Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/53/20241290

Application of machine learning in lung cancer prediction

Dingjingmu Yang

University of Alberta, 116 St & 85 Ave, Edmonton, AB, Canada T6G 2R3

dingjing@ualberta.ca

Abstract. Lung cancer is a life-threatening disease that is mainly caused by long-term smoking, and genetic reasons. This disease is terribly difficult to treat, but the survival rate can be largely increased by an early diagnosis. However, most people with lung cancer are not detected until the late stage. In recent years, many researchers have developed effective pre-diagnosis methods based on machine learning techniques. Machine learning technique enables the computer to learn from data and perform tasks. This review paper lists machine learning models that can be applied to lung cancer probability prediction. The models are trained by datasets of three types of backgrounds: genetic data, clinical data, and histological data. Each model uses different machine learning algorithms, and all of the models perform excellent ability in predicting. This paper suggests that machine learning models can be applied in screening for lung cancer.

Keywords: machine learning, lung cancer, prediction.

1. Introduction

Lung cancer ranked as the second most common cancer in the world and the five-year survival rate is about 18.6% [1]. 238,000 new cases and 127,000 deaths are estimated for lung cancer in America in 2023 [2]. Smoking, secondhand smoke, radiation, and air pollution are all risk factors for lung cancer.

Most of the time, lung cancer is asymptomatic, making it impossible for a patient to detect it early. Most people with lung cancer are not diagnosed until the late stage [3]. Therefore, a way to identify a potential patient is necessary. Nowadays, Low Dose Computerized Tomography (LDCT) is a widely used technique for detecting lung cancer. It is a medical imaging approach that uses several X-ray photographs of a patient's body obtained from various angles to provide slices of comprehensive information through computer processing. However, the primary problems with LDCT are the significant number of overdiagnoses and false positives [4]. Over-diagnosis refers to incorrectly identifying harmless tumors as cancer. Both over-diagnosis and false-positive may lead to unnecessary treatments that harm patients' bodies. In addition, there are only 2.6 doctors per 1000 people in America, implying that not everyone will have a chance to do LDCT.

According to data from the American Lung Association, in 2021, only 5.8% of the 8 million Americans at high risk of lung cancer received LDCT screening. In this wat, a method to pre-diagnose more people in a more efficient and less costly way is needed. In recent years, many researchers have developed prediagnosis methods based on machine learning algorithms. Machine learning involves training computers to learn from data and perform tasks. To properly build a machine learning model, researchers need to determine the features that contribute to the prediction and acquiring data from the

DOI: 10.54254/2755-2721/53/20241290

real world. Then, the researchers need to select the model that is appropriate to the task. Usually, multiple models are selected so that a clear comparison between models will be presented. Next, the models will learn from the data and automatically abstract the relationships between inputs and outcomes. Once trained, the models' performances are evaluated and the researchers will select the best one among them to conduct further research.

An overview of the machine learning models that can be used for lung cancer prediagnosis is provided in this research. This study focuses on machine learning models that are trained by datasets of three factors of lung cancer: genetic factors; clinical data; and histology. These variables have a strong correlation with lung cancer, therefore the models can make predictions with a high degree of accuracy.

2. Factors related to lung cancer and machine learning tools for prediction

2.1. Key factors of Lung cancer prediction

2.1.1. Gene expression. The gene expressions provide important information for estimating the likelihood of developing lung cancer. The gene determines how a body reacts to the environment, and some individuals are more likely to develop cancer because of genetic factors. Cancer is a genetic disease, and there are a lot of genes discovered as causes of cancer [5]. Also, certain gene leads to gene mutations that cause the growth of cancer cell [6].

2.1.2. Clinical data. Clinical data is the information that is collected while a patient is receiving therapy, and is an important resource for medical research. Typically, clinical data contains data about patients' age, gender, weight, height, blood analysis, and so on. In the study about lung cancer, information on smoking status is also what researchers gather because it is one of the most indispensable features of cancer due to the large number of carcinogens contained in cigarettes [7].

2.1.3. Histology. Histology is the study of the microscopic structure of cells, tissues, and organs. It is an essential tool in predicting a patient's future disease.

2.2. Common Machine Learning Algorithms Used in Prediction

Here are some machine learning algorithms that are applied in later applications.

Multi-layer perceptron is a type of neural network model that has numerous hidden layers. The model is used for recognising complex patterns.

Random subspace is an ensemble method that uses random subsets of features to build different classification models.

The decision tree is a type of tree structure model where each node represents a test, each branch denotes an outcome of the test, and each leaf means an outcome. The model is built by repeatedly splitting the input data into subsets based on several features.

Random forest algorithms is a combination of multiple decision trees, and produce outcome by a voting mechanism.

A support vector machine is a type of model that finds a hyperplane that best separates the inputs into two classes. It is effective even if trained by high-dimensional data. Sequential minimal optimization is an algorithm that optimizes support vector machine.

K-means clustering is a technique that assigns data into different clusters based on similarity. It is efficient for large datasets.

The convolutional neural network is an algorithm for image analysis, which automatically learns features and patterns from the input. Also, this model is the most widely used deep learning algorithm in applications in medical imaging [8].

The linear model is a simple type of model that learns linear relationships between features from training data but can not handle complex non-linear relationships.

DOI: 10.54254/2755-2721/53/20241290

K-nearest neighbour is a classification algorithm that assigns a data point to a cluster based on the majority cluster of the nearest neighbours of the data point.

Extreme gradient boosting is a technique that constructs a strong learner by combining weaker learners. The model is usually constructed by multiple decision trees. In addition, the model can handle missing values which often occur in real-world datasets.

Logic learning machine is a rule-based machine learning approach that uses logic-based representation to learn from the data. It takes the features as input and returns a series of "rules".

3. Application of Models that are trained on different settings

3.1. Gene expression data as the training set

In 2018, Jayadeep Pati developed machine learning models using the micro-array gene data which is a subset of genes to predict the probability of lung cancer. Random subspace, sequential minimal optimization (SMO), and multilayer-perceptron (MLP) are selected as the algorithms to train the models. The outcomes of prediction fall into four categories: positive examples labelled as positive (Ture Positive, TP); negative examples labelled as positive (False positive, FP); negative examples labelled as negative (Ture negative, TN); positive examples labelled as negative (False negative, FN). The models' performance is analyzed using a recall value by the formula: Recall = TP/(TP+FN). As a result, the SMO performs the best with a recall value of 0.9029 which means it correctly identifies 90.29% of positive examples [9].

In 2018, Nicolas Coudray and the team applied a Convolutional Neural Network (CNN) algorithm to predict the gene mutation using images. Analysing gene mutation The dataset is from the NCI Genomic Data Commons. The researchers used Area Under the ROC Curve (AUC) value to assess the performance of the model. In short, if the AUC value is equal to 0.5, the predictions are not more accurate than random guessing, and if the AUC value is equal to 1.0, the model is predicting perfectly. After training and testing, six of ten common mutated genes are concluded to be predictable due to high AUC values with a range between 0.733 and 0.856 [10]. In this way, the model can help pathologists to assess the condition of lung tumours.

In 2017, Alicia Hulbert and the team improved the accuracy of lung cancer diagnosis by detecting DNA methylation in sputum and plasma. Three random forest models were built in this study. All of them are trained on clinical risk factors and methylation data but each had different splits of factors. The model trained using clinical variables has an AUC value of 0.64, which is not good enough. However, the model trained using sputum sample data has an AUC value of 0.85 and the model trained using plasma data has an AUC value of 0.89 [11]. Thus, improving the accuracy of lung cancer diagnosis using machine learning models that is trained by data from sputum and plasma is profitable.

In 2022, Abhishek Choudhary and the team used the K-Nearest Neighbour(KNN) model, a random forest model, a support vector machine (SVM), a linear model (LM), and a model that combines all these four algorithms to analyze the probability of lung cancer. The researcher took five single nucleotide polymorphisms of DNA repair gene XRCC1 and smoking status data as inputs to train the models. The performance of the ensemble-based approach surpassed the four individual models by an AUC value of 0.93 [12]. The high AUC value indicates that the ensemble-based model is powerful for predicting the probability of lung cancer.

In 2019, Masih Sherafatian and Fateme Arjmand applied a decision tree algorithm as a substitute for statistical methods to classify lung cancer status. The model was trained on data from miRNAs involved in cancer and had an AUC value of 91.2% [13]. This shows that it is a practical way to analyze lung cancer status using models trained by miRNA data.

3.2. Clinical data as the training set

In 2020, Michael K. Gould and the team developed an extreme gradient boosting (XGBoost) model for lung cancer prediction. More than 200 thousand data with 834 features including BMI data, smoking information, and spirometric results were collected from Kaiser Permanente Southern

DOI: 10.54254/2755-2721/53/20241290

California. After training and testing, the XGBoost model has an AUC value of 0.856, which shows that this model is conducive to pre-diagnosing lung cancer in the early stage [14].

In 2022, Ruiyuan Yang and the team predict epidermal growth factor receptor (EGFR) mutation in lung cancer via machine learning models. The researchers trained the models with data on some clinical features including blood markers from West China Hospital. Among several algorithms that are selected in this study, the random forest model and XGBoost model outplayed the other models with AUC values of 0.825 and 0.826 respectively [15]. This illustrates that using machine learning models to predict a gene mutation related to lung cancer is a successful way.

3.3. Histology data as the training set

In 2021, Margarita Kirienko and the team applied an automatic segmentation method for image data acquisition and processing using k-means clustering and a threshold-based algorithm. Also, a logic learning machine (LLM) is applied and trained on data from 151 patients to predict histology and tumour recurrence. As a result, the LLM produced 6 rules. Remarkably, the two rules reached an accuracy of 93% and 81% respectively [16], where accuracy value is calculated using Accuracy = (TN+TP)/(FP+TN+TP+FN) formula. Due to the high accuracy of the outcome, the model can be applied to analyze information on histology.

In 2018, Pooya Mobadersany and the team built a genomic survival convolutional neural network (GSCNN) model. This model is a combination of traditional survival models and CNN models. The model can learn visual patterns and molecular biomarkers related to lung cancer from genomic data and histology data. The researchers used Harrell's c index to measure the accuracy of the model. Harrell's c index assesses how well a model can distinguish between pairs of observation, a 1.0 means the model correctly discriminates between pairs of individuals. The c index for GSCNN is 0.801, and the accuracy of the model exceeds the accuracy of human doctors using current clinical standards [17].

4. Conclusion

This review paper lists several machine learning models trained by different backgrounds in recent years. It is extremely challenging to treat lung cancer, but early detection will greatly improve the chances of survival. Machine learning technique provides a new and promising pre-diagnosis method. To predict the probability of lung cancer via machine learning models, the doctors simply need to collect certain biological data from the patient. On average, the time and energy doctors spend on patients will be greatly shortened and saved, while the accuracy of diagnosis remains the same or even higher. The field of machine learning in healthcare is continuously evolving, meaning that more factors that are related to lung cancer prediction will be discovered and applied. Also, multiple models can be used as cross-validation to provide a more accurate prediction. The high accuracy allows the machine learning models to give a persuasive risk prediction of lung cancer, and the low cost brings a chance of screening to a great number of people. In this way, models can be applied as a tool for large-scope screening for lung cancer in the future, and the patient can decide whether to do an LDCT for double screening. On the other hand, it takes a massive amount of data to train a high-quality machine learning model, but people might not feel comfortable contributing their personal medication data due to the awareness of privacy protection. However, people's attitudes towards machine learning models will change as more and more people are helped by the models. In closing, the technology of machine learning is constantly improving, and as a result, its use in the medical industry will expand and become more effective in the future.

References

[1] Lung Cancer Group 2023 Lung Cancer Survival Rate
[2] Lung Cancer Group Research Foundation 2023 Lung Cancer Facts 2023
[3] Latimer KM 2018 Lung Cancer: Clinical Presentation and Diagnosis FP Essent. Jan
[4] Pinsky PF 2014 Assessing the benefits and harms of low-dose computed tomography screening for lung cancer Lung Cancer Manag

DOI: 10.54254/2755-2721/53/20241290

[5] Vogelstein B. and Kinzler K 2004 Cancer genes and the pathways they control Nat Med **10**
[6] Petljak M and Dananberg A 2022 Mechanisms of APOBEC3 mutagenesis in human cancer cells Nature
[7] Hoffmann D and Hecht SS 1990 Advances in tobacco carcinogenesis Handbook of experimental pharmacology
[8] Marie MK and Georgios K 2021 Semi. in Nucl. Med. 51 2 L143-156
[9] Pati J 2019 IEEE. 7 L4232-4238
[10] Coudray N and Ocampo P S *2018* Nat. Med. 24 L1559-1567
[11] Alicia H and Ignacio J 2017 Clin. Cancer. Res. L1998–2005.
[12] Choudhary Abhishek and Anand Adarsh Taylor Francis L0739-1102
[13] Sherafatian M and Sherafatian M 2019 Oncology Letters L2125-2131
[14] Gould MK and Huang BZ et al 2021 Am J Respir Crit Care Med
[15] Yang R and Xiong X 2022 n Lung Cancer Front Oncol
[16] Kirienko M and Sollini M 2021 Eur. J. Nucl. Med. Mo.l Imaging 48 L3643-3655
[17] Mobadersany P and Yousefi S 2018 Proc. of the Nat. Aca. of Sci.