Lung Cancer Prediction Using Machine Learning: A Systematic Review

Vikas*, Dr. Prabhpreet Kaur

Computer Science & Technology, Guru Nanak Dev University, Amritsar, India

*Corresponding author

doi: https://doi.org/10.21467/proceedings.114.3

Abstract

One of the large spread diseases in a human being is Lung Cancer. It remains a threat to society and is the cause of thousands of deaths worldwide. Early detection cause of lung cancer is an understandable perspective to maximize the opportunity of the existence of the patients. This paper is about the observation of lung cancer. Here, Computed Tomography (CT) is used for the observation of lung cancer. Various Algorithms are used to search out lung cancer prediction correctly like K Nearest Neighbor, SVM, Decision Tree, and many more. An Aim of the introduced analysis to design a model that can reduce the likelihood of lung cancer in a patient with maximum accuracy. We began by surveying various machine learning techniques, explaining a concise definition of the most normally used classification techniques for identifying lung cancer. Then, we analyze survey representable research works utilizing learning machine classification methods in this field. Moreover, an elaborated comparison table of surveyed paper is introduced.

Keywords: Lung Cancer, Lung Cancer Prediction, Machine Learning, Machine Learning Classification Techniques.

1 Introduction

Lung cancer is examined by the unmanageable cell in the lungs. Lung Cancer is recognized as the main objective of cancer-associated death. During the current year, A lot of progress has been made in the area of the treatment of cancer. However, it has remained the danger of society and reason for the death of thousands of individuals in about the world. This growth can spread to the surrounding tissue and various organs of the body apart from the lungs by the process of metastasis. The major cause of lung cases is prolonged tobacco smoking with 85% and about 10% to 15% cases occur in people who won't ever smoke. Such cases are usually brought about by revealing inherited components and radon gas, secondhand smoke, or different types of air pollution. It can be seen through chest radiographs and CT scans are used to detect lung cancer. Therefore, initial disclosure of cancer in the lungs is the earliest method for the cure of lung cancer. Hence, machine learning is the process that is used to classify the existence of lung cancer.

2 Machine Learning

Machine learning is a category of Artificial Intelligence that gives the capability directly to study an improvement from experience without clearly programming the system. Machine learning is used to examine certain patterns to provide good learning to machines and to handle data in an extra efficient way. The purpose of Machine learning is to understand some knowledge from data itself. All the algorithms along with their description have been clarified in the forthcoming content of this paper. Machine Learning Techniques: The important Machine learning techniques that can be labeled are as follows:

Proceedings DOI: 10.21467/proceedings.114; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-947843-8-8

2.1 Decision Tree

It is a kind of supervised learning method which is utilized for both classification and regression. Although mostly approved to solve classification issues. A Decision Tree is a Graphical representation to obtain all possible solutions based on the given problem. The decision tree has Two nodes that are the decision node and the leaf node. The decision node is utilized to create conclusions and has various branches while the leaf node is the outcome of those choices which don't have additional branches.

2.2 Naive Bayes

Naive-based is a classification technique that depends on the Bayes theorem. In a basic word, Naive Bayes calculation is a kind of technique that relies on the probabilistic suggestion, depends on the Bayesian principle. It relies on the likelihood of right according to usable information and selects the most extreme likelihood of information from any class. It has the least inaccuracy rate think about the next classifier.

2.3 Support Vector Machine

Supervised Machine learning algorithms include Support Vector Machine. The support vector machine aims to make the best line and decision range which separate the classes for which simply insert newly information points into the accurate range. This accurate range is known as a hyperplane. SVM picks the vertex points/vectors that help to create hyperplanes. Such maximum cases are known as Support Vector, and therefore the method is called a Support Vector Machine. The study of SVM is as follows:

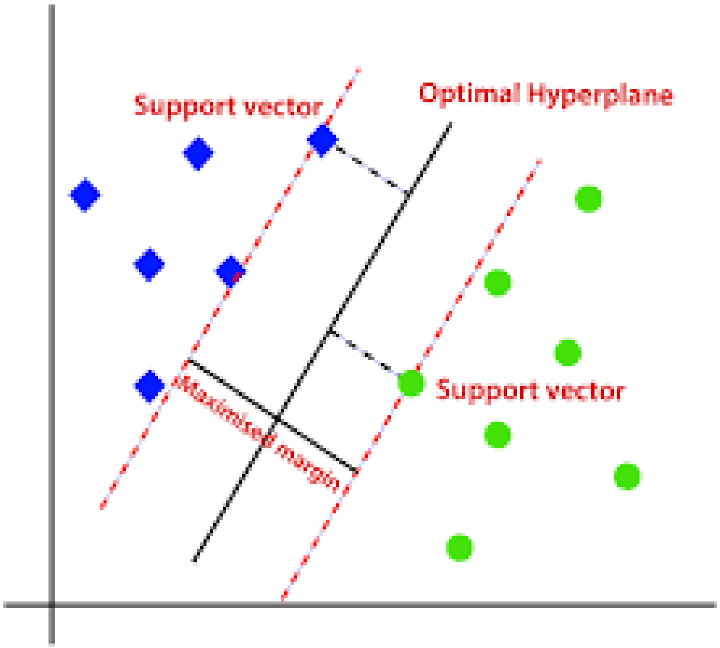


Fig: Support Vector Machine

2.4 K-Mean Clustering

K-mean is the easiest unsupervised machine learning algorithm that calculates popular clustering situations. It defines that the points in the cluster must be equal to each other. So, here, our goal is to limit the distance between the points and a cluster. Is an algorithm that attempts the least distance of points in a cluster with its centroid. The principal goal of the K-means algorithm is to reduce the sum of a path between the points and their particular cluster centroid.

2.5 K-Nearest Neighbor (KNN)

The Algorithm utilizes feature similarity to estimate new data points. In KNN, the training data (which is well-known data) is provided to the learner.

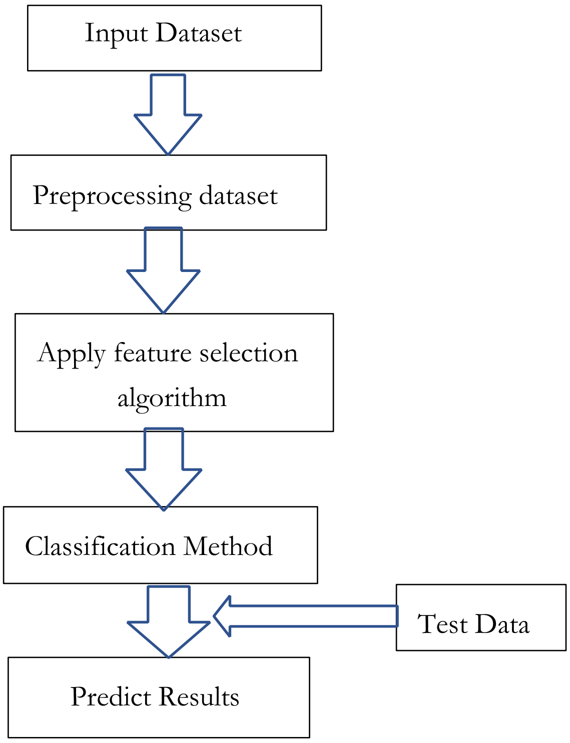
3 Problem Definition

The researchers have neglected the use of the statistical test for feature selection. Most of the researchers have focused on the population-based meta-heuristic algorithm, Genetic algorithm is very complex and computationally costly that is time-consuming.

4 Methodology Flow Chart

For Lung cancer prediction following steps have been used.

4.1 Data Preprocessing

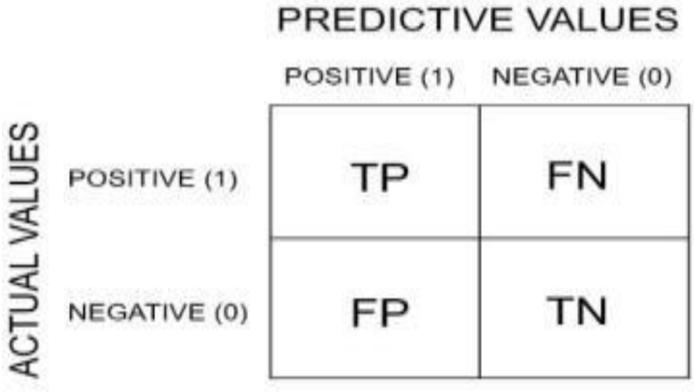


Flow chart of working metrology

In addition to the accuracy of predictive models, the actual overall performance is not only affected by the actual algorithms implemented, but is also solved by the expertise of the data set. In today's world, the data is incomplete, inappropriate, and incorrect, often a shortage of particular values. There is where data comes to the preprocessing scenario which helps to clean, format, and organize raw data so that it is ready for machine learning. Preprocessing phase is very important because it works on the dataset and puts on the idea to the reality that understands the algorithm. In other words, Data processing is a procedure that can be utilized in an effective and efficient format. Data Preprocessing steps include data learning, data transformation, data normalization also other steps regarding the nature of the dataset.

4.2 Performance Evaluation Matrix

Evaluation metrics describe the overall performance of the model. It aims at the predictive ability of a model. The given below are the metrics through which different researchers analyze the prediction models and justify the performance of their outcomes. We have given small definitions for each metrics are as follows:



		PREDICTIVE VALUES	
		POSITIVE (1)	NEGATIVE (0)
ACTUAL VALUES	POSITIVE (1)	TP	FN
	NEGATIVE (0)	FP	TN

: Evaluation matrix

1. **Accuracy**: Accuracy metric means that how many data points are estimated correctly.
   (TP+TN) / TP+TN+FP+FN
2. **Precision:** Precision metric shows the percentage of your results that are appropriate.
   TP / TP+FP
3. **Recall**: Recall mention the percentage of total appropriate results correctly.
   TP /TP+FN
4. **F-Measure:** Combination of precision and recall.
   2TP /2TP+FP+FN

5 Literature Review

D.Jayaraj et al.[1] To predict lung cancer and the dataset was occupied from Lung Image Database Consortium (LIDC) is applied. As a result, the Random forest model gives the highest accuracy with 89.90% and sensitivity with 90.85% also specificity with 88.32%. individually. S. Baskar et al.[2] provides a Support Vector machine model (SVM) in which they detect lung cancer. Therefore, SVM is advised as the highest appropriate algorithms for the diagnosis or detection of cancer. Kymelia Roy et al.[3] recommended making use of Random forest and SVM algorithm to identify the lung cancer disease. By using SVM classification, we get the highest result. Pradeep K R et al. [4] predicted lung cancer by using several algorithms such as Naïve Bayes, SVM, and C4.5. In this paper, various machine learning techniques are compared with the dataset i.e. North Central Cancer Treatment Group had been used in this model to detecting lung cancer. Negar Maleki et al. [5] provided research to help in the detection of lung cancer disease by making use of genetic algorithms. The Dataset taken from the Data world site contains 1000 samples. We also implement 10-fold cross-validations for the training set. As a result, the accuracy attained was 100%. Nikita Bangerjee et al. [6] diagnosed lung cancer prediction by making use of the various machine learning algorithms. In this proposed model, three kinds of machine learning algorithms are used like Random Forest, SVM, and Artificial Neural Network. By comparing the performance of these algorithms, Random forest has achieved the highest accuracy with 96%. Badrul Alam Miah et al. [7] Paper deals with Image processing and Neural Networks. Firstly, Feature extraction is used to select Features that are used to train and test the neural networks. CT scan image is used and achieved the accuracy of 96.67 %. Radhika P R et al. [8] in their proposed paper various outcomes are generated for every algorithm which exists in the lung cancer dataset. The classifier like Naïve Bayes, SVM, decision Tree, and Logistic Regression was executed. SVM achieved the maximum accuracy of 99.2%. Binila Mariyam Boban et al. [9] in this paper, different algorithms are used for lung cancer prediction like MLP (Multiplayer Perceptions), K-Nearest Neighbor, and Support Vector Machine. It comprises 400 CT scan lug disease images. By using various Machine learning algorithms, MLP (Multilayer perceptron) gives the highest accuracy with 98%. Ibrahim M. Nasseret et al. [10] diagnosed lung cancer prediction by using Artificial Intelligence includes Machine Learning that gives the capability directly to study is similar to the Neural Network which provides the best technique that resolves a classification issue and prediction. The dataset was taken from the user sta427 ceyin on the data world website. ANN model can detect existence with 96.67% accuracy of lung cancer. T. Maria Patrica Peeris et al. [11] highly effective optimizing classification techniques Tomography (CT) images have been used to predict lung cancer. approach for the prediction of lung cancer. This research is used to combine KNN and Naïve Bayes algorithm. The Naïve Based algorithm at first uses two hidden layers to extract features from the nearest neighbor. The aim of optimization will allow the model to modify the feature extraction process as the input image given in the network. Given, any motion of the image, the model will be trained for prediction. Syed Saba Raoof et al. [12] used Convolutional Neural Network to produce an accurate. In this paper, Deep learning is used to predict lung cancer. X-ray and Computed Tomography (CT) images are used to detect lung cancer on CT images. Wasudeo Rahane et al. [13] diagnosed lung cancer through SVM and Image Processing. SVM and Image Processing are used to improve the accuracy. Ozge GUNAYDIN et al. [14] used various algorithms for the prediction of lung cancer such as KNN, SVM, Decision tree, Naive Bayes, ANN, and Principal Component study. Decision Tree achieves the best accuracy obtains an accuracy of 93.24%. 247 CT scan images are used in this paper. Mr. Babu Kumar S et al. [15] developed a Convolutional Neural Network Technique to predict Lung cancer. This paper will present a short idea of several techniques used with CNN to diagnose Lung Cancer. The Dataset is taken from LIDC and LUNA. Nada S. El-Askary et al. [16] used the Random Forest to diagnose Lung Cancer. The Random Forest technique provides the best outcomes. CT Images are generally used Dataset in the CAD system. Pranamita Nanda et al. [17] predict lung cancer by using several Methods including SVM, Naïve Bayes, Decision Tree, Random Forest As well as Improved Random Forest from which Improved Random Forest has achieved an accuracy of 98.4%. The Dataset is taken https://www.cancerdata.org with 509 patients. Chethan K S et al. [18] focuses on Convolutional Neural networks (CNN) for the prediction of lung cancer and CT Images are obtained from Kaggle and LUNA websites. At First, Preprocessing and Segmentation are applied to the input image which will divide it, and at last, we train our Data by using the CNN algorithm and improve our accuracy. Emrana Kabir Hashi et al. [19] use the Decision Tree and K Nearest Neighbors Technique For predicting lung cancer. Then, the System evaluates and equates the accuracy of KNN and c4.5.and obtained the highest accuracy of C4.5 with 90.43% for predicting the disease. In this model, PIMA Indians Dataset is used. Qing Wu et al. [20] detect Lung Cancer by using the Entropy Degradation Method (EDM). The author had Predicted by using Computed Tomography and the algorithm obtained an accuracy of 77.8%.

Table 1: Comparison of machine learning classification for lung cancer disease prediction

Reference	Author	Year	Dataset or CT image	Classification technique used	Accuracy achieved
[7]	Badrul Alam Miah	2015	CT image	Neural Network	96.67%
[19]	Emrana Kabir Hashi	2017	PIMA Dataset	KNN, Decision Tree	90.43%
[20]	Qing Wu	2017	CT scan Image	Entropy Degradation Method	77.8%
[8]	Radhika P R	2018	Data world site contains 1000 samples.	Naïve Bayes, Decision tree, SVM, and Logistic Regression	99.2%
[13]	Wasudeo Rahane	2018	200 CT Scan Image	SVM	Not Mentioned
[1]	D.Jayaraj	2019	Lung Image Database consortium (LIDC)Dataset	SVM, KNN, Random Forest, MLP proposed model	89.90%
[10]	Ibrahim M.Nasser	2019	User 427 ceyin on data website	Artificial Neural Network	96.67%
[2]	S. Baskar	2019	CT Image	SVM	90%
[3]	Kymelia Roy	2019	CT Image	Random Forest and SVM	94.5%
[4]	Pradeep K R	2019	North edental Cancer Treatment Group	Naïve Bayes, SVM, C4.5	Not Mentioned
[14]	Ozge Gunaydin	2019	247 CT Scan images	SVM, KNN, Naïve Bayes, Decision Tree, PCA, ANN	93.24%
[16]	Nada S. El-Askary	2019	CT images	Random Forest	Not Mentioned
[6]	Nikita Bangerjee	2020	CT image	SVM, ANN, Random Forest	96%
[5]	Negar Maleki	2020	The data world site contains 1000 samples.	Decision tree, KNN And genetic algorithm.	100%
[9]	Binila Mariyam Boban	2020	CT Scan	KNN,SVM and MLP	98%
[11]	Peeris T.M.P	2020	CT Image	Naïve Bayes, KNN	Not Mentioned
[12]	Syed Saba Raoof	2020	CT Image	Convolutional Neural Network	Not Mentioned
[15]	Mr. Babu Kumar S	2020	LIDC and LUNA Dataset	Convolutional Neural Network	Not Mentioned
[17]	Pranamita Nanda	2020	Cancer data website with 509 patients.	Random Forest, Naïve Bayes, Decision Tree, and Improved Random Forest, SVM	98%
[18]	Chethan K S	2020	CT Scan from Kaggle and LUNA	Convolutional Neural Network	Not Mentioned

6 Gaps in Literature

From the existing literature, it is revealed that existing machine learning models suffer from at least one of the following problems:

1) The Genetic Algorithm suffers from very complexity and hence takes a lot of time in processing with high cost.
2) Maximum existing researchers have ignored feature selection techniques using statistical tests in the time of training and testing. It is observed that by using the feature selection technique, we can increase the performance of machine learning models.
3) The previous researcher states that future works may use the machine learning classification algorithm and feature selection comparing their performances with the previous one.

7 Conclusion

This paper justifies the literature of different Machine learning to predict lung cancer. The accuracy of the introduced model can be different and depends upon the quality of the dataset or CT images used, the tools used by various researchers. It depends upon the model if they use the feature selection technique or not. From the comparison table, we conclude the researcher who produced the highest accuracy was Negar Maleki that use the genetic algorithm with a dataset was taken from the data world site contains 1000 samples. The dataset must be interpreted to obtain good results. Also, we use an appropriate algorithm to develop predictive models. Finally, machine learning is used to diagnose lung cancer disease that helps both health professionals and patients. It is still working for different fields. As viewed from the comparison table, most researchers found a more CT image. There is a need for more high-quality CT images that will be published by various researchers will be a good source for different diseases for their prediction which helps to obtain good results with high accuracy.

References

[1] D.Jayaraj and S.Sathiamoorthy "Random Forest-based Classification Model for Lung Cancer Prediction on Computer Tomography Images." International Conference on Smart Systems and Inventive Technology (ICSSIT 2019) IEEE.

[2] S. Baskar, P. Mohamed Shakeel, K. P. Sridhar and R. Kanimozhi "Classification System for Lung Cancer Nodule Using Machine Learning Technique and CT Images." International Conference and Electronics System (ICCES 2019) IEEE Xplore.

[3] Kyamelia Roy, Sheli Sinha Chaudhury and Madhurima **Burman "**A Comparative of Lung Cancer detection using supervised neural network." (2019) IEEE.

[4] Pradeep K R and Naveen N C "Lung Cancer Survivability Prediction based on Performance Using Classification Techniques of Support Vector Machines, C4.5 and Naive Bayes Algorithms for Healthcare Analytics." International Conference on Computational Intelligence and Data Science (ICCIDS 2018) IEEE Xplore.

[5] Negar Maleki, Yasser Zeinali and Seyed Taghi Akhavan Niaki "A KNN method for lung cancer prognosis with the use of a genetic algorithm for feature selection." (2020) Elsevier Ltd.

[6] Nikita Banerjee and Subhalaxmi Das "Prediction Lung cancer-In Machine Learning Perspective." California Institute of Technology, on July 04, 2020, from IEEE Xplore.

[7] Md. Badrul Alam Miah and Mohammad Abu Yousuf "Detection of Lung Cancer from CT Image Using Image Processing and Neural Network." 2nd International Conf on Electrical Engineering and Information & Communication Technology (ICEEICT) 20 IS Jahangirnagar University, Dhaka1342, Bangladesh, 21-23 May 2015.

[8] Radhika P R, Rakhi. A. S. Nair and Veena G "A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms." (2018) IEEE.

[9] Binila Mariyam Boban and Rajesh Kannan Megalingam "Lung Diseases Classification based on Machine Learning Algorithms and Performance Evaluation." International Conference on Communication and Signal Processing. July 28 - 30, 2020, India (IEEE).

[10] Naseer I.M. and Abu-Naseer "Lung cancer detection using artificial neural network." International Journal of Engineering and Information Systems (IJEAIS). Vol. 3 Issue 3, March – 2019, Pages: 17-23.

[11] Peeris T. M. P. and Brundha "Optimizing Classification Techniques for lung cancer detection on CT images." EPRA International Journal of Multidisciplinary Research (IJMR) Volume: 6, Issue: 3 March 2020.

[12] Syed Saba Raoof, M A.Jabbar, and Syed Aley Fathima. "Lung Cancer Prediction using Machine Learning: A Comprehensive Approach." Second International Conference on Innovative Mechanisms for Industry Applications (ICIMIA 2020) IEEE Xplore, Part Number: CFP20K58-ART; ISBN: 978- 1-7281-4167-1.

[13] Wasudeo Rahan, Himali Devi, Yamini Magar, Anjali Kalane and Satyajeet Jondhale, "Lung Cancer Detection using Image processing and Machine learning Healthcare." International Conference on Current Trends Towards Converging Technologies, Coimbatore, India. 2018 (IEEE).

[14] Ozge Gunaydin, Melike Gunay, and Oznur Sengel, "Comparison of Lung cancer detection algorithm." (2019) IEEE.

[15] Mr. Babu Kumar S and Dr. M VinothKumar. "Detection of Lung Nodules using Convolution Neural Network: A Review." Second International Conference on Inventive Research in Computing Application (ICIRCA-2020) IEEE Xplore.

[16] Nada S. El-Askary and "Lung Nodule Detection and Classification Using Random Forest: A Review." Ninth International Conference on Intelligent Computing and Information systems (ICICIS-2019) IEEE.

[17] Pranamita Nanda and Dr. N. Duraipandian **"** Prediction of Survival Rate from Non-Small Cell Lung Cancer using Improved Random Forest." Fifth International Conference on Inventive Computation Technologies (ICICT-2020) IEEE Xplore.

[18] Chethan K S, Vishwanath S, Rakshith V Patil and Vijetha K A "Segmentation and Prediction from CT Images for detecting Lung Cancer." 11th ICCCNT 2020 July 1-3, 2020 - IIT – Kharagpur (IEEE).

[19] Emrana Kabir Hashi, MD. Shahid Uz Zaman and MD. Rokibul Hasan. "An Expert Clinical Decision Support System to Predict Disease Using Classification Techniques." International Conference on Electrical, Computer, and Communication Engineering (ECCE), February 16-18, 2017, Cox's Bazar, Bangladesh. (IEEE).

[20] Qing Wu and Wenbing Zhao. "Small-Cell Lung Cancer Detection Using a Supervised Machine Learning Algorithm." 2017 International Symposium on Computer Science and Intelligent Controls. (IEEE)