

MLP-Based Lung Cancer Prediction and Feature Importance Evaluation

Zonglin Jiang^a

Computer Science, Arizona State University, Tempe, U.S.A.

Keywords: Machine Learning, Lung Cancer Prediction, MLP.

Abstract: Lung cancer remains one of the deadliest cancers globally, with high mortality rates due to the challenges of early detection. Traditional diagnostic methods, such as CT scans and biopsies, have limitations, including the risk of human error and patient discomfort. With the advent of machine learning (ML) technologies, early detection has improved significantly. This paper investigates the importance of features in lung cancer prediction using a Random Forest model and a Multilayer Perceptron (MLP) model. The dataset used consists of 309 clinical samples and 15 features, with binary classification into cancerous and non-cancerous cases. After data preprocessing, the models were trained and evaluated to assess the contribution of different features. Age, Allergy, and Swallowing Difficulty were found to be the most important features in both models. The study highlights the impact of dataset imbalance on feature importance and model performance. Future work will focus on addressing this imbalance to improve prediction accuracy and reliability in clinical applications.

1 INTRODUCTION


Cancer is well known for how painful and challenging it is to cure. Lung cancer is the cancer that has the highest fatality rate. According to the World Health Organization (WHO), in 2020, lung cancer caused 1.8 million deaths. At the same time, it is not only hard to cure but also hard to find (World Health Organization: WHO & World Health Organization: WHO, 2023). The early stage of lung cancer is mostly asymptomatic; by the time there are symptoms, likely, the cancer has already progressed to later stages with limited treatment options. Therefore, screening high-risk individuals and achieving early detection is important for more treatment options and higher survival rates.

However, traditional methods of diagnosing lung cancer, like chest X-rays, Computed Tomography (CT) scans, or biopsies have several limitations. CT scans rely on the professionalism of radiology doctors; this introduces human mistakes and subjective decision possibilities to the test. However, those methods can recognize visible abnormal situations. But when the cancer is too small, it is hard for doctors to tell if it is cancer or inflammation and whether it is cancer or something else before tissue

diagnosis with only a few exceptions (Connolly et al., 2003). At the same time, biopsy will increase the discomfort of the patient and the risk of infection. Therefore, a method to eliminate human mistakes and increase accuracy is required.

In recent years, the rise of artificial intelligence (AI) and machine learning (ML) has provided a new opportunity to improve the diagnostic method (Kononenko, 2001; Erickson, 2017; Giger, 2018). Those technologies use huge datasets to automatically analyses medical images and clinical data and recognize possible patterns and abnormalities that may represent cancer. ML models, like Neural Networks, support vector machines, and random forests, are used for the recognition of lung cancer and have made huge progress in improving accuracy and comprehensibility (Pacurari et al., 2023). Take convolutional neural networks (CNNs) as an example; they have achieved significant improvements in early detection and diagnosis accuracy (Javed et al., 2024). It is possible that those models can detect early-stage lung cancer with patterns that are hard for humans to understand instantly; this provides a way for early intervention.

However, one of the main challenges in developing efficient ML lung cancer detection is

^a <https://orcid.org/0009-0009-0296-7614>

choosing relative features from complicated and high-dimension datasets. Selecting features is vital for improving the performance of the model because unrelated or redundant features will introduce noise, which will cause overfitting and lower the accuracy of the model. Also, different types of data, for example, imaging data, clinical records, and genome information lead to challenges in integration and analysis. A successful model not only needs to recognize cancer accurately but also needs to spread among different patient groups and clinical environments. This research discussed the importance of different lung cancer dataset features for Multilayer Perceptron (MLP). This paper has designed a Random Forest model and an MLP model for lung prediction based on clinical datasets. The evaluation standard is based on the importance of features based on their contribution to the accuracy of the diagnosis.

2 METHODS

2.1 Dataset Preparation

The data used in this study to evaluate the feature importance is Lung Cancer (Aswad, 2022). There are 309 specimens and 15 features. The classification task of those data is binary classification, data has been split into 2 categories which are those who have lung cancer and those who do not have lung cancer. This paper pre-processed the data by changing the 1 and 2 of the features into 0 and 1 which is easier to understand for others, then using one hot encode to code the Gender feature from M and F to 2 features, which are Gender_F and Gender_M this way 1 and 0 can be used to represent female and male, since this dataset don't have any missing values or missing features. Therefore, this paper didn't have any preprocess measure for those situations. After those processes, using the train_test_split method to split the dataset into train 80% train set and 20% testing set.

2.2 Random Forest

The first model used in this paper to assess the importance of features is a random forest. Random forest is a frequently used machine learning algorithm, mainly for classification and regression tasks. As an ensemble learning method, random forest builds multiple decision trees during the training process and obtains the final prediction result by voting or averaging the results of these decision trees.

The main reason for choosing random forest as the feature importance evaluation model is that it can measure the feature importance by calculating the contribution of each feature to the reduction of impurity. Each tree in a random forest chooses the feature that minimizes impurity when splitting nodes, which allows the model to automatically evaluate which features are most important in the prediction task. In terms of the selection of model hyperparameters, most of the hyperparameters in this paper are determined by using random search cross-validation. RandomizedSearchCV is verified by randomly selecting several combinations in the hyperparameter space to find the optimal hyperparameter configuration. For $n_{\text{estimators}}$, however, this article takes a loop approach by adding 10 values at a time from values between 1 and 200 and loop-training the model to determine the optimal number of trees. Through this method, the paper finds the best hyperparameter configuration suitable for the data set while ensuring the performance of the model, so that the importance of features can be more accurately evaluated.

2.3 MLP Model

The other model this study used is the MLP model (Taud, 2018; Pinkus, 1999), MLP is a Neural Network model that is widely used in classification and regression tasks. The main idea is to process the input data layer by layer through multiple hidden layers of neurons, and finally output the predicted result. The structure of an MLP consists of an input layer, one or more hidden layers, and an output layer, with neurons in each layer undergoing nonlinear transformations via activation functions to capture complex patterns in the data. In this paper, the proposed MLP model architecture consists of two hidden layers, each containing 64 and 32 neurons respectively. To avoid the model training time being too long, the maximum number of iterations is set at 1500. The activation function uses the Rectified Linear Unit (ReLU), which can effectively mitigate the gradient disappearance problem in deep networks. In terms of optimizer, this paper explores different learning rates, regularization parameters, and optimization algorithms when using random search cross-validation to optimize hyperparameters, and finally determines the optimal configuration of hyperparameters. Using RandomizedSearchCV, this paper finds the best model suitable for the dataset and task in a search space containing 100 different parameter combinations. This method ensures the stability of the model in generalization performance

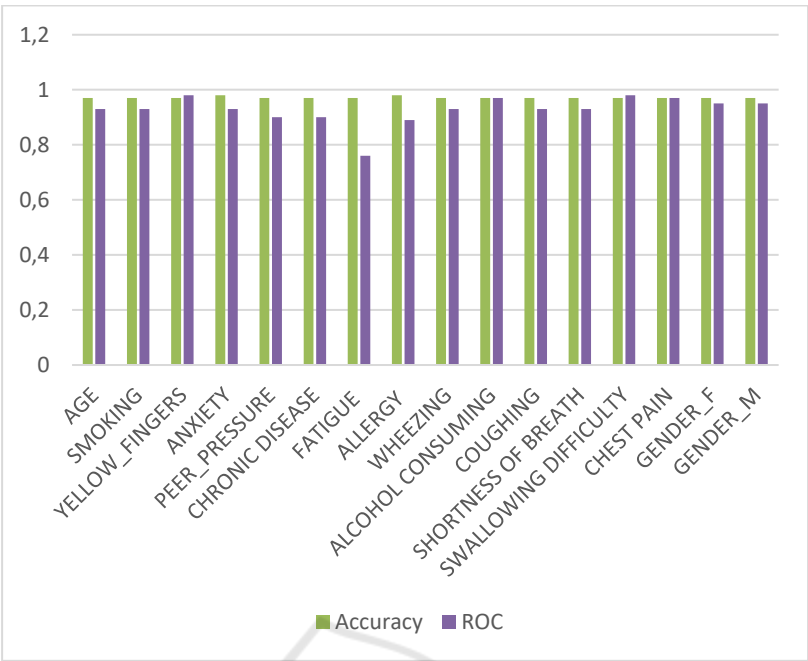


Figure 1: Accuracy and ROC of MLP model after removing the corresponding features (Photo/Picture credit: Original).

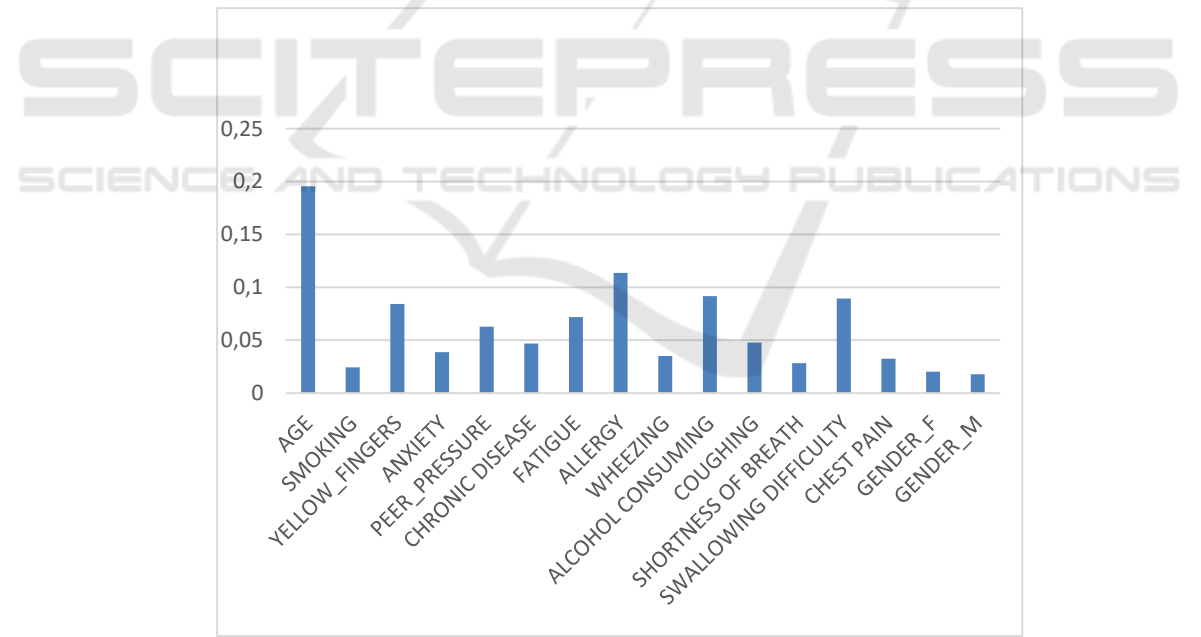


Figure 2: Features Importance in Random Forest (Photo/Picture credit: Original).

and improves the model's prediction accuracy. Finally, the MLP model constructed in this paper performs well on the test set, which proves its effectiveness in complex classification tasks. The classification report, accuracy, and Receiver Operating Characteristic (ROC) curve of the model

further verified the reliability and prediction ability. This study will then put the feature into the MLP model based on the feature importance in the Random Forest model, then this study calculates the feature importance to the MLP model by comparing the contribution to the accuracy.

3 RESULTS AND DISCUSSION

This study evaluated the importance of features in a lung cancer dataset using Random Forest and Multilayer Perceptron (MLP) models. The dataset comprises 309 samples and 15 features. These features were selected to identify better factors related to lung cancer and improve the model's predictive accuracy. The results are provided in Figure 1 and Figure 2.

3.1 Feature Importance Evaluation

Using the Random Forest model, this study calculated the contribution of each feature to the model's performance. The results of feature importance are shown in Figure 1. Age (AGE) was identified as the most significant feature with an importance score of 0.1955, highlighting its crucial role in lung cancer prediction. The next most important features were Allergy and Swallowing Difficulty, with importance scores of 0.1136 and 0.0893, respectively, indicating a strong correlation with lung cancer.

In the MLP model, this study assessed the contribution of features to the model's accuracy by using the feature importance from the Random Forest model. Detailed data on feature importance in the MLP model are presented in Figure 2. This study observed that removing the Age feature decreased model accuracy to 0.94 (ROC of 0.90) while removing the Allergy feature decreased accuracy to 0.96 (ROC of 0.93).

In summary, both Random Forest and MLP models highlight the importance of features such as Age, Allergy, and Swallowing Difficulty in lung cancer prediction. The identification and weighting of these features are crucial for enhancing early detection accuracy. These findings help optimize model performance in practical applications and provide a foundation for further research and feature selection strategies.

The results of this study indicate that both the Random Forest and MLP models' most important features are Age, Allergy, and Swallowing Difficulty in lung cancer prediction. However, it is important to consider the impact of dataset imbalance on the evaluation of feature importance.

Additionally, features with lower importance scores, such as Fatigue and Wheezing, might still have clinical significance in specific patient groups or disease stages. Therefore, it is important for future research to address dataset imbalance through techniques like over sampling, and under sampling. These methods can provide a more

balanced view of feature importance and improve the model's overall performance

4 CONCLUSIONS

This study demonstrates the features and importance of lung cancer prediction in the MLP model using a random forest model. Features such as Age, Allergy, and Swallowing Difficulty are important for the diagnostic process. However, the significant imbalance in this dataset, with only 39 non-cancerous samples, may impact the accuracy of feature importance evaluations. Future research should focus on using dataset imbalance with advanced sampling techniques and validating findings with larger, more balanced datasets. Overcoming these challenges will enhance the accuracy of predictive models, improve early detection strategies for lung cancer, and benefit patient outcomes.

REFERENCES

- Connolly, J. L., et al. 2003. Role of the surgical pathologist in the diagnosis and management of the cancer patient. Holland-Frei Cancer Medicine - NCBI Bookshelf.
- Erickson, B. J., Korfiatis, P., Akkus, Z., & Kline, T. L. 2017. Machine learning for medical imaging. *Radiographics*, 37(2), 505-515.
- Giger, M. L. 2018. Machine learning in medical imaging. *Journal of the American College of Radiology*, 15(3), 512-520.
- Javed, R., Abbas, T., Khan, A. H., Daud, A., Bukhari, A., & Alharbey, R. 2024. Deep learning for lungs cancer detection: A review. *Artificial Intelligence Review*, 57(8).
- Kononenko, I. 2001. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, 23(1), 89-109.
- Lung cancer. 2022. Kaggle. <https://www.kaggle.com/datasets/nancyalaswad90/lung-cancer/data>
- Pacurari, A. C., et al. 2023. Diagnostic Accuracy of Machine Learning AI architectures in detection and Classification of lung Cancer: A Systematic review. *Diagnostics*, 13(13), 2145.
- Pinkus, A. 1999. Approximation theory of the MLP model in neural networks. *Acta numerica*, 8, 143-195.
- Taud, H., & Mas, J. F. 2018. Multilayer perceptron (MLP). *Geomatic approaches for modeling land change scenarios*, 451-455.
- World Health Organization: WHO & World Health Organization: WHO. 2023. Lung cancer. Retrieved September 4, 2024, from <https://www.who.int/news-room/fact-sheets/detail/lung-cancer>