MACHINE LEARNING APPROACH OF LUNG CANCER PREDICTION USING MULTI MACHINE LEARNING MODELS

Conference Paper · March 2025

Luxshi Karunakaran
Sabaragamuwa University of Sri Lanka

Data Science Track

ComURS2025 Computing Undergraduate Research Symposium 2025

19th February 2025, Sabaragamuwa University of Sri Lanka, Belihuloya

Paper ID:44

Publication: https://www.comurs.sab.ac.lk/abstractBook.html

MACHINE LEARNING APPROACH OF LUNG CANCER PREDICTION USING MULTI MACHINE LEARNING MODELS

K. Luxshi*, Professor R.M.K.T. Ratnayaka*, Miss. AMCM. Aruppola*

* *Department of Physical Sciences and Technology, Faculty of Applied Sciences,*
Sabaragamuwa University of Sri Lanka, P. O. Box 02, Belihuloya.
klluxshi99@gmail.com

ABSTRACT

Lung cancer is still one of the leading causes of cancer deaths worldwide. Cancer detection is a complex and challenging process; however, when identified at an incipient stage, it is amenable to curative interventions. Machine learning is one of the most promising artificial intelligence methods widely used in oncological diagnosis and detection of early stages of disease. Thus, this study assesses the application of machine learning to predict lung cancer for symptom-based diagnosis. We introduce a novel approach by evaluating deep learning methods, whereas previous research primarily relied on traditional machine learning models. The numerical dataset in the Kaggle repository was preprocessed to ensure the quality of inputs and the holdout method was used to evaluate the model performance. Various implemented like Logistic Regression, Decision Trees, Gradient Boosting, SVM, ANN, Random Forest, XGBoost, Linear Regression, Naive Bayes, and LSTM. The models were evaluated using Accuracy, Sensitivity, Specificity, and ROC-AUC matrices. The SVM model outperformed all the other models with an accuracy value of 96.98% followed by ANN (96.82%) and LSTM (96.52 %) models. SVM recorded higher sensitivity compared to ANN (98.48%) and LSTM (98.99%). However, SVM and LSTM recorded a lower specificity value of 74.24%, whereas ANN recorded a highest value of 81.81%. The ROC-AUC was highest for both LSTM and ANN (99.32%) while SVM resulted in 98.58%. These results show that machine learning algorithms can classify lung cancer with acceptable accuracy which opens the way toward the improvement of clinical diagnosis. This study emphasizes the usage of machine learning models like SVM in clinical practice to improve the detection rate of early lung cancer as a binary classification task. Advanced machine learning algorithms can be further finetuned and coupled with different cross-validation methods to check the suitability to detect lung cancers in the future.

Keywords: Lung cancer, Machine learning, Numerical data, Predictive analysis, SVM

I. INTRODUCTION

Worldwide, lung cancer accounts for the majority of fatalities caused by the disease. People who smoke and people who don't smoke are both affected by the repercussions. Carcinoma is the medical term for lung cancer. Cancer originates in cells referred to as epithelial cells. When lung cells mutate or when tissues experience unregulated cell proliferation, this may lead to the development of lung cancer. The lungs are the main organs of respiration. The human body has two lungs, one on each side of the chest. The left lung is smaller than the right, leaving room for the heart. During breathing, the chest rises and falls. That is because by inhalation, the lungs swell, and by exhalation, they shrink. The lungs are responsible for enriching the blood with oxygen. The heart sends to the lungs blood that is low in oxygen and rich in carbon dioxide. The blood inside the lungs is “cleansed”, absorbs oxygen and leaves carbon dioxide. Carbon dioxide is eliminated during exhalation, while oxygen enters the lungs during inhalation

In many countries, the number of former smokers is high, and many types of lung cancer concern former smokers as well. In the United States alone, there are more than millions of former smokers (people who have already stopped smoking), approaches such as lung cancer screening are evidence-based measures to detect and cure lung cancer before the development of lethal metastatic spread in current and former smokers. Supporting smoking cessation is important for current smokers, but lung cancer is a lifelong risk for every smoker. The patient's risk of dying of lung cancer is determined by the advanced stage of cancer. If someone identifies it in the early stages, it can even be cured, while, at an advanced stage, median survival is less than two years. The early-stage detection of lung cancer is associated with a high frequency of cure, whereas lung cancer detected in higher stages is often associated with a median survival of less than years.

Nowadays, artificial intelligence (AI) and machine learning (ML) techniques play a critical role in healthcare. Due to the wide applicability of AI/ML in numerous health conditions' risk prediction, a variety of regulations should be determined a sin to evaluate and support the practical development of AI/ML-based software tools for the early prediction and diagnosis of a disease.

In the context of this study, for this particular disease, many scientific studies have been executed from the perspective of ML. Here, a methodology for designing effective ML classification models is presented to predict lung cancer occurrence with the aid of the most common habits and symptoms/signs as input features to the models. Our contribution is a comparative assessment of numerous classifiers to develop the intended model with the highest sensitivity and discrimination ability in identifying those at high risk. For the evaluation of the models, we considered the performance metrics precision, recall, F-Measure, accuracy and AUC. Moreover, AUC ROC curves are also captured and presented. Finally, from various aspects, the performance analysis revealed that SVM is the most efficient model, and therefore constitutes the main proposition of this article.

II. RELATED WORK

Like other domain, researchers have successfully implemented machine learning techniques to construct prediction models context of certain diseases such as lung cancer.

Firstly, the authors demonstrated Lung cancer [14] types using machine learning approach employed prediction and classification, different techniques are used such as K-Nearest Neighbor (KNN), Decision Tree (DT), Random Forest, Adaptive Boosting (AdaBoost). The proposed mechanism used machine learning approach different types of lung cancer based on statistical and textural features. The dataset that observed images are Adenocarcinoma-325, large cell carcinoma-191, squamous cell carcinoma-272 and normal images are 79. After the extraction normalize the sample using min-max normalization technique. In this paper focused on supervised machine learning models. The overall testing accuracy value of decision tree 87.26%, random forest 87.27%, random forest 87.27%, KNN 87.26%, and AdaBoost 90.74%. The AdaBoost is best suitable for the detecting different types of lung cancer. The overall accuracy is 90.74% which is best as compared to DC, RF and KNN. The sensitivity is 81.80%, specificity is 93.99%, F1 is 0.81 and Kappa is 0.753 and AUC is 0.93.

Similarly, the authors applied Lung Cancer risk prediction using Text dataset [9]work with several ML models used such as Logistic regression, Decision tree, Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Navie Bayes. The data preprocessing involved find out the missing and noisy data techniques used SMOTE (Synthetic Minority Over-sampling Technique). In this paper, Random Forest (RF) model best performed other models. The accuracy of RF model 90.32%, precision 89.82%, recall 90.32%, and F1 score 89.29%, making effective model for the lung cancer risk.

Other domain Prediction of lung cancer [13] using WEKA implements algorithms for data pre-processing, feature reduction, classification such as Naive Bayes, Bayesian Network, J48. The performances of the algorithms for lung cancer disease are analyzed using visualization tools. Class is predicting the risk attributes for distinct in Label Low, Medium and High. In this risk prediction 30 patients in low level risk, 33 patients in medium level risk, 37 patients in high level risk. In Navie bayes 0.01s, 100 instances are correctly, The Bayesian network algorithm in 0.03s, all the instances are correctly classified and J48 algorithm builds the prediction in 0.06s, all the instances are correctly classified. In this time taken algorithms to build model in WEKA tool. Navie Bayes algorithm is the best performance algorithm based on the time.

Moreover, in the Evaluation of Utilization for Lung cancer classification [8] based on Gene Expression levels are the International Agency for research on cancer there were about 13% (1,825 thousand) of new lung cancer cases of the total number of new cancer cases and about 19.4% (1,590 thousand) deaths of the total number of deaths owing to lung cancer in the world in 2012. The population screening machine learning methods are used to differentiate between benign and malignant lung nodules based on low-dose computed tomography which is considered as a widespread standard in detecting and analysis of lung diseases. The dataset processed four publicly available dataset related to gene expression. In Dana-Farber Cancer Institute, Harvard Medical School (Bhattacharjee) consists of 203 samples, 139 correspond with adenocarcinoma, 21 squamous cell lung carcinomas. The University of Michigan consists of 96 samples, 86 primary adenocarcinoma, 10 non-neoplastic tissue. The University of Toronto, Ontario, Canada consists of 39 samples of non-cancer cell lung cancer 24 samples correspond to patients with lung cancer, remaining 15 patients are disease-free. Then Brigham and Woman's Hospital, Harvard Medical School consists of 181 samples of malignant tissue. Where 31- malignant pleural mesothelioma and 150 adenocarcinomas. The expression level of each and every Institute have 12600, 7129, 2880,12533 genes. Comparing machine learning methods such as AUC, MCC for each institute and using machine learning algorithms like k-NN (k=1), k-NN (k=5), k-NN (k=10), NB-normal, NB-histogram, SVM, C4.5 Decision tree. Compared machine learning algorithms SVM tends to the most appropriate auxiliary tool in lung cancer screening while others showed sufficient effectiveness to use in the tasks of gene expression level assessment.

The authors in aimed to build a Machine learning approach [4] in Lung cancer types non-small cell lung cancer (NCLC) divides into three sub-parts adenocarcinoma (AC), squamous cell carcinoma (SCC), large cell carcinoma (LCC). The most common type of lung cancer is non-small cell lung cancer it will take 85%-90% and 10-15% of the cases are diagnosed with small cell lung cancer. The prediction and classification, different techniques are used such as K-Nearest Neighbor (KNN), Decision Tree (DT), Random Forest, Adaptive Boosting (AdaBoost). The proposed mechanism used machine learning approach different types of lung cancer based on statistical and textural features. The dataset that observed images are Adenocarcinoma-325, large cell carcinoma-191, squamous cell carcinoma-272 and normal images are 79. After the extraction normalize the sample using min-max normalization technique. In this paper focused on supervised machine learning models. The overall testing accuracy value of decision tree 87.26%, random forest 87.27%, random forest 87.27%, KNN 87.26%, and AdaBoost 90.74%. The AdaBoost is best suitable for the detecting different types of lung cancer. The overall accuracy is 90.74% which is best as compared to DC, RF and KNN. The sensitivity is 81.80%, specificity is 93.99%, F1 is 0.81 and Kappa is 0.753 and AUC is 0.93.

Finaly, the authors designed to mechanism [7] to the two main types are small-cell lung carcinoma (SCLC) and non-small-cell lung carcinoma (NSCLC). The most common symptoms are coughing (including coughing up blood), weight loss, shortness of breath and chest pains. The vast majority (85%) of cases of lung cancer are due to long-term Tabacco smoking. Lung cancer may be seen on chest radiographs and computed tomography (CT) scans. The diagnosis is confirmed by biopsy, which is usually performed by bronchoscopy or CT-guidance. The proposed lung cancer prediction model consists of four different steps such as image pre-processing, image segmentation, feature extraction and image classification. Classification of images is a basic task that seeks to interpret a picture as a whole. By assigning it to a particular label the purpose is to identify the image. The dataset consists of 14 features like Gender, Age, Smoker, tumor location, t-stage, n-stage, stage, timing, diabetes, status, meal.cal, wt.loss, ph,ecog, and ph.karno using UCI machine learning repository. The proposed lung cancer prediction model is implemented for the python programming with keras and TensowrFlow2. The effectiveness of the proposed technique is demonstrated using different evaluation metrics measuring true and misclassification of lung cancer positive/negative cases. The ML techniques such as SVM, RF, NB, ANN, and GNB. The performance analysis shows GNB prediction model achieves 98% accuracy, 92% precision, 97% sensitivity, 98% specificity compare to other machine learning techniques.

This review seeks to address the following Research Questions (RQs).

1. *RQ1: Which algorithms of machine learning work best for predicting lung cancer?*
2. *RQ2: Which data preprocessing steps are important to improve the accuracy when predicting lung cancer?*
3. *RQ3: What do we use in measuring the performance of the lung cancer prediction models?*
4. *RQ4: What is the difference in performance of various models in regard to accuracy, sensitivity, and specificity while predicting lung cancer?*
5. *RQ5: What obstacles do one come across when deploying a particular machine learning models for lung cancer diagnosis?*

III. METHODOLOGY

The major methods followed for the cancer prediction study are discussed in the section.

1. Sources

Literature search is an important part while doing research and it is always better to get hold of research papers for the study first. The following standard data libraires are used to find research papers article and study-related to the topic.

- ResearchGate (https://www.researchgate.net/search)
- ScienceDirect (https://www.sciencedirect.com)
- IEEE Xplore (https://ieeexplore.ieee.org/Xplore/home.j)
- Google Scholar (https://scholar.google.com/)

Support Vector machine (SVM), Data Preprocessing in healthcare, Cost and Benefit analysis in machine learning, forecasting of lung cancer using ML, Other mechanisms of ML in the diagnosis of cancer. Hybrid Machine learning models related to the detection of lung cancer. Every data pre-processing technique necessary in healthcare

AI innovation was presented in this study. Limit of detection, upper limit of quantification, carryover contamination in lung cancer models challenges of Applying AI for Diagnosis of Cancer before starting research.

2. Terms and Search Strings

2.1. Key Terms

- Lung Cancer Predictions
- Machine Learning models
- Artificial Neural Networks (ANN)
- Support Vector machine (SVM)
- Random Forest
- Gradient Boosting
- XGBoost
- Logistic Regression
- Decision Tree
- Linear Regression
- Navie Bayes
- LSTM (Long-Short Term Memory)
- Data Preprocessing in healthcare
- Evaluation metrics in machine learning.

2.2. Search Strings

- Lung cancer prediction using machine learning
- Machine learning algorithms for cancer diagnosis
- Hybrid Machine learning models for lung cancer detection.
- Data preprocessing techniques in healthcare AI
- Accuracy, sensitivity, specificity in lung cancer models
- Challenges in implementing AI for cancer diagnosis.

3. Inclusion and Exclusion Criteria

3.1. Inclusion Criteria

- Study Relevance: Studies that focus on lung cancer prediction, Studies that implement and discuss machine learning models in healthcare, particularly for cancer diagnosis.
- Algorithm used: Papers that use relevant machine learning algorithms such as Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost, SVM, Navie Bayes, Linear Regression and Artificial Neural Network and LSTM.
- Evaluation metrics: Studies reporting performance metrics such as accuracy, sensitivity, specificity, precision, recall, and AUC-ROC for model assessment.
- Data preprocessing: Studies that include data processing steps like a categorical data transformation for relevant to improving model performance.

3.2. Exclusion Criteria

- Irrelevant Cancer Types: Studies focused solely on cancers other than lung cancer (e.g., breast, liver, or skin cancer).
- Non-Machine Learning Approaches: Papers that do not utilize machine learning or AI methods for prediction or diagnosis.
- Lack of Performance Metrics: Studies that do not provide sufficient evaluation metrics (e.g., studies without accuracy, sensitivity, or specificity).

The strategy followed in this work is drawn in Figure 1.

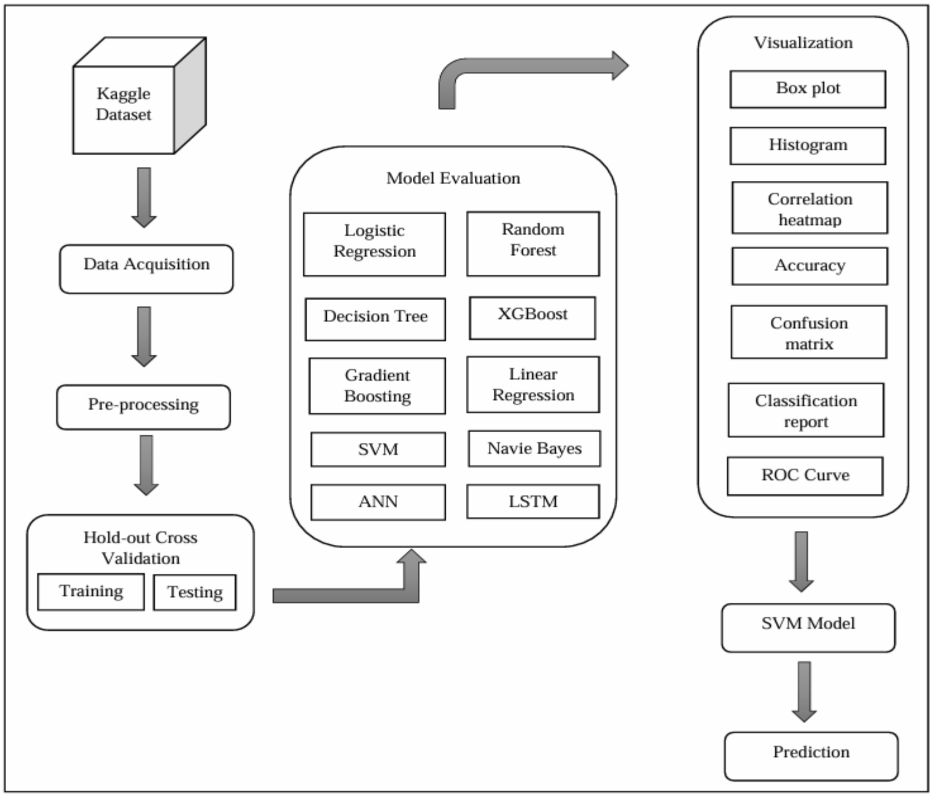


Figure 1 : Block Diagram of Strategy of Lung cancer prediction

1. Dataset Acquisition

The dataset used in this article is obtained from Lung cancer dataset Kaggle website. This dataset consists 3309 instances with one target class which mean whether the person is lung cancer or not. The features listed in this csv dataset have 16 attributes are described as follows: (All the features are numerical values)

- GENDER: Gender of the patient [M/F]
- AGE: Age of the patient
- SMOKING: Whether the patient is a smoker [Y/N]
- YELLOW_FINGERS: Yellow fingers due to smoking [Y/N]
- ANXIETY: Anxiety levels [Y/N]
- PEER_PRESSURE: Peer pressure experienced [Y/N]
- CHRONIC DISEASE: Presence of chronic diseases [Y/N]
- FATIGUE: Fatigue levels [Y/N]
- ALLERGY: Presence of allergies [Y/N]
- WHEEZING: Wheezing sounds [Y/N]
- ALCOHOL CONSUMING: Alcohol consumption [Y/N]
- COUGHING: Presence of coughing [Y/N]
- SHORTNESS OF BREATH: Shortness of breath [Y/N]
- SWALLOWING DIFFICULTY: Difficulty swallowing [Y/N]
- CHEST PAIN: Chest pain [Y/N]
- LUNG_CANCER: Lung cancer diagnosis [Y/N]

2. Preprocessing

Note that, no preprocessing was performed on this dataset, relied on as there are no missing values or outliers. The categorical data are changed into the numerical values. (AGE and LUNG_CANCER are the categorical values).

3. Cross validation

In this dataset used Hold out cross validation method. Hold out method mean when split the dataset into training and testing, the training set is what the model is trained on and the testing set is used to see how well that model performs on unseen data. In this section the hold-out method using 80% for training and the remaining 20% of the data for testing.

4. Model Evaluation

In this article, for the topic under consideration, various machine learning models were employed in order to identify which one performs better than other algorithms evaluating their performance.

4.1. Logistic Regression

Logistic regression machine learning is a statistical technique used to building machine learning models where the dependent variable is dichotomous which mean binary classification. It is used to describe the data between one dependent variable and one or more independent variables. It can work with categorical and numerical data making it versatile for various applications. The sigmoid or logistic function is essential for converting predicted values into probabilities in logistic regression.

4.2. Random forest

Random forest is a popular machine learning algorithm that belong to the supervised learning technique. It can be used for both classification and regression problems in ML. It is a concept known as “ensemble learning”, which mean the process of combining multiple classifiers to solve the complex problem and improve the performance of the model. Random forest is a classifier that contains a number of decision trees on various of subsets of the given dataset and takes the average to improve the predictive accuracy of the dataset.

Figure 2 shows that Random Forest works in two phases, first into create the random forest by combining N decision trees and second is to make predictions for each tree created in the first phase. The process can be explained in the below the steps:

- Step 1: Select random K data points from the training set
- Step 2: Build the decision tree associated with the selected data points (Subsets)
- Step 3: Choose the number N for decision trees that want to build.
- Step 4: Repeat Step 1 & 2
- Step 5: For the new data points, find the predictions for each decision trees, and assign the new data points to the category wins the majority wins.

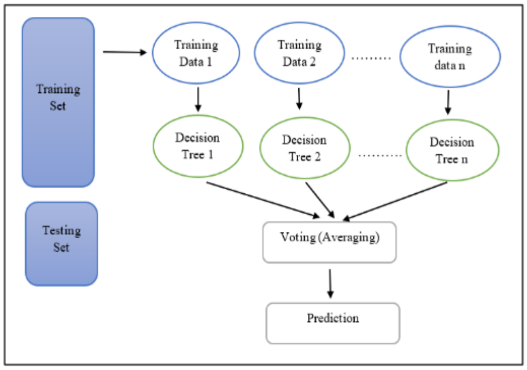


Figure 2: Random Forest Algorithm

4.3. *Decision Tree*

Decision tree is a supervised learning technique that can be both classification and regression problems but mostly used to classification technique. The decision tree for predicting the class of the given dataset, the algorithm starts from the node of the tree (Figure 3). This algorithm compares the values of root attribute with the recorded attribute and based on the comparison follows branch and jumps to the next node. For the next node, the algorithm again compares the attribute value with the other sub-notes and move further. It continues the process until it reaches the leaf node of the tree.

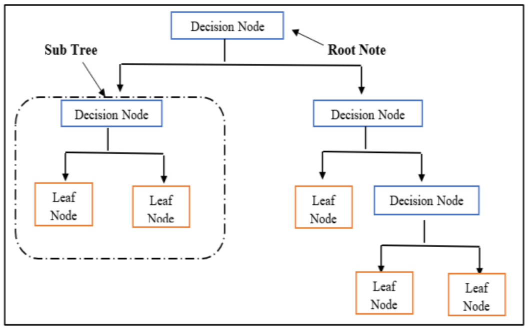


Figure3: Decision Tree

4.4. *XGBoosting*

XGBoost is an optimized distributed gradient boosting library designed for efficient and scalable training of machine learning models. It is an ensemble learning method that combine the predictions of multiple weak models to produce stronger prediction. It stands for “Extreme Gradient Boosting” and its efficient handling of missing values and handle real world data without requiring significant pre-processing.

4.5. *Gradient Boosting*

Gradient Boosting is a powerful boosting algorithm that combines several weal learners into string learners, new model is trained to minimize the loss function such as mean squared error or cross-entropy of the previous models using gradient descent.

- Step 1: Difference between the actual and the predicted variables (Input features X, target variable Y)

$$L(f) = \sum_{i=1}^{N} L(y_i, f(x_i))$$

- Step 2: Minimize the loss function L(f).

$$\hat{f}_0(x) = \underset{f}{argmin}\ L(f) = \underset{f}{argmin} \sum_{i=1}^{N} L(y_i, f(x_i))$$

$$\hat{y}_i = F_{m+1}(x_i) = F_m(x_i) + h_m(x_i)$$

- Step 3: Steepest Descent
  The steepest descent finds the loss function,

$$h_m = -\rho_m g_m$$

$$g_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x_i)=f_{m-1}(x_i)}$$

- Step 4: Solution
  The gradient similarly for M trees,

$$f_m(x) = f_{m-1}(x) + \left(\underset{h_m \epsilon H}{argmin}\left[\sum_{i=1}^{N} L(y_i, f_{m-1}(x_i) + h_m(x_i))\right]\right)(x)$$

  Current solution is,

$$f_m = f_{m-1} - \rho_m g_m$$

4.6. *Linear Regression*

Linear regression is one of the easiest and most popular machine learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continues/real or numeric variables. It shows the linear relationships between a dependent (Y) and one or more independent variables(X).

4.7. SVM (Support Vector Machine)

SVM is one of the most popular supervised learning algorithms, used for the classification and regression problems. The goal of this algorithm to create a best line or decision boundary that can be segregate n-dimensional space into classes that can easily put the new data point in the correct category.

4.8. Navie Bayes

Navie Bayes is one of the supervised learning algorithms and mainly used for the classification algorithms which help in building the fast machine learning models that can be make quick predictions. The Navie Bayes comprised of two words Navie and Bayes. Navie means it occurrence of a certain feature is independent of the occurrence of other features and Bayes depends on the principle of Baye's Theorem. It is a probabilistic classifier, which mean it predicts on the basis of the probability of an object.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

4.9. ANN (Artificial Neural Network)

ANNs are basically massive parallel computational models that imitate the function of human brain. An ANN consists of large number of simple processors linked by weighted connections. The processing nodes called "neurons". Each node output depends only on the information that is locally available at the node.

ANN can be defined based on three characteristics

- The Architecture indicating the number of layers and number of nodes in each of the layer.
- The learning mechanism applied for updating the weights of the connection.
- The activation functions used in various layers.

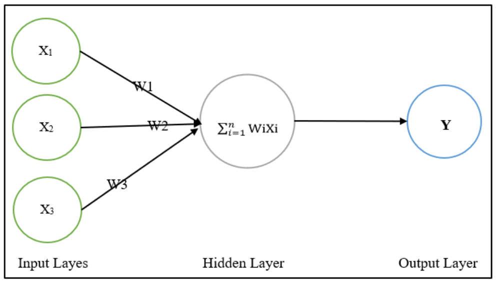


Figure 4: ANN block diagram

Figure 4 shows that information is fed into the input layer which transfers it to the hidden layer. The interconnections between the two layers assign weights to each input randomly. A bias added to every input after weights are multiplied with them individually and the weighted sum is transferred to the activation function. The activation function determines which nodes it should fore for feature extraction. The model applies an application function to the output layer to deliver the output. Weights are adjusted and the output is back-propagated to minimize error. The figure 5 is shown details about the ANN structure.

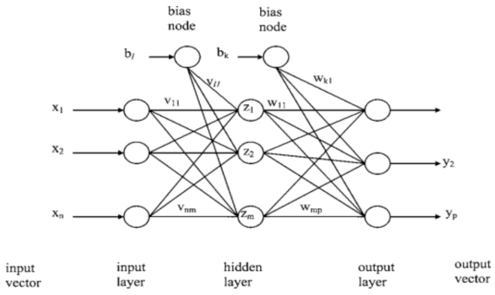


Figure 5: ANN Structure

4.10. LSTM (Long Short-Term Memory Networks)

The LSTM involves the memory cell which is controlled by three gates: Input gate, Forget gate and the output gate. These gates decide what the information to add to remove from output from the memory cell. The LSTM networks to selectively retain or discard information as it flows through the network, which allows them to learn long-term dependencies.

Advantages and Disadvantages of Machine learning models.

Algorithms	**Advantages**	**Disadvantages**
Logistic Regression	• Easier to implement, interpret and very efficient to train. • It is easily extended to multiple classes (multi-model regression).	• It may lead overfitting, • Limitation of linearity between the dependent variable and independent variable
Random Forest	• High Accuracy • Robustness to noise	• It consumes lot of memory • It takes longer time to prediction
Decision Tree	• Lesser effort for data preprocessing • Does not require normalization of data	• High variance algorithm • Highly time-consuming in training process
XGBoosting	• Fast and efficient for large datasets • Prevent the overfitting	• It can be memory-intensive for large datasets • Lack of transparency
Gradient Boosting	• Support handling categorical features • Train faster on large datasets	• Prone to overfitting • Computationally expensive and take a long time to training.
Linear Regression	• Simple Implement and easier to interpret • Susceptible to overfitting.	• Prone to Underfitting • Assumes that data is independent
SVM	• More effective in high dimensional spaces	• It is not suitable for the large dataset • It does not perform well when the dataset has more noise
Navie Bayes	• Simple and easy to implement • Fast and used for real-time predictions	• Zero-frequency problem
ANN	• Ability to handle complex data • Non-linear modeling capabilities	• Need large amounts for training data • Black box nature • Prone to overfitting
LSTM	• Avoiding Vanishing Gradient Problems • Handling variable length sequences	• Prone to overfitting • Computational complexity

Table 1 : Pros and Cos of Machine Learning Algorithms

5. *Visualization*

5.1. Box plot

The Age by lung cancer using box plot ([Figure 6](#)) which mean the relationship between the Age and Lunch cancer patient. In figure 6 shows the 40 to 70 age people are affected to the lung cancer.

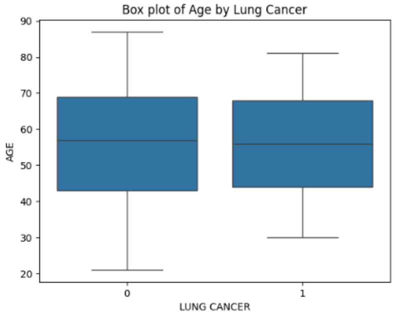


Figure 6: Box plot

5.2. Histogram

The Feature distribution for the numerical features and [Figure 7](#) shows the histogram of the numerical features of lung cancer.

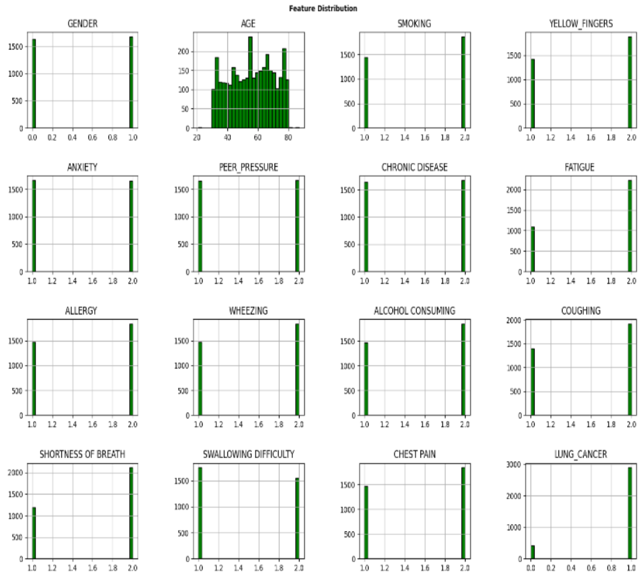


Figure 7: Histogram of features

5.3. Correlation Heatmap

Feature correlation is the inevitable step which indicates the contribution of each feature to the target class. Correlation can be positive or negative of relevant features with the class. The correlation coefficient between -1 to +1, +1 indicates perfect positive correlation which mean stronger correlation and direct proportionally of features with the class while negative correlation tells above the inverse proportionality which mean weaker correlations. The 0 indicates no correlation between the features. To visualize the correlation values together on a plane, Heatmap is generally used by the analysis. The Heatmap for the lung cancer prediction model is shows in [Figure 8.](#)

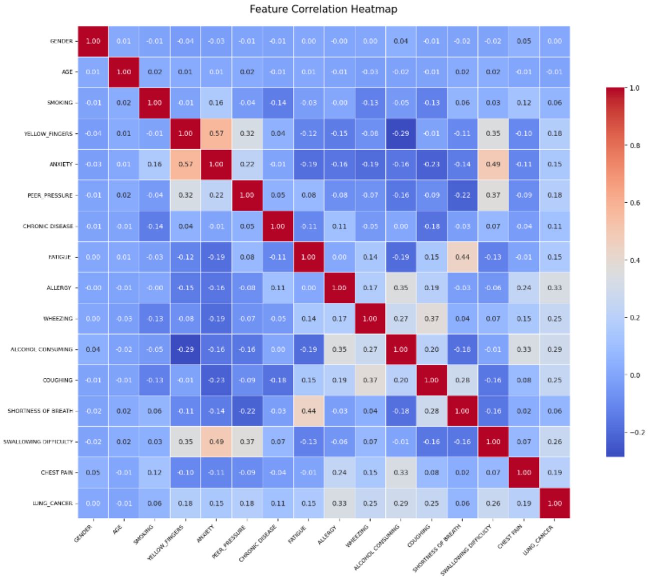


Figure 8: Heatmap

5.4. Comparison of LSTM and ANN models training history

[Figure 9](#) shows that ANN and LSTM training history which mean ANN graph shows that accuracy over 50 epochs for both training and validation part. Both accuracies are going to increasing but the validation accuracy higher than the training accuracy. When consider the LSTM depicts accuracy of model over 50 epochs, both validation and training accuracies improve over time but with more fluctuations compared to the ANN. The accuracy stabilizes close to 96%

with validation accuracy generally higher but more volatile than training accuracy.

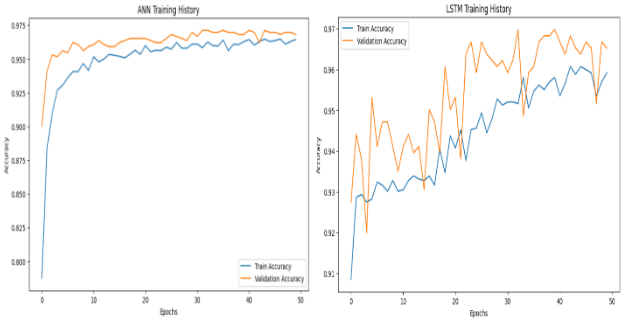


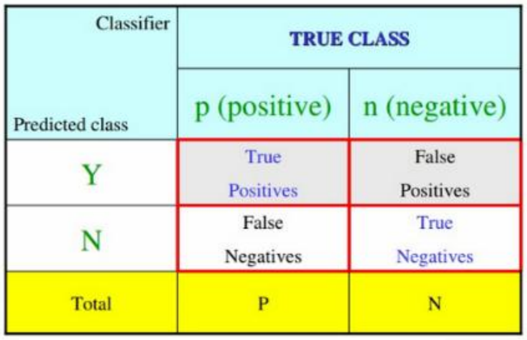
Figure 9: LSTM and ANN graph

IV. RESULTS AND DISCUSSION

Results

Results are obtained through confusion matrix (Figure 10) and classification report. Confusion matrix is generated from the binary classification outcomes. The highlighted parameters like accuracy, error rate, truth positive rate (TPR), false positive rate (FPR), truth negative rate (TNR) and false negative rate (FNR) can be calculated on the basis of this matrix. Parametric definitions are in Table 2.

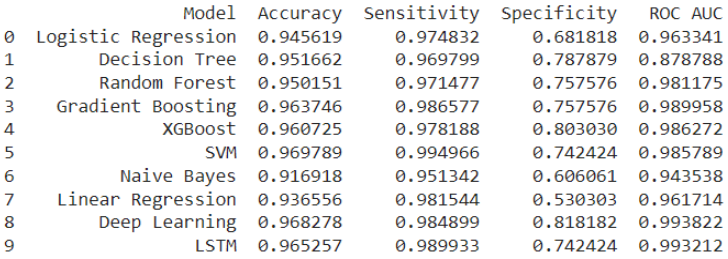
The performance of the machine models was evaluated and using variety of libraries for data preprocessing, classification, prediction and visualization. In the context for work, plenty of machine learning models such as including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost, SVM, Navie Bayes, Linear Regression, Artificial Neural Network(ANN), Long-Short Term Memory (LSTM) are evaluated in the terms of accuracy, sensitivity, specificity (Figure 11 and Figure 12), classification report and confusion matrix (Table 3), ROC curve (Figure 13), with help of these evaluation finally find out the best model (Figure 14) and finally based on the best model predict the new patient outcome (Figure 15).



Classifier Predicted class	TRUE CLASS p (positive)	 n (negative)
Y	True Positives	False Positives
N	False Negatives	True Negatives
Total	P	N

Figure 10: confusion matrix

Evaluation



```
                 Model  Accuracy  Sensitivity  Specificity   ROC AUC
0  Logistic Regression  0.945619     0.974832     0.681818  0.963341
1        Decision Tree  0.951662     0.969799     0.787879  0.878788
2        Random Forest  0.950151     0.971477     0.757576  0.981175
3    Gradient Boosting  0.963746     0.986577     0.757576  0.989958
4              XGBoost  0.960725     0.978188     0.803030  0.986272
5                  SVM  0.969789     0.994966     0.742424  0.985789
6          Naive Bayes  0.916918     0.951342     0.606061  0.943538
7    Linear Regression  0.936556     0.981544     0.530303  0.961714
8        Deep Learning  0.968278     0.984899     0.818182  0.993822
9                 LSTM  0.965257     0.989933     0.742424  0.993212
```

Figure 11: Comparing of machine learning models

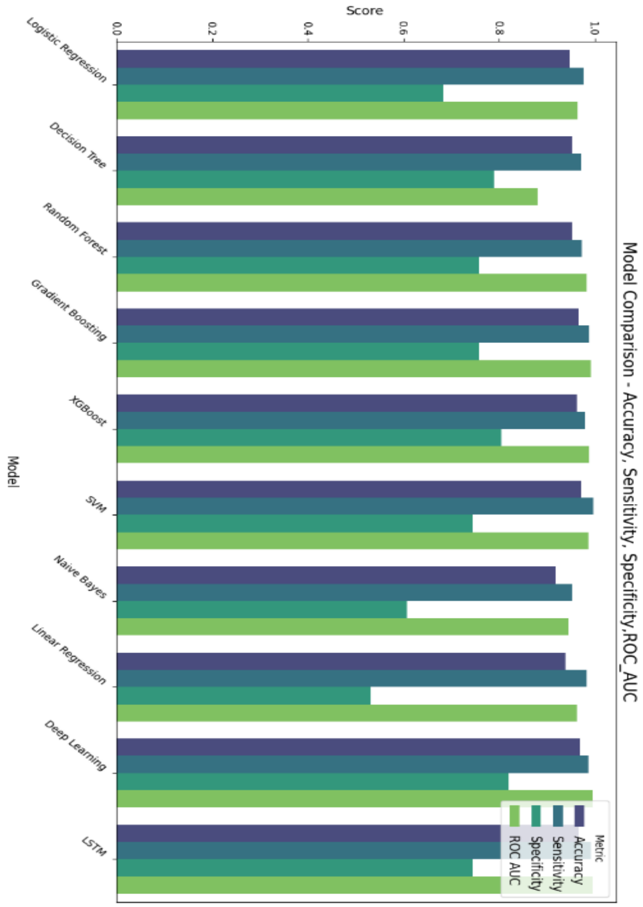


Figure 12: comparison graph for all machine learning models

No	Parameters	Definitions
1	True Positive (TP)	Correct positive predictions
2	False Positive (FP)	Incorrect Positive predictions
3	True Negative (TN)	Correct negative predictions
4	False Negative (FN)	Incorrect negative predictions
5	Error Rate	Total number of incorrect predictions $ER = (FP + FN) / P + N$
6	Accuracy	Total number of correct predictions $Accuracy = (TP + TN) / P + N$
7	Sensitivity	The number of correct positive predictions (TP) divided by total number of positives (P). $Sensitivity = TP/P$
8	Specificity	The number of correct negative predictions divided by the total number of negatives. $Specificity = TN/N$
9	Precision	The number of correct positive predictions divided by the total number of positive prediction. $Precision = TP / (TP + FP)$
10	F-score	Harmonic mean of precision and recall. F1-score = (2 x precision x recall) / (precision + recall)

Table 2 :Parameters definition

Confusion matrix, classification report for machine learning models

Algorithms	**Precision**		**Recall**		**F1-score**		**Support**	
	0	**1**	**0**	**1**	**0**	**1**	**0**	**1**
Logistic Regression	0.75	0.97	0.68	0.97	0.71	0.97	66	**596**
Decision Tree	0.74	0.98	0.79	0.97	0.76	0.97	66	**596**
Random Forest	0.75	0.97	0.76	0.97	0.75	0.97	66	**596**
Gradient Boosting	0.86	0.97	0.76	0.99	0.81	0.98	66	**596**
XGBoost	0.80	0.98	0.80	0.98	0.80	0.98	66	**596**
SVM	0.94	0.97	0.74	0.99	0.83	0.98	66	**596**
Navie bayes	0.58	0.96	0.61	0.95	0.59	0.95	66	**596**
Linear Regression	0.76	0.95	0.53	0.98	0.62	0.97	66	**596**
ANN	0.86	0.98	0.82	0.98	0.84	0.98	66	**596**
LSTM	0.89	0.97	0.74	0.99	0.81	0.98	66	**596**

Table 3 : confusion matrix and classification report for machine learning models

ROC curve

Receiver Operating Characteristic (ROC) curve along with the Area Under the curve (AUC) values used to evaluate the performance of classification models. Figure 13 shows four individual ROC curves with varying AU curves. In the AUC 100% is a perfect classifier that indicating the model distinguishes perfectly between classes, AUC 50% is a random classifier as the curve is a diagonal line meaning the model makes random predictions. And AUC 90% is a high-performing classifier with the ability to distinguish between classes, when shows the AUC 65% has a moderately performing classifier, showing better than random but still limited performance.

Based on the Figure 13, Figure 14 shows that comparing all machine learning models ROC curve based on their AUC curves. Table 4 shows that approximately values of each model.

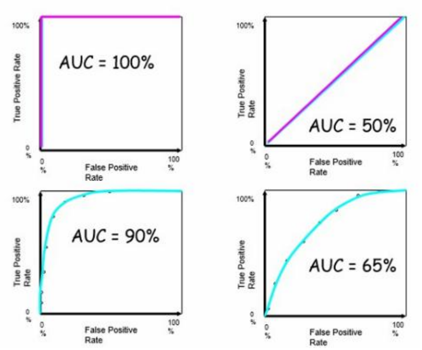


Figure 13: AUC under ROC curve

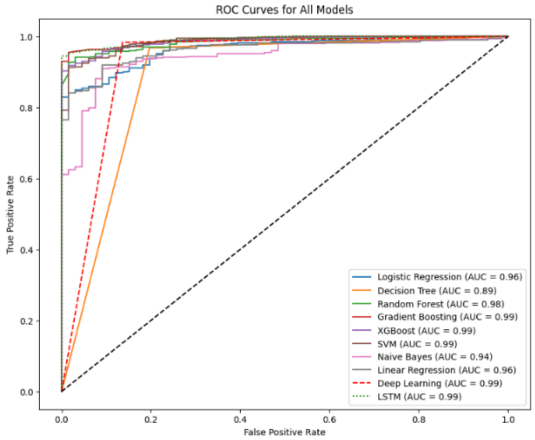


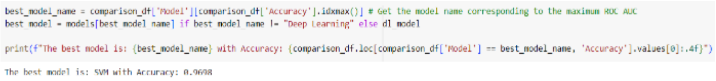
Figure 14: ROC curve for all machine learning models

Table 4 : AUC values for machine learning models

Models	**AUC values**
Logistic Regression	**0.9633**
Decision Tree	**0.8787**
Random Forest	**0.9811**
Gradient Boosting	**0.9899**
XGBoost	**0.9862**
SVM	**0.9857**
Navie Bayes	**0.9435**
Linear Regression	**0.9617**
ANN	**0.9938**
LSTM	**0.9932**

SVM Model

With the help of the machine learning algorithm to find out the best model for predicting the lung cancer. Figure 15 shows the SVM is getting the accuracy value **0.9698** rather than other models.



```
best_model_name = comparison_df['Model'][comparison_df['Accuracy'].idxmax()] # Get the model name corresponding to the maximum ROC AUC
best_model = models[best_model_name] if best_model_name != "Deep Learning" else dl_model

print(f"The best model is: {best_model_name} with Accuracy: {comparison_df.loc[comparison_df['Model'] == best_model_name, 'Accuracy'].values[0]:.4f}")
```

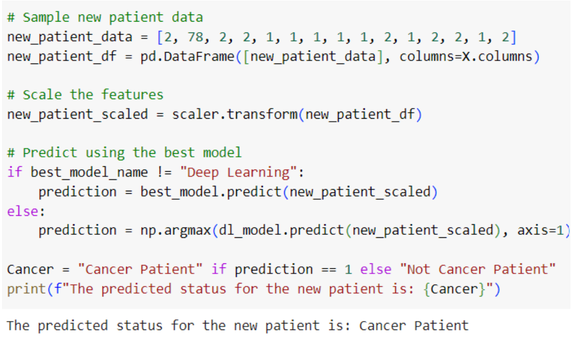
```
The best model is: SVM with Accuracy: 0.9698
```

Figure 15: Best model

Prediction

The SVM model is high accuracy, now predict the lung cancer patient whether the patient is lung cancer or not with help of machine learning concept. Here, using the best model SVM to predict the lung cancer patient.

Below the Figure 16 shows the prediction of new patient outcome with help of the attributes.



```
# Sample new patient data
new_patient_data = [2, 78, 2, 2, 1, 1, 1, 1, 1, 2, 1, 2, 2, 1, 2]
new_patient_df = pd.DataFrame([new_patient_data], columns=X.columns)

# Scale the features
new_patient_scaled = scaler.transform(new_patient_df)

# Predict using the best model
if best_model_name != "Deep Learning":
    prediction = best_model.predict(new_patient_scaled)
else:
    prediction = np.argmax(dl_model.predict(new_patient_scaled), axis=1)

Cancer = "Cancer Patient" if prediction == 1 else "Not Cancer Patient"
print(f"The predicted status for the new patient is: {Cancer}")
```

```
The predicted status for the new patient is: Cancer Patient
```

Figure 16: Final outcome

1. ***RQ1: Which algorithms of machine learning work best for predicting lung cancer?***

   Some works give Logistic Regression, Decision Tree, Random Forest, GBDT, XGBoost, SVM, Navie Bayes, Linear Regression Artificial Neural Network and LSTM as some piece of work on the prediction of lung cancer. SVM is particularly highlighted for its accuracy of correctly identifying cancerous or non cancerous cases, thereby making it a perfect model in healthcare predictive modeling. ANN, Random Forest, and XGBoost also show a reliable performance level both in this research and prior studies, while SVM produces the best records owing to the stability of the method in dealing with the featured classification problems.

2. ***RQ2: Which data preprocessing steps are important to improve the accuracy when predicting lung cancer?***

   The role of data preprocessing in increasing the efficacy of Machine learning models and results accuracy. Said steps include operations such as converting categorical data into numerical forms or applying normalization to the data. The above techniques assist model to gain deeper insights of patterns into the data hence enhancing outcomes in prediction. While there is emphasis on the necessity of this approach, researchers caution that when working with medical datasets such preprocessing is crucial because the variability of formatting of data can negatively impact the model.

3. ***RQ3: What do we use in measuring the performance of the lung cancer prediction models?***

   While comparing the performance of the models for the prediction of lung cancer, authors usually assess the indexes including accuracy, sensitivity, specificity, and AUC-ROC. These metrics are really crucial because they all give different measures on how well or bad the model performs in discriminating between cancerous and non-cancerous cases. For example, sensitivity and specificity are crucial in medical diagnosis to reduce the rate of false negatives and false positives, which are very risky to the patient as well as such a diagnostic technique.

4. ***RQ4: What is the difference in performance of various models in regard to accuracy, sensitivity, and specificity while predicting lung cancer?***

   Most studies analyze models to determine the best approach for the lung cancer prediction process. SVM always has the highest value of accuracy, sensitivity, as well as specificity. ANN and XGBoost algorithms are also efficient, which can be expected from the further usage of these tools for predictive diagnostics. However, the literature indicates that since SVM possess better classification performance as compared with other models of methods, the latter is recommended in healthcare context because it often surpass the other models in coping with the health care data.

5. ***RQ5: What obstacles do one come across when deploying a particular machine learning models for lung cancer diagnosis?***

   Some difficulties appear when applying machine learning techniques for classification of lung cancer: the necessity of high performances due to large datasets, a risk of overfitting, and therefore the need in large amounts of data, and interpretative problems associated with some algorithms, such as SVM. These challenges can restrict its use in clinical environment where resources are tight and hence accuracy has to be balanced with the process of identifying it, in addition, it has to be transparent to other health care professionals. The above limitations suggest that more investigation into such methods is useful and should continue continuously for example in regards to choosing the most efficient algorithms.

Discussion

The proposed methodology in the current study is based on a dataset consisting of features that capture human habits (such as smoking, and alcohol consumption) and signs/symptoms as risk factors that lung cancer patients usually incur. However, these signs are not necessarily related to lung cancer disease, as we observed from features analysis of Methodology. Unlike other cancers, lung cancer cannot be seen with the naked eye, and its symptoms are often accompanied by other disease symptoms. The most frequent symptoms are allergies, asthma, shortness of breath, and coughing. In this work, we selected to train several classifiers on various risk factors related to such symptoms to be able to correctly identify the class (Lung Cancer or Non-Lung Cancer) of an unknown instance, and thus the associated risk.

V. CONCLUSION AND FUTURE WORK

The lungs are the main organs of respiration. Humans never stop breathing until they die because the lungs supply their blood with oxygen, which is vital for human life. Lung cancer is the leading cause of death from malignancies in both genders. The patient's lifespan is determined by the advanced stage of cancer. The earlier the diagnosis, the longer the life expectancy.

In this research work, we exploit supervised learning and deep learning to develop models for identifying individuals with lung cancer manifestation based on several features–symptoms. Various machine learning models, including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost, SVM, Navie Bayes, Linear Regression, Artificial Neural Network (ANN), Long-Short Term Memory (LSTM) were evaluated in terms of accuracy, precision, recall, F-Measure and AUC. From the experiment results the SVM outperformed the other models with an accuracy, sensitivity, specificity results are approximately 97%, 99%, 74% and an AUC of 98%. Additionally, the proposed models performed with better results in comparison to the models as shown in Figure 11 and Figure 12. When consider the results of SVM Precision, Recall, f1-score, support are shown the Table 3.

In future work, we aim to extend the using other machine learning models like KNN, K-means, Adaboost etc. The evaluation used cross validation method assuming the Hold-out method apart from that using other cross validation method like k-fold validation method for data-splitting method for the models' validation.

REFERENCES

1. Yang, Y., Xu, L., Sun, L., Zhang, P., & Farid, S. S. (2022). Machine learning application in personalised lung cancer recurrence and survivability prediction. *Computational and Structural Biotechnology Journal*, *20*, 1811–1820. https://doi.org/10.1016/j.csbj.2022.03.035.
2. *2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST) : 21st-22nd December 2018 (Friday & Saturday) : Venue: Main Auditorium, IQRA University, Defence View, Shaheed-e-Millat Road Extension, Karachi-75500, Pakistan*. (2018). IEEE.
3. AN EXTENSIVE REVIEW ON LUNG CANCER DETECTION USING MACHINE LEARNING TECHNIQUES. (2020). *Journal of Critical Reviews*, *7*(14). https://doi.org/10.31838/jcr.07.14.68
4. Anita, C. S., Vasukidevi, G., Rajalakshmi, D., Selvi, K., & Ramesh, T. (2022). Lung cancer prediction model using machine learning techniques. *International Journal of Health Sciences*, 12533–12539. https://doi.org/10.53730/ijhs.v6ns2.8306

5. Banerjee, N. (2020). *Prediction Lung Cancer-In Machine Learning Perspective*.
6. Christopherp, T., & Jamera Banup, J. (2016). Study of Classification Algorithm for Lung Cancer Prediction. In *IJISET-International Journal of Innovative Science, Engineering & Technology* (Vol. 3, Issue 2). www.ijiset.com
7. Dritsas, E., & Trigka, M. (2022). Lung Cancer Risk Prediction with Machine Learning Models. *Big Data and Cognitive Computing*, *6*(4). https://doi.org/10.3390/bdcc6040139
8. Ingle, K., Chaskar, U., & Rathod, S. (2021). Lung Cancer Types Prediction Using Machine Learning Approach. *Proceedings of CONECCT 2021: 7th IEEE International Conference on Electronics, Computing and Communication Technologies*. https://doi.org/10.1109/CONECCT52877.2021.9622568
9. Kumar Mohan, & Bhraguram Thayyil. (2023). Machine Learning Techniques for Lung Cancer Risk Prediction using Text Dataset. *International Journal of Data Informatics and Intelligent Computing*, *2*(3), 47–56. https://doi.org/10.59461/ijdiic.v2i3.73
10. Manju, B. R., Athira, V., & Rajendran, A. (2021). Efficient multi-level lung cancer prediction model using support vector machine classifier. *IOP Conference Series: Materials Science and Engineering*, *1012*(1), 012034. https://doi.org/10.1088/1757-899x/1012/1/012034
11. Nemlander, E., Rosenblad, A., Abedi, E., Ekman, S., Hasselström, J., Eriksson, L. E., & Carlsson, A. C. (2022). Lung cancer prediction using machine learning on data from a symptom e-questionnaire for never smokers, formers smokers and current smokers. *PLoS ONE*, *17*(10 October). https://doi.org/10.1371/journal.pone.0276703
12. Nisha Jenipher, V., & Radhika, S. (2020a). A study on early prediction of lung cancer using machine learning techniques. *Proceedings of the 3rd International Conference on Intelligent Sustainable Systems, ICISS 2020*, 911–916. https://doi.org/10.1109/ICISS49785.2020.9316064
13. Nisha Jenipher, V., & Radhika, S. (2020b). A study on early prediction of lung cancer using machine learning techniques. *Proceedings of the 3rd International Conference on Intelligent Sustainable Systems, ICISS 2020*, 911–916. https://doi.org/10.1109/ICISS49785.2020.9316064
14. Podolsky, M. D., Barchuk, A. A., Kuznetcov, V. I., Gusarova, N. F., Gaidukov, V. S., & Tarakanov, S. A. (2016). Evaluation of machine learning algorithm utilization for lung cancer classification based on gene expression levels. *Asian Pacific Journal of Cancer Prevention*, *17*(2), 835–838. https://doi.org/10.7314/APJCP.2016.17.2.835
15. Sivanagireddy, K., Yerram, S., Kowsalya, S. S. N., Sivasankari, S. S., Surendiran, J., & Vidhya, R. G. (2022). Early Lung Cancer Prediction using Correlation and Regression. *2022 1st International Conference on Computer, Power and Communications, ICCPC 2022 - Proceedings*, 24–28. https://doi.org/10.1109/ICCPC55978.2022.10072059