

LUNG CANCER PREDICTION USING MACHINE LEARNING

Natrah binti Ahmad Thohawi, Norpah binti Mahat

Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Perlis Branch, Malaysia

e-mail: natrah683@gmail.com, norpah020@uitm.edu.my

Abstract

Lung cancer continues to be one of the main causes of cancer deaths all over the world, majorly because of the advanced diagnosis (Bertolaccini et al., 2017). The research project seeks to improve early detection by carrying predictive modelling with several different machine learning algorithms. Python was used to analyse the data set consisting of 309 entries of health-related survey questions age, gender, smoking pattern and signs and symptoms like cough, chest pain and shortness of breath. There are five models of classification which are Artificial Neural Network (ANN), Logistic Regression, Decision Tree, Naive Bayes and Gradient Boosting and compared each one. The performance of each model was measured using accuracy, precision, recall, F1 scores, confusion matrix and 10-fold cross-validation. The overall highest performance was achieved by Gradient Boosting (accuracy of 89.32%) and Logistic Regression obtained the most consistent cross-validation score (91.92%). The ANN had flawless recall but a major issue with the false positives. Decision Tree and Naive Bayes have proved also to be reliable, which makes their implementation into clinical measures to be possible. This study proves that machine learning algorithms can prove to be of major help in predicting lung cancer when tuned more to sensitivity or specifically (Raoof et al., 2020). Even though some of the models showed trade-offs (tendency of ANN to over-predict positive cases), combination of several algorithms and the highly directional parameter tuning can lead to better reliability of diagnostics. Predictive modelling is effective in early detection that not only raises the chances of survival but also minimizes the burden on the healthcare systems (Raoof et al., 2020).

Keywords: *Lung Cancer, Machine Learning, Artificial Neural Network, Logistic Regression, Gradient Boosting, Classification Model*

1. Introduction

Lung cancer has become a large issue in the world today and the ability to survive this disease has been greatly influenced by the early diagnoses (Bertolaccini et al., 2017). Nevertheless, the present-day ways of diagnostics analyse the disease when it is already in developed stages, which restricts success of treatment. The study addresses how machine learning methods are applied to enhance the early detection of lung cancer. The main objective is to predict lung cancer using machine learning based on the most important patient characteristics like age, smoking status and such symptoms as coughing or pain in the chest. The paper will also compare how well the five machine learning algorithms, namely Artificial Neural Network (ANN), Logistic Regression, Decision Tree, Naive Bayes and Gradient Boosting, perform. Through these results, it could be evaluated concerning the accuracy, precision, recall and F1 score to determine the most appropriate model to be used in lung cancer prediction. The scale does not exceed a well-organized data with 309 records and is devoted more to the algorithmic research, not to clinical implementation

2. Methodology

This paper predicts lung cancer development in terms of health-related factors through a systematic machine learning procedure. The data which is provided by Kaggle has 309 examples and 15 variables such as age, cigarette use, anxiety, returning cough, chest pain, breathlessness. Raw data were initially pre-processed while the categorical data were encoded as numeric, and the quality of the data with missing values was tested. This dataset was later split into training (70%) and testing (30%) dataset. The Scikit-learn library available in Python implemented five learning algorithms, namely, Artificial Neural Network (ANN), Logistic Regression, Decision Tree, Naive Bayes, and Gradient

Boosting. Four metrics (accuracy, precision, recall and F1 score) were used to evaluate the model. To understand which classifications were incorrect and correct, each model was provided with a confusion matrix. To make the results more dependable, 10-fold cross-validation was used where an average performance was computed over several partitions. This is a systematic procedure that allowed a fair comparison regarding the effectiveness of the models to predict lung cancer and determine those that had the best sensitivity and specificity combination to be used practically in assisting early detection.

3. Results / Findings

This study used a machine learning algorithm to for predict lung cancer using five machine learning algorithms on a data set of 309 patient records. The accuracy, precision, recall, F1 score, and confusion matrix comprehensions were used to assess each model. Gradient Boosting scored the best with accuracy of 89.32 percent followed by Logistic Regression with 88.35 percent. The Artificial Neural Network (ANN) had a hundred percent (100%) recalls, which meet all the positive cases of lung cancer. However, it had shown a large amount of false positive results, which means that it is an overprediction. The performance levels between Naive Bayes and Decision Tree were about equal and models performed approximately at 86.41 percent accuracy. The model stability was also confirmed using 10-fold cross-validation, with Logistic Regression-based model having an average score of 91.92%, which confirmed its consistency in splitting data. These results show that ANN is very sensitive, yet the Logistic Regression and Gradient Boosting have overall performance superior to misclassifications. This makes them better applicable in practical application in early detection of lung cancer where there is a call to have accuracy and reliability.

4. Novelty of Research

This research has the potential to be novel because the machine learning models have been used in predict lung cancer using only basic and symptomatic data without relying on expensive tests or scans. The model features, including age, smoking status, chest discomfort, and shortness of breath, allow it to be used at the early screening stage, particularly in less resourceful environments. It is unique in the sense that all five machine learning algorithms, ANN, Logistic Regression, Decision Tree, Naive Bayes, and gradient boosting are compared on a single dataset. The study provides reliable findings due to the use of 10-fold cross-validation that also indicates the possibility of creating effective, broadly available, and scalable diagnostic tools.

5. Conclusion or Future Work

This research shows that machine learning could be successful in predicting lung cancer based on simple health data. In the future, the model could be enhanced through larger and more varied data sets and possibly be transformed into a user-friendly system of early screening in clinics particularly in low resource regions.

REFERENCE LIST

- Bertolaccini, L., Solli, P., Pardolesi, A., & Pasini, A. (2017). An overview of the use of artificial neural networks in lung cancer research. *Journal of Thoracic Disease*, 9(4), 924–931.
<https://doi.org/10.21037/jtd.2017.03.157>
- Raoof, S. S., Jabbar, M. A., & Fathima, S. A. (2020). Lung cancer prediction using machine learning: A comprehensive approach. In *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)* (pp. 108–115). IEEE.
<https://doi.org/10.1109/ICIMIA48430.2020.9074947>