Journal of Current Science and Technology, September-December 2024 Vol. 14 No. 3, Article 55
Copyright ©2018-2024, Rangsit University
ISSN 2630-0656 (Online)

Cite this article: Pechprasarn, S., Suechoey, N., Pholtrakoolwong, N., Tanedvorapinyo, P., & Toboonliang, Y. (2024). Optimizing lung cancer diagnosis with machine learning and feature selection methods. *Journal of Current Science and Technology*, *14*(3), Article 55. https://doi.org/10.59796/jcst.V14N3.2024.55

Journal of Current Science and Technology

Journal homepage: https://jcst.rsu.ac.th

Optimizing Lung Cancer Diagnosis with Machine Learning and Feature Selection Methods

Suejit Pechprasarn[1,*], Nichapha Suechoey[2], Nutchareeya Pholtrakoolwong[2], Pattaraporn Tanedvorapinyo[2], and Yanisa Toboonliang[2]

[1]College of Biomedical Engineering, Rangsit University, Pathum Thani 12000, Thailand
[2]Satriwithaya School, Wat Bowon Niwet, Phra Nakhon, Bangkok 10200, Thailand
The last four authors have contributed equally.

[*]Corresponding author; E-mail: suejit.p@rsu.ac.th

Received 12 March 2024; Revised 9 April 2024; Accepted 10 April 2024
Published online 1 September 2024

Abstract

Lung cancer is a prevalent disease, with nearly 238,000 new cases diagnosed in 2023. This study utilizes clinical predictors from a Kaggle dataset containing 309 observations across 15 variables to aid in lung cancer diagnosis. The variables include swallowing difficulty, peer pressure, gender, allergy, yellow fingers, anxiety, wheezing, alcohol consumption, chronic disease, chest pain, coughing, fatigue, smoking, age, and shortness of breath. The research aims to develop and compare various supervised machine learning models for classifying and predicting lung cancer, while also identifying key clinical tests and parameters using unsupervised statistical models. The dataset was divided into training and test sets, balanced, and preprocessed for unbiased training. Feature selection and machine learning models were applied to identify crucial predictors. The study explored tree models, logistic regression, Naïve Bayes, support vector machine (SVM), ensemble, neural network, and kernel models. Among these, the linear SVM achieved the highest accuracy of 93.75% with 5-fold cross-validation. However, it showed overfitting, with a lower test accuracy of 82.55%. The Gaussian Naïve Bayes model emerged as the optimal choice, providing consistent performance between validation and test cases. It achieved the highest cross-validation classification accuracy of 82.81% using only 9 variables: swallowing difficulty, peer pressure, gender, allergy, yellow fingers, anxiety, wheezing, alcohol consumption, and chronic disease. This model allows for effective training with fewer predictors without compromising classification performance.

Keywords**:** *lung cancer; machine learning; smart diagnosis tool; artificial intelligence; feature selection; classification*

1. Introduction

Lung cancer is currently one of the leading causes of death in humans. Over the past years, lung cancer has been diagnosed globally at an increasing rate through time (Thandra et al., 2021; Li et al., 2023). Scientists believe that smoking is responsible for nearly 80% of lung cancer causes; however, non-smoking individuals are also likely to develop lung cancer due to other external predictors, such as being a second-hand smoker or fine dust in the atmosphere (Walser et al., 2008).

As technology has drastically developed throughout the years, experts can now classify lung cancer into 2 brief classifications: small-cell and non-small-cell lung cancer (Sankar et al., 2023; Ketkomol et al., 2024). Those in which are non-small cell types are far more common than small cell types and can be divided into many subtypes, whether it is squamous cell carcinoma, adenocarcinoma, or large cell carcinoma (Lareau et al., 2021). Once diagnosed with lung cancer, physicians run additional tests to determine its spread and stage in the body (Sherry, 2022).

The symptoms that an individual with lung cancer can have include many things, for example, coughing, weight loss, chest pain, and many more (Ruano-Raviña et al., 2020). When a physician evaluates a patient, they will inquire about those symptoms to determine if the patient requires further cancer treatment. Most patients with lung cancer present with digital clubbing, hemoptysis, coughing, weight loss, loss of appetite, chest pain, and fatigue (Kim et al., 2022).

Physicians often spend much time investigating possible symptoms that can indicate cancer (Vidaver et al., 2016); hence, it will be beneficial to rule out some insignificant symptoms to help clinical experts reduce diagnosis time.

A lung cancer diagnosis is usually carried out using imaging techniques such as CT scans, MRI scans, and PET scans (Bukhari et al., 2017). A physician will use the images from these tools to determine whether there is a possibility of cancer. The current diagnosis of lung cancer involves the use of spiral CT technology to assist with chest radiography, as well as different types of imaging complemented with pathological assessment of biopsies (Cegla et al., 2023). However, the most common diagnostic tool is white light bronchoscopy (WLB) (Nooreldeen, & Bach, 2021). Bronchoscopy in the lung is used for direct visualization of pathological changes in the trachea and bronchi. It has been modified as a flexible fiberoptic bronchoscopy that can visualize the tracheobronchial tree to the level of segmental and subsegmental divisions for diagnosing lung cancer (Arroliga, & Matthay, 1993).

In the past several decades of technological development, artificial intelligence (AI) has been utilized and implemented in many fields (Cao, 2017). There are several types of artificial intelligence these days, such as machine learning, expert systems, neural networks, and pattern recognition (Pechprasarn et al, 2023a). Machine learning has grown and developed rapidly in recent years and gained worldwide attention, and it usually provides systems or programs the ability to learn from experiences or data without being explicitly programmed (Sarker, 2021).

Support vector machines (SVM) have the advantage of handling structured data, data with multiple dimensions, and complex functions, which is suitable for determining kernel functions (Sowmya et al., 2021). In the 21st century, human life has been dramatically involved with AI, and it has also extended to the medical field. Numerous AI models have been published and proven to play an essential role in clinical decision-making as they can help clinicians predict treatment response, side effects, and prognosis (Chiu et al., 2022). Integrating AI into clinical workflows is promising and generally has satisfying results.

Artificial intelligence (AI) improves lung cancer diagnosis and treatment by classifying lung cancer types with high accuracy using techniques like deep learning and enhancing machine learning models for better diagnostic performance (Singh, & Gupta, 2019). With personalized treatment plans and precise prognostic predictions from AI, physicians can hope for improved patient outcomes (Pereira et al., 2020). As research in this field advances, collaboration between AI and medical expertise holds promise for a brighter future in the fight against lung cancer (Lakshmanaprabu et al., 2019).

Machine learning (ML) has accelerated several research fields related to the medical field, and the techniques are continuously updated. The ML model is adapted to learn from previous diagnoses and medical records to produce reliable results when new datasets are inserted into the model repeatedly (Rana, & Bhushan, 2023). Up until now, AI detection methods have also shown great potential in lung cancer patient care (Wang et al., 2019), as lung cancer is the best field for AI applications due to its heterogeneity (Chiu et al., 2022). One example of machine learning in lung cancer diagnosis includes screening standard criteria and laboratory results of patients to predict if they have developed lung cancer (Gould et al., 2021).

Furthermore, AI has also been used in the radiomics field, turning datasets into numbers that can be analyzed by AI, which is particularly helpful in studying lung cancer. After the analysis, the data will be able to diagnose, predict the progression of cancer, and monitor the treatment process as a helping tool for clinical physicians (Tunali et al., 2021).

The purpose of this study is to use machine learning models from the available dataset to identify crucial predictors of lung cancer causes. This developed model was aimed to scope down the predictors and reduce the time for physicians to evaluate and diagnose patients suspected to have lung cancer with only at-sight symptoms. Furthermore, this model should help individuals to observe their primary symptoms. We performed several training sessions to develop ML models, whether it was to use the feature selection method along with the available machine learning model in MATLAB2022b to identify crucial predictors of lung cancer.

2. Objectives

1) To identify critical symptoms indicating lung cancer and reduce diagnosis time for clinical technicians, the identified symptoms will be tested and validated by training and testing the model using only the identified symptoms and show that the trained model can provide accurate lung cancer diagnosis and prediction.

2) To use machine learning models to accurately predict lung cancer from the collected dataset. Different machine learning models will be trained and tested to compare their performance with results in the literature.

3) To apply feature selection methods and machine learning models to train a less complex lung cancer prediction model with similar performance.

3. Materials and methods

This section demonstrates the methodology, data source, data curation, classification training, and feature selection process. The process of this study is shown in Figure 1, provided below. Here, we propose a systematic approach to preprocess the dataset, preparing for unbiased training and compare different types of supervised machine learning models using performance metrics, including accuracy, precision, recall, specificity, F1-score, and area under the curve of receiver operating characteristic (ROC) curve with 5-fold cross-validation. After identifying an appropriate model, unsupervised feature selection methods (Karabulut et al., 2012) based on statistical analysis: ANOVA, Kruskal-Wallis, Chi^2, and MRMR are utilized to identify statistically significant parameters, and these clinical features are later verified by training the model with fewer predictors to show that the trained model is capable of predicting the lung cancer without compromising the classification performance.

3.1 Dataset source and details

The lung cancer dataset we used was collected from the Kaggle database website (Sasaki, 2020) (accessed February 27). The dataset contains data from 309 observances, including yes or no labels with details as described in Table 1 below.

3.2 Data Curation

The 309 observational data available were classified into 2 types: negative individuals and individuals with lung cancer. There are 39 cases for negative individual cases and 270 cases for individuals with lung cancer.

3.3 Datasets for training, validation, and test datasets

We reduced the dataset to a training dataset with a total of 64 rows, 32 for ordinary individuals and 32 for individuals with lung cancer, ensuring the training set was balanced and unbiased. The remaining original data were then separated into an unseen testing dataset, consisting of 7 rows of non-cancer individual data and 238 rows of lung cancer data, totaling 245 cases.

3.4 Machine Learning Training, Validation, Testing, and Performance Metrics

We used the Classification Learner application available in MATLAB 2022b to train the machine learning models listed in Table 2. The models were trained to compare their performance on the separated dataset (Pechprasarn et al., 2023b).

Classification performances were evaluated using a 5-fold cross-validation computing accuracy, precision, recall or sensitivity, specificity, and F1-score, using the equations 1 to 5 as expressed below.

$$Accuracy = \frac{T_p + Tn}{T_p + Tn + Fp + Fn} \quad (1)$$

$$Precision = \frac{T_p}{T_p + Fp} \quad (2)$$

$$Recall = sensitivity = \frac{T_p}{T_p + Fn} \quad (3)$$

$$F_{1\text{-}score} = 2\frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

$$Specificity = \frac{T_n}{T_n + F_p} \quad (5)$$

where T_p represents true positive cases
T_n represents true negative cases,
F_p represents false positive cases
F_n represents false negative cases.

The classification performance was then evaluated using the unseen test dataset to determine if the models were efficient and accurately predict other datasets, i.e., to test whether they could generalize to unseen data.

.

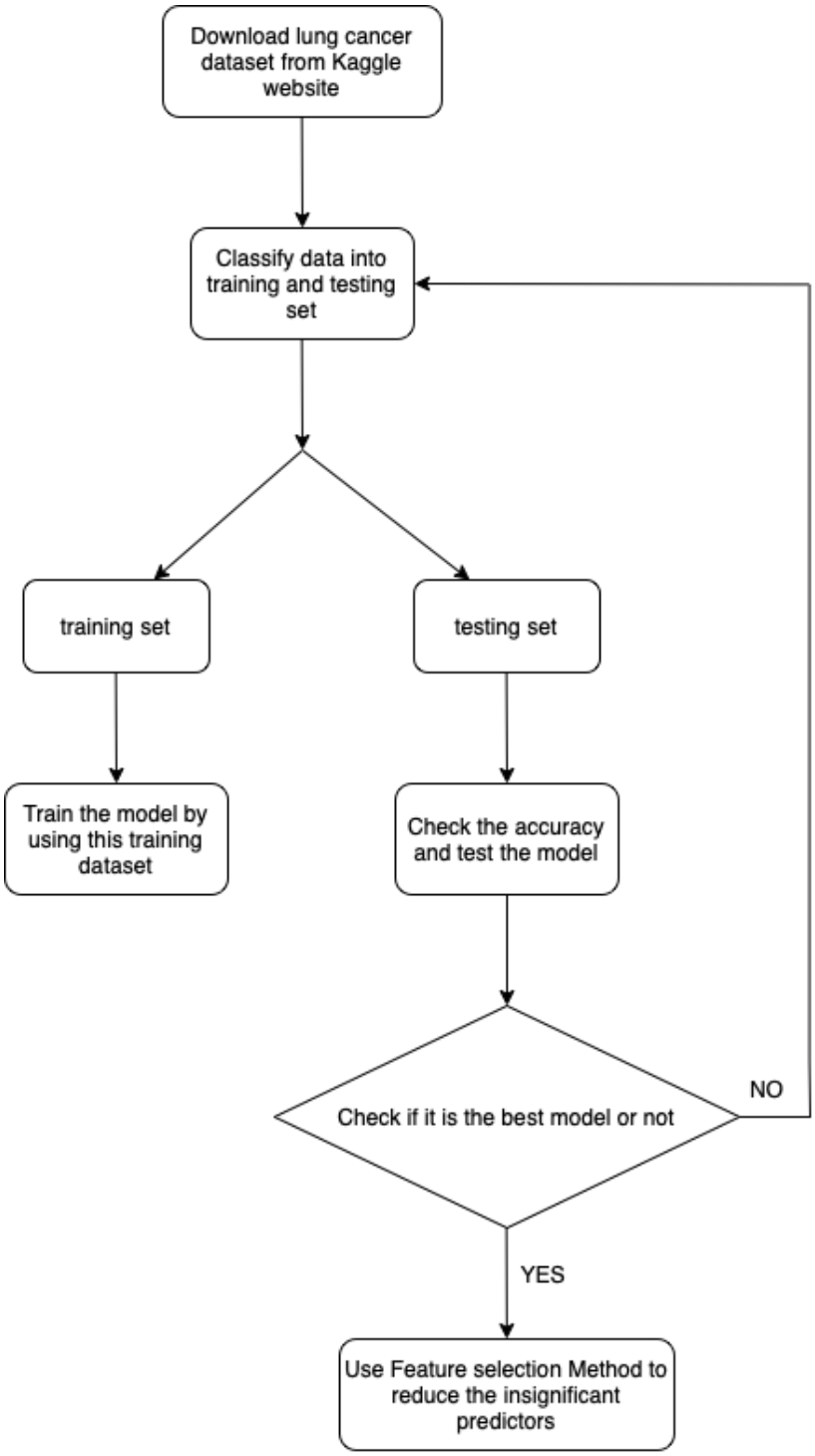

Figure 1 Process flow in this study includes data curation, data separation, training and testing, and feature selection. Adapted from "Identification of Important Factors in the Diagnosis of Breast Cancer Cells Using Machine Learning Models and Principal Component Analysis", published in *J. Curr. Sci. Technol, 13*(3), p. 645

Table 1 Predictors and labels' details from the downloaded dataset

Variable	Variable Detail	Values	Type
Swallowing Difficulty	Experience of pain while swallowing food, drink, or any substances	Yes: 2 No: 1	Predictor
Peer Pressure	Experience of being influenced in an activity or behavior by surrounding people	Yes: 2 No: 1	Predictor
Gender	Male or Female sexuality	Yes: 2 No: 1	Predictor
Allergy	Symptoms of being hyperimmune to a specific thing(s)	Yes: 2 No: 1	Predictor
Yellow Fingers	Experience of having yellowish fingers	Yes:2 No: 1	Predictor
Anxiety	Experience of being excessively worried or nervous	Yes: 2 No: 1	Predictor
Wheezing	Experience of having a high-pitched sound from lungs while breathing	Yes: 2 No: 1	Predictor
Alcohol Consuming	Personal drinking habits, including frequency and amount of beverage consumed	Yes: 2 No: 1	Predictor
Chest Pain	Experience having pain or discomfort around the chest	Yes: 2 No: 1	Predictor
Coughing	Experience coughing excessively	Yes: 2 No: 1	Predictor
Fatigue	Experience of being exhausted	Yes: 2 No: 1	Predictor
Smoking	Personal smoking habit, whether one has been smoking or not	Yes: 2 No: 1	Predictor
Age	The length of time one has been used on Earth ever since their labor	Yes: 2 No: 1	Predictor
Shortness of Breath	Experience of being unable to breathe regularly or feeling suffocated	Yes: 2 No: 1	Predictor
Lung Cancer Class	Yes: Positive lung cancer case No: Negative lung cancer case	Yes/No	Label

Table 2 The machine learning models utilized in the study

Model	Detail	Model	Detail
Tree	Fine tree	K-Nearest neighbor	Fine KNN
	Medium tree		Medium KNN
	Coarse tree		Coarse KNN
Logistic regression	Logistic Regression		Cubic KNN
Naïve Bayes	Gaussian Naïve Bayes		Weighted KNN
	Kernel Naïve Bayes	Ensemble	Boosted trees
SVM	Linear SVM		Subspace Discriminant
	Quadratic SVM		Subspace KNN
	Cubic SVM		RUS Boosted trees
	Fine Gaussian SVM	Neural network	Narrow Neural network
	Medium Gaussian SVM		medium Neural network
	Coarse Gaussian SVM		Wide Neural network
	Logistic regression Kernel		Tri-layered Neural network

3.5 Feature Selection

Feature selection models, including ANOVA, Kruskal Wallis, Chi^2, and MRMR, were applied to identify crucial predictors contributing to model classification accuracy using the Feature Selection analysis tool available on MATLAB 2022b. After identifying the essential predictors, the model was trained to show that they still performed well, similar to the models trained using all 15 variables. The results section will show that ML models prepared using fewer predictors can provide comparable performance to the models trained with all available predictors.

4. Results

4.1 Machine Learning Classification accuracy based on 15 parameters

Table 3 shows the comparison of the confusion matrix, accuracy, specificity, precision, recall, and F1-score computed from the trained models using 5-fold cross-validation and the training dataset of all the models available in MATLAB 2022b. As can be seen, the best performance model is the Linear SVM followed by Quadratic SVM, as their accuracy, specificity, precision, recall, and F1-score were exceptionally outstanding.

Table 4 shows the comparison of confusion matrix rate, accuracy, specificity, precision, recall, and F1-score computed from the trained models using 5-fold cross-validation and the test dataset of all the models in MATLAB 2022b as percentage. Since the number of data entries in each class was imbalanced, for a fair comparison, the confusion matrices were visualized as rate or normalized by the number of rows in each class.

Table 3 Confusion matrix, accuracy, specificity, precision, recall, and F1-score computed from the trained models using 5-fold cross-validation and the training dataset

Model	Detail	T_p	T_n	F_p	F_n	Accuracy	Specificity	Precision	Recall	F_1-score	AUC
Tree	Fine Tree	23	30	2	9	82.81%	93.75%	92.00%	71.88%	80.7%	84.18%
	Medium Tree	23	30	2	9	82.81%	93.75%	92.00%	71.88%	80.7%	84.18%
	Coarse Tree	23	31	1	9	84.38%	96.88%	95.83%	71.88%	82.14%	84.04%
Logistic Regression	Logistic Regression	26	27	5	6	82.81%	84.38%	83.88%	81.25%	82.54%	83.3%
Naïve Bayes	Gaussian Naïve Bayes	29	27	5	3	87.50%	84.38%	85.29%	90.63%	87.88%	91.02%
	Kernel Naïve Bayes	27	30	2	5	89.06%	93.75%	93.1%	84.38%	88.53%	91.31%
SVM	Linear SVM	29	31	1	3	93.75%	96.88%	96.67%	90.63%	93.55%	97.17%
	Quadratic SVM	29	30	2	3	92.19%	93.75%	93.55%	90.63%	92.07%	96.68%
	Cubic SVM	26	30	2	6	87.5%	93.75%	92.86%	81.25%	86.67%	95.02%
	Fine Gaussian SVM	32	12	20	0	68.75%	37.5%	61.54%	100%	76.19%	77.44%
	Medium Gaussian SVM	28	29	3	4	89.06%	90.63%	90.32%	87.5%	88.89%	96.0%
	Coarse Gaussian SVM	25	32	0	7	89.06%	100%	100%	78.13%	87.72%	96.19%
Ensemble	Boosted Trees	12	18	14	20	46.88%	56.25%	46.15%	37.5%	41.38%	0%
	Bagged Trees	29	27	5	3	87.5%	84.38%	85.29%	90.63%	87.88%	90.53%
	RUS Boosted Trees	12	18	14	20	46.88%	56.25%	46.15%	37.5%	41.38%	0%
Neural Network (NN)	Narrow NN	27	27	5	5	84.38%	84.38%	84.38%	84.38%	84.38%	89.26%
	Medium NN	27	26	6	5	82.81%	81.25%	81.82%	84.38%	83.08%	90.14%
	Wilde NN	26	28	4	6	84.38%	87.5%	86.67%	81.25%	83.87%	86.28%
	Bilayered NN	26	27	5	6	82.81%	84.38%	83.87%	81.25%	82.54%	83.84%
	Trilayered NN	27	29	3	5	87.5%	90.63%	90.00%	84.38%	87.1%	86.67%
Kernel	SVM Kernel	26	20	12	6	71.88%	62.5%	68.42%	81.25%	74.29%	72.56%
	Logistic regression Kernel	26	21	11	6	73.43%	65.63%	70.27%	81.25%	75.36%	78.91%

Table 4 Confusion matrix rate, accuracy, specificity, precision, recall, and F1-score computed from the trained models using 5-fold cross-validation and the test dataset

Model	Detail	T_p	T_n	F_p	F_n	Accuracy	Specificity	Precision	Recall	F_1-score	AUC
Tree	Fine Tree	61.80%	71.40%	28.60%	38.20%	66.60%	71.40%	68.36%	61.80%	64.92%	69.06%
	Medium Tree	61.80%	71.40%	28.60%	38.20%	66.60%	71.40%	68.36%	61.80%	64.92%	69.06%
	Coarse Tree	57.10%	71.40%	28.60%	42.90%	64.25%	71.40%	66.63%	57.10%	61.50%	60.29%
Logistic Regression	Logistic Regression	50.80%	100.00%	0.00%	49.20%	75.40%	100.00%	100.00%	50.80%	67.37%	78.36%
Naïve Bayes	Gaussian Naïve Bayes	80.70%	85.70%	14.30%	19.30%	83.20%	85.70%	84.95%	80.70%	82.77%	94.21%
	Kernel Naïve Bayes	53.80%	100.00%	0.00%	46.20%	76.90%	100.00%	100.00%	53.80%	69.96%	94.45%
SVM	Linear SVM	65.10%	100.00%	0.00%	34.90%	82.55%	100.00%	100.00%	65.10%	78.86%	95.89%
	Quadratic SVM	63.90%	100.00%	0.00%	36.10%	81.95%	100.00%	100.00%	63.90%	77.97%	95.35%
	Cubic SVM	63.00%	100.00%	0.00%	37.00%	81.50%	100.00%	100.00%	63.00%	77.30%	95.95%
	Fine Gaussian SVM	94.10%	57.10%	42.90%	5.90%	75.60%	57.10%	68.69%	94.10%	79.41%	89.17%
	Medium Gaussian SVM	68.10%	100.00%	0.00%	31.90%	84.05%	100.00%	100.00%	68.10%	81.02%	95.59%
	Coarse Gaussian SVM	64.30%	100.00%	0.00%	35.70%	82.15%	100.00%	100.00%	64.30%	78.27%	95.83%
Ensemble	Boosted Trees	0.00%	100.00%	0.00%	100.00%	50.00%	100.00%	N/A	0.00%	N/A	N/A
	Bagged Trees	35.30%	100.00%	0.00%	64.70%	67.65%	100.00%	100.00%	35.30%	52.18%	79.89%
	RUS Boosted Trees	0.00%	100.00%	0.00%	100.00%	50.00%	100.00%	N/A	0.00%	N/A	N/A
Neural Network (NN)	Narrow NN	63.00%	100.00%	0.00%	37.00%	81.50%	100.00%	100.00%	63.00%	77.30%	85.41%
	Medium NN	79.00%	100.00%	0.00%	21.00%	89.50%	100.00%	100.00%	79.00%	88.27%	95.17%
	Wilde NN	56.70%	100.00%	0.00%	43.30%	78.35%	100.00%	100.00%	56.70%	72.37%	89.50%
	Bilayered NN	59.20%	100.00%	0.00%	40.80%	79.60%	100.00%	100.00%	59.20%	74.37%	85.53%
	Trilayered NN	62.60%	71.40%	28.60%	37.40%	67.00%	71.40%	68.64%	62.60%	65.48%	72.30%
Kernel	SVM Kernel	50.00%	100.00%	0.00%	50.00%	75.00%	100.00%	100.00%	50.00%	66.67%	86.28%
	Logistic regression Kernel	64.30%	71.40%	28.60%	35.70%	67.85%	71.40%	69.21%	64.30%	66.67%	81.06%

The best performing machine learning model was the linear SVM model with an accuracy of 93.75%, precision of 96.67%, recall of 90.63%, specificity of 96.88%, and F_1 score of 93.55%, as shown in Table 3 in comparison to the other models. The confusion matrix of the validation case using a 5-fold cross-validation approach and their corresponding ROC for this model are shown in Figure 2a and Figure 2b below. The area under the ROC curves (AUC) was 0.9717 for the two classes. However, the accuracy percentage of this Linear SVM model was not yet satisfied when tested, as shown in Table 4, using the separated test dataset, as established in Figure 2c and Figure 2d. The performance metrics of the tested Linear SVM indicated high overfitting, with discrepancies of over 10%. Table 5 highlights the differences in classification performance metrics between the validation and test performance, where positive and negative values indicate overfitting and underfitting performance, respectively.

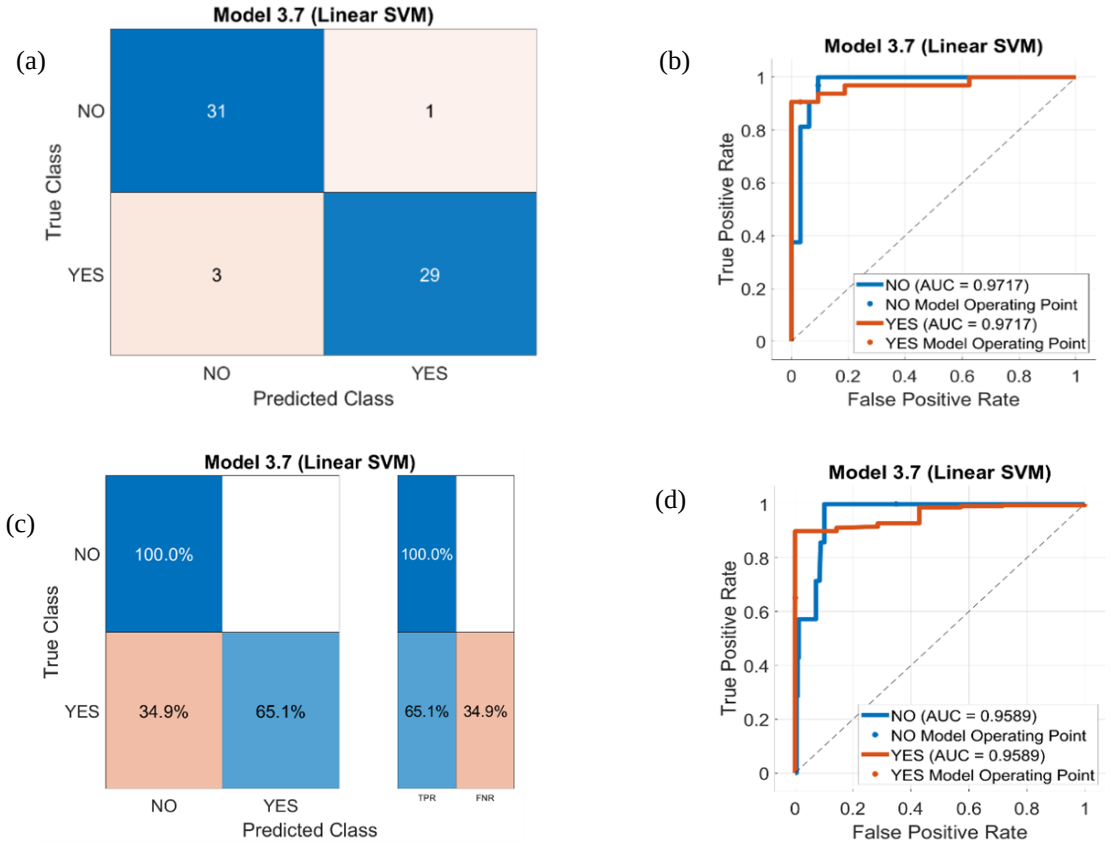

Figure 2 (a) Confusion matrices of the trained Linear SVM validation using the training dataset, (b) ROC curve plots of the trained Linear SVM validation using the training dataset, (c) Confusion matrices in percentage and True-False Rate Graph of the trained Linear SVM when tested using the test dataset, and (d) ROC curve plots of the trained Linear SVM using the test dataset.

Table 5 shows that the only model that did not have performance differences between the validation and test was Gaussian Naïve Bayes. The Gaussian Naïve Bayes model can predict with an accuracy of 87.50%, a precision of 84.38%, a recall of 85.29%, an F_1 score of 90.63%, and a specificity of 87.88% for the validation. For the test case, the model can provide similar performance with an accuracy of 83.20%, precision of 84.95%, recall of 80.70%, F_1 score of 82.77%, and specificity of 85.70%, respectively. The confusion matrices of the trained Gaussian Naïve Bayes for the validation and test cases are shown in Figures 3a and 3c. The trained model has an overfitting rate of 4.134% on average. Figures 3b and 3d show the ROC curves with their AUC values; the discrepancies were well within 2% between the validation and the test cases. The test performance of our trained model was similar to the performance reported in the literature; for example, Liu et al., (2020) employed logistic regression and achieved an accuracy of 86.2%. Pacurari et al., (2023), using an artificial neural network (ANN) and support vector machine (SVM), achieved accuracy from 77.8% to 100%. Teramoto et al., (2017) can classify approximately 71% of lung cancer-suspected images correctly by utilizing a deep convolutional neural network (DCNN). As you can see, our model discrepancies are not significantly different from the existing reported performance in the literature.

Table 5 Discrepancies in performance metrics calculated by subtracting Table 3 from Table 4

Model	Detail	ΔAccuracy	ΔSpecificity	ΔPrecision	ΔRecall	ΔF1-score	ΔAUC
Tree	Fine Tree	16.21%	22.35%	23.64%	10.08%	15.78%	15.12%
	Medium Tree	16.21%	22.35%	23.64%	10.08%	15.78%	15.12%
	Coarse Tree	20.13%	25.48%	29.20%	14.78%	20.64%	23.75%
Logistic Regression	Logistic Regression	7.41%	-15.62%	-16.12%	30.45%	15.17%	4.94%
Naïve Bayes	Gaussian Naïve Bayes	4.30%	-1.32%	0.34%	9.93%	5.11%	-3.19%
	Kernel Naïve Bayes	12.16%	-6.25%	-6.90%	30.58%	18.57%	-3.14%
SVM	Linear SVM	11.20%	-3.12%	-3.33%	25.53%	14.69%	1.28%
	Quadratic SVM	10.24%	-6.25%	-6.45%	26.73%	14.10%	1.33%
	Cubic SVM	6.00%	-6.25%	-7.14%	18.25%	9.37%	-0.93%
	Fine Gaussian SVM	-6.85%	-19.60%	-7.15%	5.90%	-3.22%	-11.73%
	Medium Gaussian SVM	5.01%	-9.37%	-9.68%	19.40%	7.87%	0.41%
	Coarse Gaussian SVM	6.91%	0.00%	0.00%	13.83%	9.45%	0.36%
Ensemble	Boosted Trees	-3.12%	-43.75%	N/A	37.50%	N/A	N/A
	Bagged Trees	19.85%	-15.62%	-14.71%	55.33%	35.70%	10.64%
	RUS Boosted Trees	-3.12%	-43.75%	N/A	37.50%	N/A	N/A
Neural Network (NN)	Narrow NN	2.88%	-15.62%	-15.62%	21.38%	7.08%	3.85%
	Medium NN	-6.69%	-18.75%	-18.18%	5.38%	-5.19%	-5.03%
	Wilde NN	6.03%	-12.50%	-13.33%	24.55%	11.50%	-3.22%
	Bilayered NN	3.21%	-15.62%	-16.13%	22.05%	8.17%	-1.69%
	Trilayered NN	20.50%	19.23%	21.36%	21.78%	21.62%	14.37%
Kernel	SVM Kernel	-3.12%	-37.50%	-31.58%	31.25%	7.62%	-13.72%
	Logistic regression Kernel	5.58%	-5.77%	1.06%	16.95%	8.69%	-2.15%

Note: Positive values indicate overfitting performance, and negative values indicate underfitting performance.

Table 6 The ANOVA values, the p-values of the Kruskal-Wallis test, the probability density of the Chi^2 test, and the MRMR values of the 15 clinical features

Predictors	Anova	Predictors	Kruskal-Wallis	Predictors	Chi^2	Predictors	MRMR
Swallowing Difficulty	15.7113	Swallowing Difficulty	13.2112	Swallowing Difficulty	13.3992	Swallowing Difficulty	0.1954
Peer Pressure	11.9894	Peer Pressure	10.6140	Peer Pressure	10.7625	Gender	0.1518
Gender	9.7403	Gender	8.8931	Gender	9.0156	Peer Pressure	0.1488
Allergy	7.4056	Allergy	6.9774	Allergy	7.0712	Allergy	0.0471
Yellow Fingers	6.1776	Yellow Fingers	5.9146	Yellow Fingers	5.9927	Yellow Fingers	0.0367
Anxiety	6.1119	Anxiety	5.8566	Anxiety	5.9339	Alcohol Consumption	0.0195
Wheezing	5.4478	Wheezing	5.2646	Wheezing	5.3332	Wheezing	0.0164
Alcohol Consumption	5.0433	Alcohol Consumption	4.8984	Alcohol Consumption	4.9617	Anxiety	0.0137
Chronic Disease	3.7416	Chronic Disease	3.6910	Chronic Disease	3.7371	Chest Pain	0.0084
Chest Pain	3.2339	Chest Pain	3.2083	Coughing	3.2476	Coughing	0.0073

Table 6 Cont.

Predictors	Anova	Predictors	Kruskal-Wallis	Predictors	Chi^2	Predictors	MRMR
Coughing	3.2339	Coughing	3.2083	Chest Pain	3.2476	Age	0.0000
Fatigue	2.0340	Fatigue	2.0422	Fatigue	2.0656	Smoking	0.0000
Smoking	1.5346	Smoking	1.5470	Age	1.9794	Chronic Disease	0.0000
Age	1.2622	Age	0.9810	Smoking	1.5641	Fatigue	0.0000
Shortness of Breath	0.0000	Shortness of Breath	0.0000	Shortness of Breath	0.0000	Shortness of Breath	0.0000

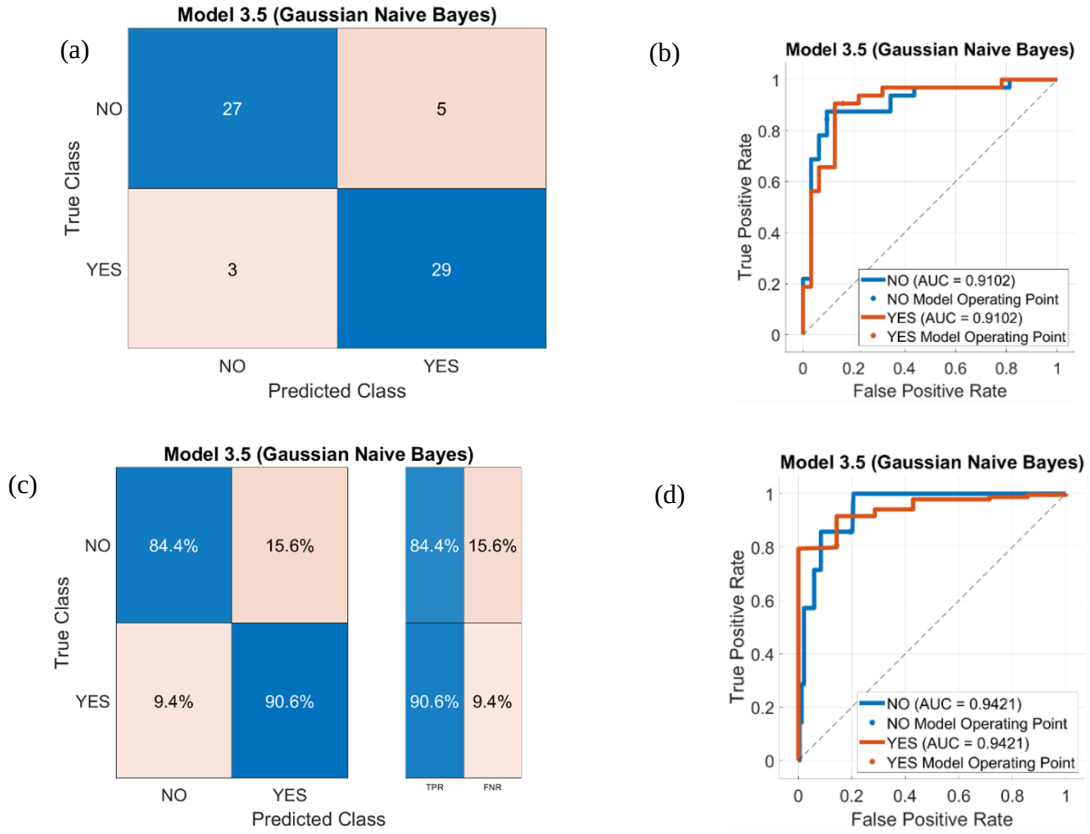

Figure 3 (a) Confusion matrices of the trained Gaussian Naïve Bayes validation using the training dataset, (b) ROC curve plots of the trained Gaussian Naïve Bayes validation using the training dataset, (c) Confusion matrices in percentage and True-False Rate Graph of the trained Gaussian Naïve Bayes when tested using the test dataset, and (d) ROC curve plots of the trained Gaussian Naïve Bayes using the test dataset

4.2 Number of Predictors Reduction Using Feature Selection Method

Several feature selection methods were then applied to the dataset using the built-in feature selection tool, including ANOVA, Kruskal-Wallis, Chi^2, and MRMR in MATLAB 2022b. The results of these unsupervised statistical analyses are shown in Table 6.

Table 6 shows that the top 3 most important features were (1) Swallowing Difficulty, (2) Peer Pressure, and (3) Gender. Although the results from different methods here show slightly different predictor rankings, the 3 highest scoring predictors remained the same for all methods. The 3 statistical methods, Chi^2, ANOVA, and Kruskal-Wallis, provided the same ranking outputs. Each predictor was employed to train Gaussian Naïve Bayes models by adding one predictor at a time and computing accuracy for validation and test cases, as explained in the materials and method section.

From Table 7, the Gaussian Naïve Bayes trained using a different number of predictors showed dramatic improvement with 3 predictors compared to 2 predictors; however, the test performance did not reach the same performance level. Therefore, the number of predictors was increased. The validation performance metrics remained within 81.25% to 85.94% for validation accuracy and 90.09% to 91.80% for validation AUC; meanwhile, the test performance gradually improved from 73.10% to 85.70% for test accuracy and 74.46% to 91.81% for test AUC. Therefore, the number of predictors that can provide optimal performance for both validation and test cases was 9 predictors, with the performance difference well below 1%. If an additional predictor was added to the training, the performance metrics deviated by nearly 3%.

Table 7 Validation and test accuracy and AUC computed from ROC curves for Gaussian Naïve Bayes models trained using 1 predictor to 10 predictors

Predictors included in the model training	Validation		Test dataset		Performance comparison	
	Accuracy	AUC	Accuracy	AUC	ΔAccuracy	ΔAUC
1 Predictors Swallowing Difficulty	79.69%	73.68%	67.45%	67.44%	12.24%	6.24%
2 Predictors Swallowing Difficulty and Peer Pressure	73.44%	84.96%	54.00%	65.88%	19.44%	19.08%
3 Predictors Swallowing Difficulty, Peer Pressure, and Gender	85.94%	90.53%	73.10%	74.46%	12.84%	16.07%
4 Predictors Swallowing Difficulty. Peer Pressure, Gender, and Allergy	81.25%	90.38%	79.85%	87.33%	1.40%	3.05%
5 Predictors Swallowing Difficulty. Peer Pressure, Gender, Allergy, and Yellow fingers	84.38%	90.48%	81.70%	87.12%	2.68%	3.36%
6 Predictors Swallowing Difficulty. Peer Pressure, Gender, Allergy, Yellow fingers, and Anxiety	82.81%	90.09%	80.05%	84.39%	2.76%	5.70%
7 Predictors Swallowing Difficulty. Peer Pressure, Gender, Allergy, Yellow fingers, Anxiety, and Wheezing	82.81%	90.97%	82.75%	88.21%	0.06%	2.76%
8 Predictors Swallowing Difficulty. Peer Pressure, Gender, Allergy, Yellow fingers, Anxiety, Wheezing, and Alcohol Consumption	81.25%	91.36%	83.40%	90.94%	-2.15%	0.42%
9 Predictors Swallowing Difficulty. Peer Pressure, Gender, Allergy, Yellow fingers, Anxiety, Wheezing, Alcohol Consumption, and Chronic Disease	82.81%	91.80%	83.80%	91.81%	-0.99%	-0.01%
10 Predictors Swallowing Difficulty. Peer Pressure, Gender, Allergy, Yellow fingers, Anxiety, Wheezing, Alcohol Consumption, Chronic Disease, Chest Pain	82.81%	91.60%	85.70%	91.45%	-2.89%	0.15%

5. Discussion

These days, the risk of getting lung cancer is increasing each year (Roland, & Rudd, 1998). We face many risks of getting lung cancer more quickly than in the past, whether from second-hand smoke or air pollution. The cost of seeing a physician to test all the symptoms and risks is unaffordable for most people. Aware of the seriousness of lung cancer, we developed this model to help physicians reduce time, predictors, and costs in diagnosing lung cancer for the patients.

However, patients may be concerned about the accuracy of diagnosing lung cancer with the existing AI models, as the accuracy rate does not reach 100%. Furthermore, teenagers and young adults can adapt to and trust modern technologies, older adults, on the other hand, prefer traditional clinical approaches over AI. However, several studies and surveys have reported that older patients are willing to adapt and employ new technologies if they are reliable, protect their data privacy (Shandilya, & Fan, 2022) and can be integrated into the existing healthcare service (Asan et al., 2020). In the next few years, technological advancements will develop drastically. Models with precise accuracy using fewer predictors help ensure the validity of traditional clinical methods, lessen the workload for physicians, and reduce healthcare costs.

Our developed artificial intelligence model can achieve a classification accuracy of 82.81% and requires only 9 predictors from the 15 variables available in the dataset. We trained the Linear SVM model due to its overall performance during the 15-variables training and found that its accuracy decreased by over 10% during testing. Therefore, we tested another model called Gaussian Naive Bayes, which produced a satisfactory outcome.

Compared to results in the literature, as discussed in the result section, similar accuracy levels were achieved, indicating that our model's accuracy is comparable for lung cancer screening. However, improvement and optimization are still needed. For future study, if anyone is interested in developing our model further, they should find appropriate datasets with more observations to achieve higher accuracy. In our study, the model we developed still has not been predicted clearly, with only 82.81% accuracy. Therefore, finding a dataset that uses more than 15 parameters and more than 309 cases will help achieve higher accuracy. We still recommend that every clinical physician or staff member use the standard tumor, node, and metastasis staging system (TNM) for a precise prediction along with our model.

6. Conclusion

We have utilized the dataset from Kaggle, containing 309 observations with 15 variables and machine learning models to identify crucial predictors for diagnosing lung cancer. First, we rearranged the data into training data (64 cases) and test data (245 cases). Then, the training data were used to train the machine learning models. First, we trained the Linear SVM dataset, resulting in 93.75% accuracy and 96.67% precision. After the training session, we tested the model again with a testing dataset; in this part, achieving only 82.55% accuracy. The accuracy of the Linear SVM model did not yet satisfy our requirement as it was overfitted. Hence, we tested the second model, Gaussian Naïve Bayes. Gaussian Naïve Bayes became the most accurate model we found. After training and testing the machine learning models, we identified 9 important variables from 15 variables and 309 case studies, with an accuracy of 82.81%. These crucial features include Swallowing Difficulty, Peer Pressure, Gender, Allergy, Yellow fingers, Anxiety, Wheezing, Alcohol Consumption, and Chronic.

7. Acknowledgements

We acknowledge the dataset used in this study, obtained from the Executive Director at Saitama City Foundation for Business Creation, Saitama, by Tetsuya Sasaki, and downloaded from a public database website, Kaggle, on February 27, 2024. We express our deepest gratitude to Satriwithaya School and the Institute of Research, Rangsit University, for the research grant.

8. References

Arroliga, A. C., & Matthay, R. A. (1993). The role of bronchoscopy in lung cancer. *Clinics in Chest Medicine*, *14*(1), 87-98. https://doi.org/10.1016/S0272-5231(21)01150-3

Asan, O., Bayrak, A. E., & Choudhury, A. (2020). Artificial intelligence and human trust in healthcare: focus on clinicians. *Journal of Medical Internet Research, 22*(6), Article e15154. https://doi.org/10.2196/15154

Bukhari, A., Kumar, G., Rajsheker, R., & Markert, R. (2017). Timeliness of Lung Cancer Diagnosis and Treatment. *Federal Practitioner, 34*(1), 24S-29S.

http://www.ncbi.nlm.nih.gov/pmc/articles/pmc6375422/

Cao, Z. (2017, May 13-14). *Development and Application of Artificial Intelligence* [Conference presentation], Proceedings of the 2nd International Conference on Mechatronics Engineering and Information Technology (ICMEIT 2017), Dalian, China. https://doi.org/10.2991/icmeit-17.2017.79

Cegla, P., Bos-Liedke, A., Burchardt, E., Konstanty, E., Piotrowski, A., Kozak, M., & Cholewinski, W. (2023). Diagnosis and treatment of lung cancer using nuclear medicine techniques—current state of the art. *Nuclear Medicine Review*, *26*, 77-84. https://doi.org/10.5603/NMR.2023.0010

Chiu, H. Y., Chao, H. S., & Chen, Y. M. (2022). Application of Artificial Intelligence in Lung Cancer. *Cancers*, *14*(6), 1-17. https://doi.org/10.3390/cancers14061370

Gould, M. K., Huang, B. Z., Tammemagi, M. C., Kinar, Y., & 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), 445-453. https://doi.org/10.1164/rccm.202007-2791OC

Karabulut, E. M., Özel, S. A., & İbrikçi, T., (2012). A comparative study on the effect of feature selection on classification accuracy. *Procedia Technology*, *1*(2012), 323-327. https://doi.org/10.1016/j.protcy.2012.02.068

Ketkomol, P., Songsak, T., Jongrungruangchok, S., Madaka, F., & Pradubyat, N. (2024). The Effect of 1'-acetoxychavicol Acetate on A549 Human Non-small Cell Lung Cancer. *Journal of Current Science and Technology*, *14*(2)., Article 43. https://doi.org/10.59796/jcst.V14N2.2024.43

Kim, J., Lee, H., & Huang, B. W. (2022). Lung Cancer: Diagnosis, Treatment, Principles, and Screening. *Clinical Presentation and Diagnosis*, *105*(5), 1-2. https://www.binasss.sa.cr/mayo/2.pdf

Lakshmanaprabu, S., Mohanty, S., Shankar, K., Arunkumar, N., & Ramírez-González, G. (2019). Optimal deep learning model for classification of lung cancer on CT images. *Future Generation Computer Systems*, 92, 374-382. https://doi.org/10.1016/j.future.2018.10.009

Lareau, S., Slator, C., & Smyth, R. (2021). Lung Cancer. *American Journal of Respiratory and Critical Care Medicine*, *204*(12), 21-22. https://doi.org/10.1164/rccm.20411P21

Li, C., Lei, S., Ding, L., Xu, Y., Wu, X., Wang, H., ... & Li, L. (2023). Global burden and trends of lung cancer incidence and mortality. *Chinese Medical Journal*, *136*(13), 1583-1590. https://doi.org/10.1097/CM9.0000000000002529

Liu, S., Liu, S., Zhang, C., Yu, H., Liu, X., Hu, Y., ... & Fu, Q. (2020). Exploratory Study of a CT Radiomics Model for the Classification of Small Cell Lung Cancer and Non-small-Cell Lung Cancer. *Frontiers in Oncology*, *10*, Article 1268, 1-11. https://doi.org/10.3389/fonc.2020.01268

Nooreldeen, R., & Bach, H. (2021). Current and Future Development in Lung Cancer Diagnosis. *International Journal of Molecular Sciences*, *22*(16), Article 8661, 1-18. https://doi.org/10.3390/ijms22168661

Pacurari, A. C., Bhattarai, S., Muhammad, A., Avram, C., Mederle, A. O., Rosca, O., ... & Mavrea, A. (2023). Diagnostic Accuracy of Machine Learning AI Architectures in Detection and Classification of Lung Cancer: A Systematic Review. *Diagnostics*, *13*(13), Article 2145. https://doi.org/10.3390/diagnostics13132145

Pechprasarn, S., Manavibool, L., Supmool, N., Vechpanich, N., and Meepadung, P. (2023a). Predicting Parkinson's Disease Severity using Telemonitoring Data and Machine Learning Models: A Principal Component Analysis-based Approach for Remote Healthcare Services during COVID-19 Pandemic. *Journal of Current Science and Technology*,*13*(2), 465-485. https://doi.org/10.59796/jcst.V13N2.2023.694465

Pechprasarn, S., Wattanapermpool, O., Warunlawan, M., Homsud, P., & Akarajarasroj, T.(2023b). Identification of Important Factors in the Diagnosis of Breast Cancer Cells Using Machine Learning Models and Principal Component Analysis. *Journal of Current Science and Technology*, *13*(3), 642-656. https://doi.org/10.59796/jcst.V13N3.2023.700

Pereira, T., Freitas, C., Costa, J. L., Morgado, J., Silva, F., Negrão, E., ... & Oliveira, H. P. (2020). Comprehensive Perspective for Lung Cancer Characterisation Based on AI Solutions Using CT Images. *Journal of*

Clinical Medicine, *10*(1), Article 118. https://doi.org/10.3390/jcm10010118

Rana, M., & Bhushan, M. (2023). Machine learning and deep learning approach for medical image analysis: diagnosis to detection. *Multimedia Tools and Applications*, *82*(17), 26731-26769.. https://doi.org/10.1007/s11042-022-14305-w

Roland, M., & Rudd, R.M. (1998). Genetics and Pulmonary Medicine, Somatic mutation in development of lung cancer. *Thorax*, *53*, 979-983. https://doi.org/10.1136/thx.53.11.979

Ruano-Raviña, A., Provencio, M., Calvo de Juan, V., Carcereny, E, Moran, T, Rodriguez-Abreu, D., …, & Cerezo, S. (2020). Lung cancer symptoms at diagnosis: results of a nationwide registry study. *ESMO Open*, *5*(6), Article e001021. https://doi.org/10.1136/esmoopen-2020-001021

Sankar, V., Kothai, R., Vanisri, N., Akilandeswari, S., & Anandharaj, G., (2023). Lung Cancer, A Review. *International Journal of Health Sciences and Research*, *13*(10), 307-315. https://doi.org/10.52403/ijhsr.20231042

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), Article 160. https://doi.org/10.1007/s42979-021-00592-x

Sasaki, T., (2020). *Lung Cancer Classify Model & Positive Clustering*. Retrieved, from https://www.kaggle.com/code/sasakitetsuya/lung-cancer-classify-model-positive-clustering

Shandilya, E., & Fan, M. (2022, October 22-23). *Understanding older adults' perceptions and challenges in using AI-enabled everyday technologies* [Conference presentation]. Proceedings of the Tenth International Symposium of Chinese CHI, Guangzhou, China. https://doi.org/10.48550/arXiv.2210.01369

Sherry, V. (2022). Lung cancer: Prevention and early identification are key. *The Nurse Practitioner*, *47(7)*, 42-47. https://doi.org/10.1097/01.NPR.0000832548.88417.be

Singh, G. A. P., & Gupta, P. K. (2019). Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans. *Neural Computing and Applications*, *31*(10), 6863-6877.. https://doi.org/10.1007/s00521-018-3518-x

Sowmya, C., Kumar, A. G., & Kumar, S. (2021). Stacked LSTM Recurrent Neural Network: A Deep Learning Approach for Short Term Wind Speed Forecasting. International Conference on Intelligent Technologies (CONIT), Hubli, India, 2021, pp. 1-7. https://doi.org/10.1109/CONIT51480.2021.9498314.

Teramoto, A., Tsukamoto, T., Kiriyama, Y., & Fujita, H. (2017). Automated classification of lung cancer types from cytological images using deep convolutional neural networks. *BioMed Research International*, *2017*(1), Article 4067832. https://doi.org/10.1155/2017/4067832

Thandra, K. C., Barsouk, A., Saginala, K., Aluru, J. S., & Barsouk, A. (2021). Epidemiology of lung cancer. *Contemporary Oncology /Współczesna Onkologia*, *25(1)*, 45-52. https://doi.org/10.5114/wo.2021.103829

Tunali, I., Gillies, R. J., & Schabath, M. B. (2021). Application of Radiomics and Artificial Intelligence for Lung Cancer Precision Medicine. *Cold Spring Harbor Perspectives in Medicine*, *11*(8), Article a039537. https://doi.org/10.1101/cshperspect.a039537

Vidaver, R. M., Shershneva, M. B., Hetzel, S. J., Holden, T. R., & Campbell, T. C. (2016). Typical time to treatment of patients with lung cancer in a multisite, US-based study. *Journal of Oncology Practice*, *12(6)*, e643-e653. https://doi.org/10.1200/JOP.2015.009605

Walser, T., Cui, X., Yanagawa, J., Lee, J. M., Heinrich, E., Lee, G., ... & Dubinett, S. M. (2008). Smoking and Lung Cancer, The Role of Inflammation. *Proceedings of the American Thoracic Society*, *5*(8), 811-815. https://doi.org/10.1513/pats.200809-100TH

Wang, S., Yang, D. M., Rong, R., Zhan, X., Fujimoto, J., Liu, H., ... & Xiao, G. (2019). Artificial Intelligence in Lung Cancer Pathology Image Analysis. *Cancers*, *11*(11), Article 1673. https://doi.org/10.3390/cancers11111673