Significant advancements have been made in the creation of diagnostic techniques for the identification of lung cancer in recent years. The researchers said that one method for detecting lung cancer in CT images is based on rule-based classifiers. This method groups pixels that have similar results for rules into clusters and then extract features from each cluster using texture analysis. The features are then fed into a machine learning algorithm (such as SVM, KNN, or Naive Bayes) to classify the clusters as normal or abnormal. The method is evaluated on a dataset of 100 CT images with 50 normal and 50 abnormal cases. The results indicate the method achieves 96% accuracy, 98% for sensitivity, and 94% for specificity. This method contributes to the literature by proposing a novel rule-based clustering technique for lung cancer detection that reduces noise and enhances contrast in CT images. The method also compares different machine learning algorithms for classification and shows that SVM performs better than KNN and Naive Bayes [9].

Another method for detecting lung cancers on chest radiographs is based on deep learning models. This method consists of two stages: first, it segments the lung regions from the chest radiographs using U-Net, a convolutional neural network (CNN) architecture; second, it classifies the segmented lung regions as benign or malignant using ResNet-50, another CNN architecture. On a dataset, the model is trained and tested with 3,000 chest radiographs with 1,500 benign and 1,500 malignant cases. The results show that the model achieves 97% accuracy, 98% for sensitivity, and 96% for specificity. This method contributes to the literature by developing and validating a deep learning model that can detect lung cancers on chest radiographs with high accuracy and a low false positive rate. The model also demonstrates that segmentation can improve the performance of classification by reducing background noise and focusing on relevant regions [10]. Machine learning models for diagnosing lung cancer using historical data is also popular. This approach predicts a future lung cancer diagnosis using a variety of machine learning techniques, including ANN, Naive Bayes, Logistic Regression, and SVM based on demographics, smoking history, comorbidities, and blood test results. Using a dataset of 6,505 non-small cell lung cancer patients and 6,505 matched controls, the algorithm is trained and tested. According to the findings, the approach has a ROC and Area Under Curve of 0.91, 0.86 for sensitivity, and 0.83 for specificity [11].

Another method for detecting lung cancer using gene expression data is based on machine learning and bioinformatics techniques. This method uses a hybrid feature selection algorithm that combines filter and wrapper methods to select the most relevant genes for lung cancer diagnosis. The method then uses various machine learning algorithms (such as SVM, KNN, Random Forest, and Decision Tree) to classify the samples as normal or abnormal. The method is tested on a dataset of 546 samples with 203 normal and 343 abnormal cases. The results show that the method achieves 98.72% accuracy, 99.13% for sensitivity, and 97.54% for specificity. This method contributes to the literature by proposing a novel hybrid feature selection algorithm for lung cancer detection that can reduce the dimensionality and complexity of gene expression data [12]. Researchers from MIT and Massachusetts General Hospital have developed a new AI tool called Sybil that can

predict whether a patient will get lung cancer up to six years in advance. Sybil uses deep learning to analyse chest X-rays and other clinical data, such as smoking history and age, to assign a risk score for each patient. The tool was trained on more than 200,000 chest X-rays from over 100,000 patients who participated in the National Lung Screening Trial. The researchers found that Sybil outperformed existing lung cancer screening methods and could potentially increase the number of early detections and reduce the number of false positivesOuyang, et al. [13]. Anwar, et al. [14] suggested random forest algorithm is the best model for lung cancer detection using lung images. The authors use the Wisconsin dataset from the UCI repository as the input for training and testing the models to compare the random forest algorithm with other existing models like Decision Tree, SVM, and Logistic Regression to show that the random forest algorithm is the best. The accuracy report will be collected from all models and compared, it is found that the algorithm proposed by the authors outperforms the other three models, achieving an accuracy of 83% Anwar, et al. [14]. Singh and Gupta [15] tested the performance of various machine learning models for detecting lung cancer and classification. The machine learning models that are used for testing are decision tree, SVM, multinomial Naive Bayes, K-Nearest Neighbour, stochastic gradient descent, Random Forest, and MLP (Multi-Level Perceptron). Using a dataset containing 15750 images to train and test each model, it is shown that the MLP (Multi-Level Perceptron) model has the highest accuracy among other models, with an accuracy score of 88.55% [15]. Meanwhile, Nasser and Abu-Naser [16] created an ANN model to detect any presence of lung cancer using the patient's symptoms. The model is then trained and tested by the authors using a dataset obtained from a data world website that contains the symptoms as the input. The model produced an accuracy of 96.67% [16].

The authors have studied and analysed recent systems used for lung cancer detection to choose the best current system based on CT scan images of lungs proposing a new model by improving on the chosen best model. The proposed model has the pre-processed images segmented using watershed segmentation and SVM (Support Vector Machine) is used to perform an additional stage of classification of whether the tumour is malignant or not by using the extracted features of the cancer nodule as training data and used to test the proposed model. This resulted in 92% accuracy with the proposed model, which is higher than the current model with an accuracy of 86.6%. However, the authors state that the model cannot classify lung cancer into different stages and suggested implementing classification in different stages and employing proper pre-processing to help further improve the proposed model in the future [17].

Researchers from this article also developed a model using random forest classification to classify cancer from CT (Computer Tomography) images to detect cancer cells much more efficiently. A set of images from LIDC (Lung Images Database Consortium) will be used as the input to train and test the model, but before that, the images will be pre-processed, segmented, and have the features extracted. To compare the model's efficiency, other existing models will be trained, and tested, and the report compared with the random tree model. It is found that the developed model has the highest accuracy, 89.9%, followed closely by KNN, at an accuracy of 89% [18].

This article has applied several classification techniques, including Logistic Regression, SVM, Decision Tree, and Naïve Bayes to detect lung cancer by using a dataset of lung cancer patients and healthy individuals. The researchers examine and compare the classification models' performance. According to the results, the Support Vector Machine algorithm, using a dataset from Data World, achieved the highest accuracy of 99.2% compared to all other classification algorithms, and Logistic Regression, using a dataset from UCI, achieved the highest accuracy of 96.9% compared to Decision Tree, which is 85.71% [19]. Similarly, the authors of the study have used some machine learning models to forecast the likelihood of getting lung cancer by using non-imaging information such as demographic, lifestyle, and medical history information. For instance, the researchers suggest feeding the machine learning models with symptom, feature, and factor data collected from the Electronic Medical Records of patients, as input to train the machine learning models. Then, the researchers will test and compare the performance of several machine learning models, including Decision Tree, Naive Bayes, Logistic Regression, Artificial Neural Network, Random Forest, KNN, and so on. Rotation Forest (RotF) has achieved the highest accuracy, sensitivity, and specificity, which are all 97.1%, as well as AUC, which is 99.3% [20]. However, this article suggests that machine learning classifiers, particularly the RBF kernel based SVM algorithm, can be an effective and cost-efficient tool for screening individuals at high risk for developing lung cancer. This is because the highest accuracy of 81.25% was achieved by the RBF kernel based SVM algorithm compared to Naive Bayes, KNN, and J48 classifiers [6]. Furthermore, the authors of the article found that ensembles of classifiers outperform individual classifiers for predicting lung cancer. Therefore, the authors used ensemble approaches to combine the predictions of three individual classifiers, such as Multi-Layer Perceptron, Gradient Boosted Tree, and Support Vector Machine, to create a majority voting-based ensemble. Although the majority voting ensemble performed admirably with an accuracy of 88.57%, Gradient Boosted Tree obtained the highest accuracy of 90%. Researchers from this article predicted early-stage lung cancer using a dataset of historical records utilizing a variety of machine learning techniques, including SVM, KNN, Random Forest, Artificial Neural Network, and a hybrid model, Voting Classifier. The hybrid model outperformed all other models, achieving an accuracy of 99.5%[21]. The authors also mentioned that all the model's performance varies depending on the size of the data gathered and the noise [22].

In this article, the authors did research the prediction of chronic disease by using Logistic Regression (LR) and the authors also indicated that the logistic regression had produced a good performance as ML models on the prediction of risk of common chronic diseases. In addition, the researchers have implied that traditional regression models such as Logistic Regression models should keep playing a significant role in predicting disease risk. Instead of machine learning, deep learning is also a technique such as using deep reinforcement learning to detect lung cancer by using the dataset [23]. The authors explore that deep reinforcement learning has the potential to be used in the detection and treatment of lung cancer with machine learning. Deep reinforcement learning models such as deep Q-learning and deep Q-Network can be used to

detect diseases [24]. This article uses convolutional neural network models in deep learning to detect lung cancer. Based on the research paper, the training accuracy of this model has been able to achieve 96.11% and the validation accuracy has been able to achieve 97.20%. Therefore, the researchers claimed that the convolutional network model is suitable to be implemented for classifying lung cancer in the earlier stage [3]. The SegNet approach is one of the deep learning approaches which can be applied to detect lung cancer. Based on the research article, the authors attempted that early lung cancer's vulnerability to detection can be improved effectively by SegNet recognition technology. The sensitivity for the SegNet model is 98.33%.and an accuracy of 92.50% which can detect lung cancer accurately and effectively [25]. Instead of machine learning and deep learning, statistical analysis is also a technique to detect lung cancer. The authors of this article have used SAS to do statistical analysis with the sample selected to determine what are the categories of people who experienced a CT examination for lung cancer within the last year. The research has shown that the rate of lung cancer screening utilization is low [26].

However, there is still a lack of studies that combine non-imaging data like patients' demographic and medical history with machine learning to predict the presence of lung cancer in patients. This is because most studies have performed research on the performance of many machine learning models for lung cancer detection using only CT images of the patient's lungs. Although valuable, CT images are unable to paint a full picture of a patient's health status. Therefore, more research must be conducted by combining both non-image data and machine learning as it may be able to develop a more accurate and effective model.

3. Methodology

Kaggle, a subsidiary of Google, started in 2010 and is an online community of data scientists and machine learning engineers, where users can publish their datasets or models created for either educational or research purposes. This allows users to compare the models with others or have a discussion with others regarding the models or dataset. Kaggle also allows users to host a competition for users to help solve data science challenges. The dataset that is used to compare the chosen machine learning algorithms is taken from Kaggle. There are 16 variables in this dataset with 309 rows of data, the variables are related to the patient's demography and symptoms diagnosed. Table 1 shows the variables' description and what data types of the variables.

Table 1.
Description of variables.

Variable	Description
Gender	Whether the patient is male or female
Age	The age of the patient, ranges between 21-87
Smoking	Whether the patient smokes or not
Yellow fingers	Whether the patient has yellow fingers or not
Anxiety	Whether the patient has anxiety or not
Peer pressure	Whether the patient has peer pressure or not
Chronic disease	Whether the patient has any chronic disease or not
Fatigue	Whether the patient has fatigue or not
Allergy	Whether the patient has any allergies or not
Wheezing	Whether the patient has wheezing or not
Alcohol consumption	Whether the patient consumes any alcohol or not
Coughing	Whether the patient has coughing or not
Shortness of breath	Whether the patient suffers from shortness of breath or not
Swallowing difficulty	Whether the patient has difficulty swallowing or not
Chest pain	Whether the patient suffers from chest pain or not
Lung cancer	Determines whether the patient is diagnosed with lung cancer or not

Figure 1 shows the steps taken to preprocess the dataset. Although there are no missing values in the dataset, there are still duplicate values and outliers. Any duplicate values will be dropped to prevent overfitting the model and producing biased results due to overweighting. Next, outliers in age will be identified, as this variable is the only one that is continuous. Any outliers detected in this variable will be replaced with the median of the variable. Furthermore, the dataset will be encoded using LabelEncoder to convert the text data into numerical values. Lastly, the data will be normalized using MinMaxScaler so that all the encoded data in the dataset is within the same range of values, which in this case is between 0 and 1.

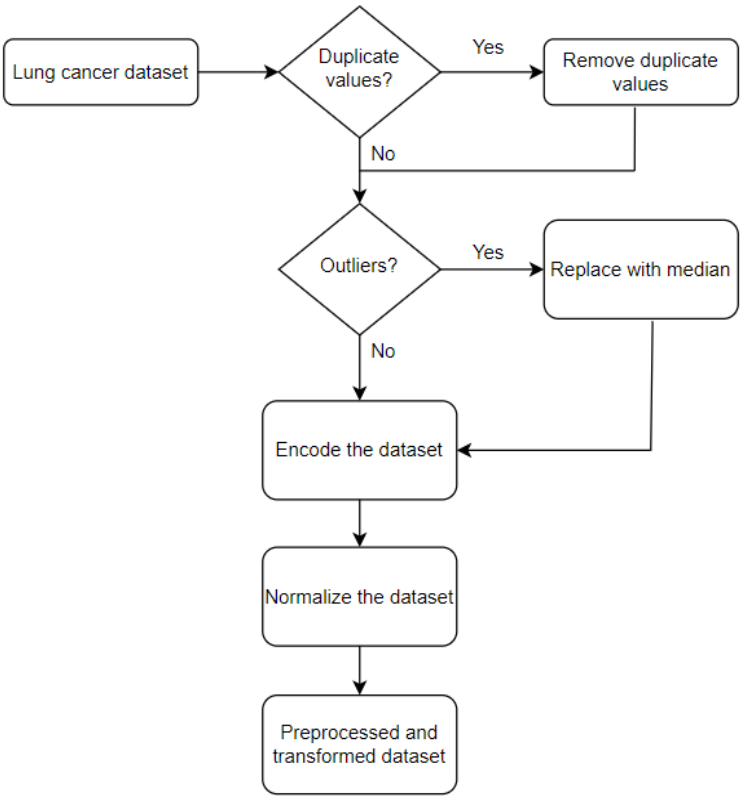

Figure 1.
Pre-processing Steps for Dataset.

After pre-processing and transforming the dataset, a few models will then be built so that each model's performance in predicting lung cancer will be tested as shown in Figure 2.

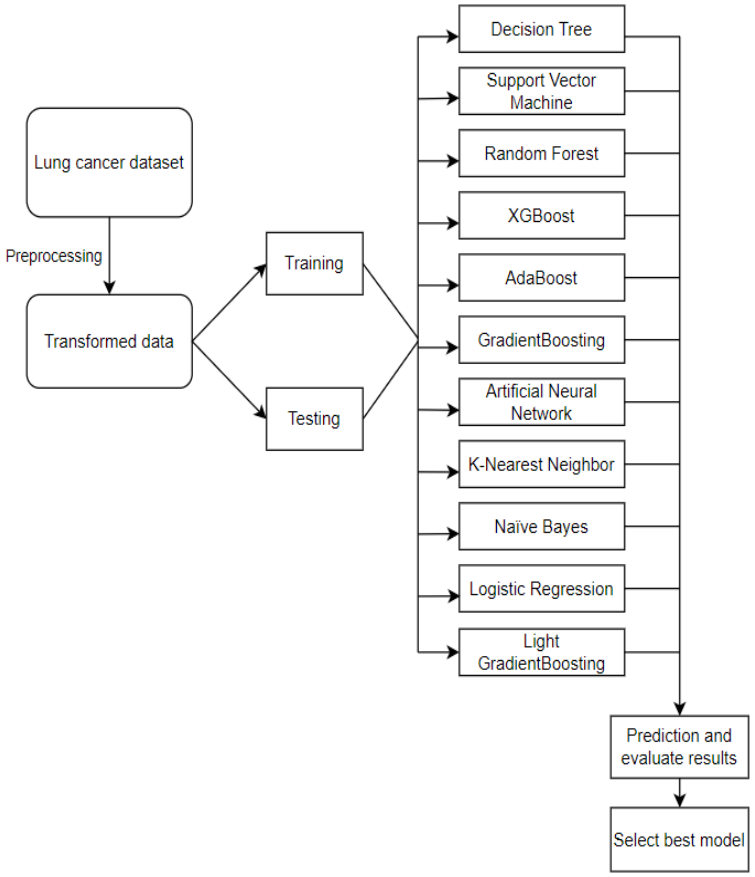

Figure 2.
Article workflow.

A few models that have been used by other researchers on this topic will also be selected and tested. The models chosen are Decision Tree, Artificial Neural Network, Logistic Regression, and Support Vector Machine, with the other models being Random Forest, XGBoost, AdaBoost, Gradient Boosting, Light Gradient Boosting, K-Nearest Neighbor, and Naive Bayes. Then, construct the models using Jupyter Notebook, import the pre-processed dataset, and split it into training and testing datasets. After that, train the model using the training dataset to achieve the desired accuracy, which is more than 80%. Finally, the fitted models will be used to predict lung cancer using the testing dataset and compare each model's performance using the model's accuracy, F1-score, precision, and recall score. A model with the best performance will be chosen.

Table 2 shows the distribution percentage of all the variables except age, this is because age is a continuous variable whilst all other variables are categorical. Based on Table 2 males are shown to be more likely to be diagnosed with lung cancer than females. Furthermore, the table shows that patients that have any kind of allergy, consumed alcohol, and have difficulty swallowing are also more likely to be diagnosed with lung cancer. Other symptoms like anxiety, yellow fingers, peer pressure, fatigue, coughing, chest pain, chronic disease, and wheezing also show that patients with these symptoms are also more likely to be diagnosed with lung cancer.

Table 2.
Feature distribution.

Feature	Lung cancer		Feature	Lung cancer	
	No	Yes		No	Yes
Gender			**Smoking**		
Male	11.97%	88.03%	No	15.80%	84.92%
Female	15.67%	84.33%	Yes	12.67%	87.33%
Yellow fingers			Anxiety		
No	21.37%	78.63%	No	18.71%	81.29%
Yes	8.18%	91.82%	Yes	8.76%	91.24%
Peer pressure			Chronic disease		
No	20.59%	79.41%	No	18.94%	81.06%
Yes	7.14%	92.86%	Yes	9.03%	90.97%
Fatigue			Allergy		
No	21.50%	78.49%	No	26.40%	73.60%
Yes	9.84%	90.16%	Yes	3.31%	96.69%
Wheezing			Alcohol consumption		
No	23.20%	76.80%	No	25.00%	75.00%
Yes	5.96%	94.04%	Yes	4.61%	95.39%
Coughing			Shortness of breath		
No	23.93%	76.07%	No	16.67%	83.33%
Yes	6.29%	93.71%	Yes	12.07%	87.93%
Swallowing difficulty			Chest pain		
No	22.45%	77.55%	No	21.31%	78.69%
Yes	3.88%	96.12%	Yes	7.79%	92.21%

4. Result

In Table 3 this research can compare the evaluation metrics of each model and select which model is the best performance model. The model with the highest accuracy is the KNN model which has 96.39%. However, this study has 7 models that have the highest recall which are Logistic Regression, SVM, Random Forest, AdaBoost, Gradient Boosting, Light Gradient Boosting, and XGBoost with a recall of 100%. Based on Table III, the KNN model has the highest precision and highest F1-score. The precision of KNN is 100% while the F1-score of KNN is 97.81%. Therefore, although the KNN model has the lowest recall, it can be said that the best model in this study will be the KNN model.

Table 3.
Model Comparison with evaluation metric.

Model	Accuracy (%)	Recall (%)	Precision (%)	F1-Score (%)
Decision tree	91.57	98.57	92.00	95.17
ANN	91.57	97.14	93.15	95.10
Logistic regression	89.16	100.00	88.61	93.96
SVM	95.18	100.00	94.59	97.22
KNN	96.39	95.71	100.00	97.81
Naïve bayes	91.57	97.14	93.15	95.10
Random forest	85.54	100.00	85.37	92.11
Ada boost	85.54	100.00	85.37	92.11
Gradient boosting	84.34	100.00	84.34	91.50
Light gradient boosting	90.36	100.00	89.74	94.59
XG boost	85.54	100.00	85.37	92.11

5. Discussion

The study aimed to create a statistical model for lung cancer detection and to evaluate the performance of various classification algorithms for this task. To accomplish this, the dataset will be pre-processed and tested using several classification algorithms, including Logistic Regression, ANN, Decision Tree, and SVM, as well as other models like KNN, Naïve Bayes, AdaBoost, Gradient Boosting, Light Gradient Boosting, XGBoost, and Random Forest.

Results shown indicated that the KNN model had the highest accuracy out of all the classification algorithms tested, with an accuracy score of 96.39%. This suggests that KNN may be a promising method for lung cancer detection, as it can accurately classify samples based on their features. It is significant that while KNN played out the best in the result, other models likewise showed great performance. SVM, Naive Bayes, Decision Tree, and ANN, for instance, achieved accuracy scores of 95.18%, 91.57%, 91.57%, and 91.57%, respectively. This suggests that there might be a variety of models for lung cancer detection that can work well, and that more research is needed to figure out which one is best. Table 4, 5, and 6 shows the comparison of the proposed model and existing studies' models for detecting lung cancer using machine learning. In terms of accuracy, the proposed model achieved an accuracy score of 91.57%, which is comparable to the Decision Tree model from the existing studies (93.7%). However, the ANN model from the existing studies achieved a higher accuracy score of 97.20% compared to the proposed model of 91.57%, nevertheless, the desirable score is achieved here.

Table 4.
Accuracy of existing and proposed model.

Models	Accuracy (%)							
	Proposed model	Dritsas and Trigka [20]	Patra [6]	Gould, et al. [11]	Makaju, et al. [17]	Munir, et al. [27]	Hatuwal and Thapa [3]	Nasser and Abu-Naser [16]
Decision tree	91.57	93.7	-	-	-	-	-	-
ANN	91.57	94.6	-	-		87.14	97.20	96.67
SVM	95.18	95.4	-	-	92	-	-	-
KNN	96.39	95.2	75	-	-	-	-	-
NB	91.57	95.0	78.13	-	-	-	-	-
Logistic regression	89.16	-	-	-	-	-	-	-
Random forest	85.54	-	-	-	-	-	-	-
Ada boost	85.54	-	-	-	-	-	-	-
Gradient boosting	84.34	-	-	-	-	-	-	-
Light gradient boosting	90.36	-	-	-	-	-	-	-
XG boost	85.54	-	-	81.00	-	-	-	-

Table 5.
Recall score for existing and proposed model.

Models	Recall (%)		
	Proposed model	Dritsas and Trigka [20]	Patra [6]
KNN	95.71	95.2	75
Decision tree	98.57	93.7	-
SVM	100.00	95.4	-
ANN	97.14	94.6	-
NB	97.14	95	78

Table 6.
Precision of existing and proposed model.

Models	Precision (%)		
	Proposed model	Dritsas and Trigka [20]	Patra [6]
KNN	100.00	95.2	73
Decision tree	92.00	93.7	-
SVM	94.59	95.4	-
ANN	93.15	94.6	-
NB	93.15	95	77.5

For the SVM model, the proposed model achieved a good accuracy score of 95.18% compared to the existing studies' models of 95.4% and 92%. In terms of KNN, the proposed model significantly outperformed the existing studies' models with an accuracy score of 96.39%, compared to the highest accuracy score achieved by the existing studies' model, which is 95.2%. The proposed model achieved an accuracy score of 91.57% in the NB model, which is comparable to one of the existing studies' models (95.0%), while the other model achieved a lower score of 78.13%. Finally, in the XGBoost model, the proposed model achieved a better accuracy score of 85.54% compared to the existing studies' model (81.00%).

Overall, the proposed model demonstrated competitive performance compared to the existing studies' models in terms of accuracy for most models, except for the ANN model, where the existing studies' model achieved a higher accuracy score. However, it is important to consider other factors such as interpretability, complexity, and scalability when selecting the appropriate algorithm for detecting lung cancer. Some of the models achieved 100% precision and 100% recall in predicting lung cancer, which is a highly unlikely result. This is because some of the features in the dataset that affect precision and recall are 100% according to the feature importance testing in the research. For example, in this research, the feature importance of the model was carried out and examined, revealing that age and alcohol consumption were the most important features for lung cancer detection using the KNN model.

6. Conclusion

This study has trained and built supervised learning models to detect patients who exhibit signs of lung cancer using non-imaging data such as various patients' historical features and symptoms. The machine learning models developed, such as ANN, Logistic Regression, Decision Tree, SVM, KNN, Naïve Bayes, Random Forest, AdaBoost, Gradient Boosting, Light Gradient Boosting, and XGBoost, were assessed for their performance in terms of accuracy, precision, recall, and F1-score. From the results, the use of classification models like K-Nearest Neighbors has proven effective in identifying lung cancer, as it displays superior performance compared to the other models, achieving the highest accuracy, precision, and F1-score.

In addition, it is essential to acknowledge this study's limitations. The dataset used in this research may not represent the diversity of patients seen in different clinical settings, as it contains small, limited samples of data. Another limitation is that the dataset is imbalanced, which can affect the performance of the models. One of the future works could involve constructing various models on larger and more diverse datasets to ensure generalizability. Additionally, deep learning models like CNN and long short-term memory (LSTM) may be utilized to detect lung cancer using both imaging and non-imaging data to improve the accuracy of lung cancer detection models in future research. Furthermore, the integration of SMOTE oversampling techniques could help address the issue of imbalanced data and enhance the performance of the models. Overall, this study presents a promising avenue for future research in the field of healthcare by utilizing machine learning to detect the early stages of lung cancer.

References

[1] C. Mattiuzzi and G. Lippi, "Current cancer epidemiology," *Journal of Epidemiology and Global Health*, vol. 9, no. 4, pp. 217-222, 2019. https://doi.org/10.2991/jegh.k.191008.001

[2] L. Corrales, R. Rosell, A. F. Cardona, C. Martin, Z. L. Zatarain-Barron, and O. Arrieta, "Lung cancer in never smokers: The role of different risk factors other than tobacco smoking," *Critical Reviews in Oncology/Hematology*, vol. 148, p. 102895, 2020. https://doi:10.1016/j.critrevonc.2020.102895.

[3] B. K. Hatuwal and H. C. Thapa, "Lung cancer detection using convolutional neural network on histopathological images," *International Journal of Computer Trends & Technology*, vol. 68, no. 10, pp. 21-24, 2020. https://doi.org/10.14445/22312803/ijctt-v68i10p104

[4] J. Elflien, "Number of cancer deaths worldwide in 2020, by major type of cancer. WHO; International agency for research on cancer," Retrieved: https://www.statista.com/statistics/288580/number-of-cancer-deaths-worldwide-by-type/, 2020.

[5] E. C. Bartlett, M. Silva, M. E. Callister, and A. Devaraj, "False-negative results in lung cancer screening—evidence and controversies," *Journal of Thoracic Oncology*, vol. 16, no. 6, pp. 912-921, 2021. https://doi.org/10.1016/j.jtho.2021.01.1607

[6] R. Patra, "Prediction of lung cancer using machine learning classifier," presented at the In Computing Science, Communication and Security: First International Conference, COMS2 2020, Gujarat, India, March 26–27, 2020, Revised Selected Papers 1 (pp. 132-142). Springer Singapore, 2020.

[7] S. Huang, J. Yang, S. Fong, and Q. Zhao, "Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges," *Cancer Letters*, vol. 471, pp. 61-71, 2020. https://doi.org/10.1016/j.canlet.2019.12.007

[8] M. Alghamdi *et al.*, "Clinical and genetic characterization of craniosynostosis in Saudi Arabia," *Frontiers in Pediatrics*, vol. 9, p. 582816, 2021. https://doi.org/10.3389/fped.2021.582816

[9] S. S. Raoof, M. A. Jabbar, and S. A. Fathima, "Lung cancer prediction using machine learning: A comprehensive approach," presented at the In 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA) (pp. 108-115). IEEE, 2020.

[10] A. Shimazaki *et al.*, "Deep learning-based algorithm for lung cancer detection on chest radiographs using the segmentation method," *Scientific Reports*, vol. 12, no. 1, p. 727, 2022. https://doi.org/10.1038/s41598-021-04667-w

[11] M. K. Gould, B. Z. Huang, M. C. Tammemagi, Y. Kinar, and R. Shiff, "Machine learning for early lung cancer identification using routine clinical and laboratory data," *American Journal of Respiratory and Critical Care Medicine*, vol. 204, no. 4, pp. 445-453, 2021. https://doi.org/10.1164/rccm.202007-2791OC

[12] E. Svoboda, "Artificial intelligence is improving the detection of lung cancer," *Nature*, vol. 587, no. 7834, pp. S20-S20, 2020. https://doi.org/10.1038/d41586-020-03157-9

[13] A. Ouyang, A. Latif Jameel, R. Barzilay, L. Sequist, F. Fintelmann, and I. Fuentes, "MIT researchers develop an AI model that can detect future lung cancer risk Deep-learning model takes a personalized approach to assessing each patient's risk of lung cancer based on CT scans," Retrieved: https://news.mit.edu/2023/ai-model-can-detect-future-lung-cancer-0120, 2023.

[14] M. Anwar, M. Bakar, H. Awais, M. Din, M. Mohsin, and M. Nazir, "Early detection of lungs cancer using machine learning algorithms," *Biological and Clinical Sciences Research Journal*, vol. 2023, no. 1, p. 187, 2023. https://doi.org/10.54112/bcsrj.v2023i1.187

[15] G. A. P. Singh and P. Gupta, "Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans," *Neural Computing and Applications*, vol. 31, no. 10, pp. 6863-6877, 2019. https://doi.org/10.1007/s00521-018-3518-x

[16] I. M. Nasser and S. S. Abu-Naser, "Lung cancer detection using artificial neural network," *International Journal of Engineering and Information Systems*, vol. 3, no. 3, pp. 17-23, 2019.
[17] S. Makaju, P. Prasad, A. Alsadoon, A. Singh, and A. Elchouemi, "Lung cancer detection using CT scan images," *Procedia Computer Science*, vol. 125, pp. 107-114, 2018. https://doi.org/10.1016/j.procs.2017.12.016
[18] D. Jayaraj and S. Sathiamoorthy, "Random forest based classification model for lung cancer prediction on computer tomography images," presented at the In 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 100-104). IEEE, 2019.
[19] P. R. Radhika, R. A. Nair, and G. Veena, "A comparative study of lung cancer detection using machine learning algorithms," presented at the In 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT) (pp. 1-4). IEEE, 2019.
[20] E. Dritsas and M. Trigka, "Lung cancer risk prediction with machine learning models," *Big Data and Cognitive Computing*, vol. 6, no. 4, 2022. https://doi.org/10.3390/bdcc6040139
[21] C. Thallam, A. Peruboyina, S. S. T. Raju, and N. Sampath, "Early stage lung cancer prediction using various machine learning techniques," in *In Proceedings of the 4th International Conference on Electronics, Communication and Aerospace Technology, ICECA 2020; Institute of Electrical and Electronics Engineers Inc., 2020; pp 1285–1292. https://doi.org/10.1109/ICECA49313.2020.9297576*, 2020.
[22] M. I. Faisal, S. Bashir, Z. S. Khan, and F. H. Khan, "An evaluation of machine learning classifiers and ensembles for early stage prediction of lung cancer," presented at the In 2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST) (pp. 1-4). IEEE, 2018.
[23] S. Nusinovici *et al.*, "Logistic regression was as good as machine learning for predicting major chronic diseases," *Journal of Clinical Epidemiology*, vol. 122, pp. 56-69, 2020. https://doi.org/10.1016/j.jclinepi.2020.03.002
[24] Z. Liu, C. Yao, H. Yu, and T. Wu, "Deep reinforcement learning with its application for lung cancer detection in medical Internet of Things," *Future Generation Computer Systems*, vol. 97, pp. 1-9, 2019. https://doi.org/10.1016/j.future.2019.02.068
[25] X. Chen, Q. Duan, R. Wu, and Z. Yang, "Segmentation of lung computed tomography images based on SegNet in the diagnosis of lung cancer," *Journal of Radiation Research and Applied Sciences*, vol. 14, no. 1, pp. 396-403, 2021. https://doi.org/10.1080/16878507.2021.1981753
[26] W. E. Zahnd and J. M. Eberth, "Lung cancer screening utilization: A behavioral risk factor surveillance system analysis," *American Journal of Preventive Medicine*, vol. 57, no. 2, pp. 250-255, 2019. https://doi.org/10.1016/j.amepre.2019.03.015
[27] K. Munir, H. Elahi, A. Ayub, F. Frezza, and A. Rizzi, "Cancer diagnosis using deep learning: A bibliographic review," *Cancers*, vol. 11, no. 9, p. 1235, 2019. https://doi.org/10.3390/cancers11091235