Lung Cancer Feature Analysis and Classification Prediction Based on Machine Learning and Deep Learning

Xin You

Zhuoyue Honors College, Hangzhou Dianzi University, Hangzhou, Zhejiang, 310000, China
21051606@hdu.edu.cn

Abstract. In recent years, lung cancer has the highest number of confirmed cases and a high mortality rate among all types of cancer, thus, it is vital to make timely and accurate lung cancer predictions. To alleviate this situation, this paper experimented with lung cancer classification prediction based on historical data using machine learning (ML) and deep learning (DL) methods. The study uses correlation analysis and F-test to perform feature selection and feature merging on the dataset. Then, K-nearest Neighbour (KNN) and eXtreme Gradient Boosting (XGBoost) in ML and Convolutional Neural Networks (CNN) in DL are applied for the prediction, classifying people into three risk levels of being diagnosed with lung cancer. The result of this experiment shows that KNN performs the best in terms of runtime and XGBoost has the best interpretability. Also, features like "obesity", "fatigue,"coughing of blood," and "air pollution" play a significant role in lung cancer classification. In contrast, others, including "age" and "gender" have little impact on the classification. This paper provides a possibility for screening potential patients with lung cancer, to some extent, alleviating the situation of delayed diagnosis of lung cancer due to limitations in existing medical technology.

Keywords: Classification prediction, Machine learning, Deep learning, Lung cancer

1 Introduction

Having nearly 2.5 million new cases and 1.8 million deaths worldwide in 2022, lung cancer has been the most significant component of cancer diagnosis and cancer deaths [1]. One of the reasons why lung cancer has high deaths is that due to the limitations of current medical technology, there often remains a delay in lung cancer diagnosis [2]. Due to these delays, patients diagnosed at advanced stages may have limited treatment options [2]. Therefore, how to predict lung cancer on a specific case based on the analysis of historical data is essential. If potential and high-risk lung cancer patients can be predicted, more medical methods can be used and the survival rates of lung cancer will be largely improved.

According to previous research, lung cancer is the result of a combination of multiple reasons [3]. People diagnosed with lung cancer often share some similarities,

Y. Wang (ed.), *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, Advances in Computer Science Research 115,
https://doi.org/10.2991/978-94-6463-540-9_8

such as having a smoking history, having a family history of lung cancer and having an overeating diet [4]. Thus, people can predict the likelihood of developing lung cancer based on the statistical data of lung cancer patients and methods in the fields of machine learning (ML) and deep learning (DL) can be applied to solve this problem. In early research, ML methods are successfully applied in cancer prediction, giving accurate decision-making results [5]. Classification models in ML, such as K-nearest Neighbour (KNN) and Support Vector Machine (SVM) have been used to predict whether a person has developed lung cancer, and deep learning models like Convolutional Neural Networks (CNN) and Radial Basis Function (RBF) have also been applied to this prediction [5-7]. These studies, to some extent, have improved the efficiency of lung cancer diagnosis, and therefore, the problem of delayed diagnosis of lung cancer has been alleviated. However, these studies primarily focus on comparing the accuracy of different models in classifying lung cancer development, neglecting in-depth analysis and visualization of patient characteristics. Besides, due to the limitation of the datasets that these previous researchers choose, these researchers cannot predict the patients' level of risk of being diagnosed as lung cancer.

Therefore, this study adds feature analysis of individuals who may be diagnosed with cancer at different risk levels and conducts more visual analysis to identify risk factors for lung cancer. In this paper, several different models, containing KNN, eXtreme Gradient Boosting (XGBoost) and CNN, are applied to do the prediction and classify people into three risk levels of being diagnosed with lung cancer. Feature engineering for reduce the runtime and improve the prediction accuracy will be applied. The runtime and accuracy of these models are compared. During the experiment, visualization methods are used to display the characteristics of people with different risk levels. The impact of each feature in modeling will also be presented visually to show the risk factors that may indicate developing lung cancer. Ultimately, this research can achieve classification prediction of hidden cancer risks and feature analysis for lung cancer and thus provide rational advice for lung cancer prevention and diagnosis.

2 Method

2.1 Data

Datasets. The dataset used in this study is first released in the Kaggle website under the "cancer" category. This dataset is related to lung cancer, containing 1000 pieces of sample and 23 different features. Compared to other similar datasets, this dataset contains a wide range of features related to lung cancer and categorizes the risk of people being diagnosed with lung cancer into three levels. Thus, it provides sufficient data for training and improves the statistical significance of the research results. Besides, the features that affected lung cancer and are contained in this dataset, including age, gender, and dust allergy, are of great analytical significance for improving the accuracy of lung cancer prediction.

Data processing. The raw data obtained directly from Kaggle is relatively clean, already evaluating patients' intensity of every feature and ranking the level. Therefore, during the data processing stage, only checking out and deleting items that contain missing data and recoding people's risk levels in lung cancer patients are needed. Due to the need for model input, the label value for "low" risk classes is set into 0, the label value for "medium" risk classes to 1, and the label value for "high" risk classes is set into 2.

2.2 Model

The process of this experiment includes data processing, feature selection, feature merging, modeling and evaluating, as shown in Fig 1. The following content will explain each part.

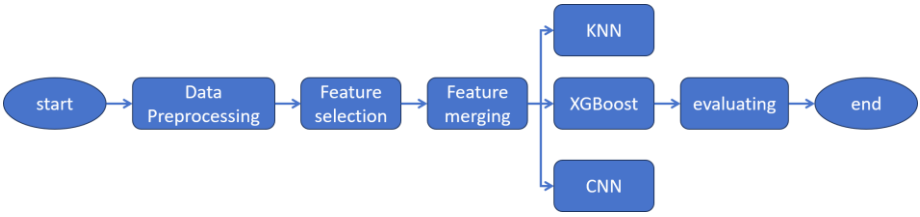

Fig. 1. Method structure diagram

Feature engineering. Since the dataset this study uses is multidimensional, filtering out and deleting the features with low relevance to labels will help reduce runtime while maintaining accuracy [8-10]. Spearman correlation coefficients are used to evaluate the correlation between various features and labels. Besides, compared to the Pearson correlation, the Spearman correlation is more suitable for the dataset in this paper, which belongs to ordinal data [11]. Using Spearman correlation coefficients,correlation values between each feature and label are got, ranging from -1 to 1 and describing the strength of the relationship by its absolute values. Any feature with a Spearman coefficient value associated with the label less than the threshold(0.4 for this dataset) is directly deleted.

Given that the dataset may have similar features, the F-test method in one-way Analysis of Variance (ANOVA) is used to determine whether there is a significant difference between every two features. F-test is a statistical test used to compare the variances of two samples. Whether or not two features need to be merged depends on whether the P-value returned by the F-test is greater than the threshold(0.01 for this dataset). The larger the P-value, the smaller the between-group difference. To ensure the number of features, features without significant differences will be merged until they are less than 20% of the total number of features.

As the values of the features are ordinal data, recoding is needed during feature merging. The value of new feature generated by feature merging in show as Eq.(1). refers to the value of the nth feature of the i-th data, the maximum data in parentheses,

and the value of the i-th data's new feature. After obtaining new merged features, delete the original features used for merging, thus achieving dimensionality reduction. The formula is as follows:

$$x_i = \frac{\sum(a_{i1}+a_{i2}+\cdots+a_{in})}{\sum(\max(a_{11},a_{21},\ldots,a_{i1})+\max(a_{12},a_{22},\ldots,a_{i2})+\cdots+\max(a_{1n},a_{2n},\ldots,a_{in}))} \quad (1)$$

Classification method. The models used in this study include KNN, XGBoost and CNN.

The KNN method is a widely used machine learning classification method. The KNN model assigns a label when K objects closest to a test object are found in the training dataset [12]. The KNN method does not have the traditional training stage. And during the prediction, K data points closest to the test data point are first found and then the label of the test data point is determined by the label that appears most frequently among the K data points [12]. The reason for choosing KNN as one of the comparative models is that it can work without any prior data distribution assumption and is simple to understand.

XGBoost is a scalable end-to-end tree-boosting system with high learning efficiency and accuracy [13]. Based on the tree structure, XGBoost has considerable interpretability that has the capability to give feature importance ranking after prediction. In this instance, the feature importance can contribute to feature analysis and provide reference for case diagnosis.

Convolutional Neural Networks have three types of layers: convolutional layers, pooling layers, and fully connected layers. CNN is formed by stacking these layers [14]. Firstly, the convolutional layer calculates the weight and determines the neurons' output [14]. Then, pooling area has the ability of downsampling and reduce subsequent calculations. Finally, the output of the connected layer is the result of the expected classification. This project applies CNN because it is suitable for high-dimensional data and can solve multi-class classification [8].

2.3 Evaluation

Evaluation for every classification method is done from two aspects: runtime and accuracy. In terms of runtime, KNN and XGBoost are run repeatedly 100 times, and the total running time divided by the number of runs will be the final runtime. For CNN, the run time of the least number of epochs will be the final runtime when obtaining the same accuracy. In terms of error analysis, accuracy is used for evaluation, as shown Eq.(2). TP refers to the number of true positives, TN refers to the number of true negative and N refers to the number of total data samples in the dataset.

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{N}} \quad (2)$$

3 Result

3.1 Result of patient characteristic

Table 1 presents some features with a strong correlation with the label, with the highest correlation coefficient of "obesity" being 0.815 and "coughing of blood" being 0.766. While some features that are not presented in the table have weak correlation with the label, with correlation coefficients for "age" and "gender" below 0.2.

Table 1. Features correlation with label

Feature	Correlation coefficient
Dust allergy	0.703
Genetic risk	0.676
Balanced diet	0.692
Obesity	0.815
Passive smoker	0.683
Coughing of blood	0.766
Chest pain	0.655

In the stage of feature independence analysis, some features that share a similar distribution are found. As show in Fig2, 0, 1 and 2 present the three level of risks of being diagnosed as lung cancer: low, medium and high. According to the Figure, the high-risk population always have a high-ranking value on features like occupational hazards, obesity, smoking and etc. In other words, people who are obese or have an addiction of smoking will be more likely to be diagnosed as lung cancer. Meanwhile, the distribution of features also indicates that the high-ranking or low-ranking value of the features shown in the figure does not always appear alone. For example, as shown in the 6th scatter plot on the first row of Fig2, a high-risk person is likely to have occupational severe hazards accompanied by serious fatigue, and a low-risk person in more likely to have very mild occupational hazards accompanied by mild fatigue.

The feature importance given by the XGBoost model based on the feature engineered data is show as Fig3. The "daily merge" feature is obtained by merging "Occupational Hazards"," Balanced Diet", "Obesity", "Smoking", "Passive Smoker" and "Fatigue". The "ill merge" feature is obtained by merging "Genetic Risk", "Coughing of blood", "Shortness of breath" and "frequent cold". According to Fig3, "daily merge" contributes the most for prediction, which suggests the feature used for composing "daily merge" is important when deciding whether a patient is at risk of being diagnosed with lung cancer. Compared to "daily merge", "illness merge" contributes less in classification but is also essential. Meanwhile, other clinical manifestations that are independent of the features covered by "ill merge" such as "chronic lung disease," also play an important role in determining a person's risk of being diagnosed as lung cancer. Finally, objective environmental factors independent of the patient's

symptoms and daily manifestations – "Air Pollution" – are also reference features in classification.

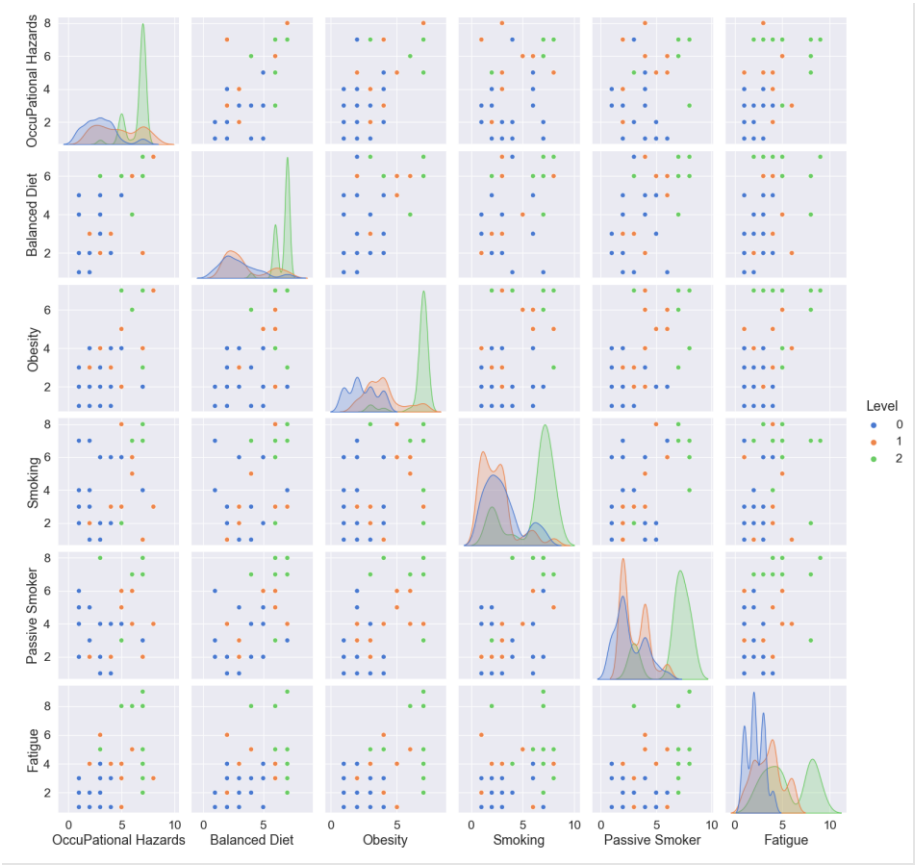

Fig. 2. Distribution of similar features

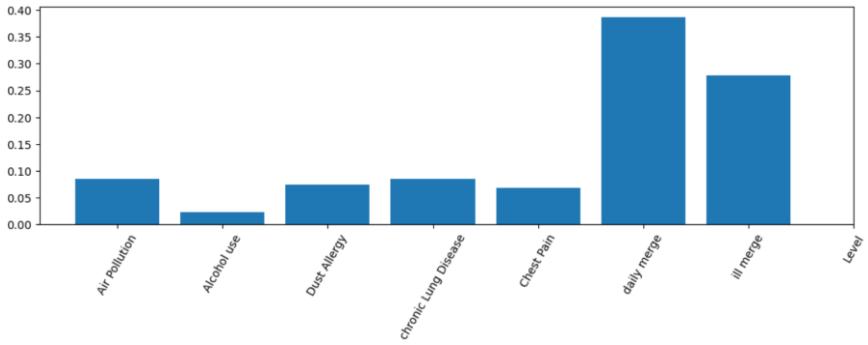

Fig. 3. Feature importance by XGBoost

3.2 Result of classification

As shown in Fig4 and Fig5, the run time and accuracy of three classifications are compared, and the two indicators of classifications using raw data and feature-engineered data are also compared.

For the machine learning methods KNN and XGBoost, feature selection and feature merging can reduce their runtime by 13.5% and 16.8%, respectively, when maintaining the same accuracy. While for the Deep learning method CNN, in order to get the same 100% accuracy, the runtime for the feature-engineered data is 3.2 times more than that of the raw data. Additionally, feature engineering also helps KNN improve its accuracy from 99.5% to 100%, showing the ability to distinguish the data with different labels better. In this experiment, it indicates that feature engineering is of use for some machine learning methods. At the same time, it may not be suitable for models like CNN with automatic feature extraction ability.

KNN, XGBoost, and CNN can all achieve 100% accuracy in classification prediction, while KNN has the shortest runtime – 30.28% of the runtime of XGBoost and 1.33% of the runtime of CNN. Regarding model interpretability, the classification principle for KNN is relatively easy to understand, and XGBoost has the best model interpretability, giving the feature importance after every prediction. In contrast, CNN has weak interpretability and cannot explain the detail for its prediction.

Fig. 4. Classification runtime comparison

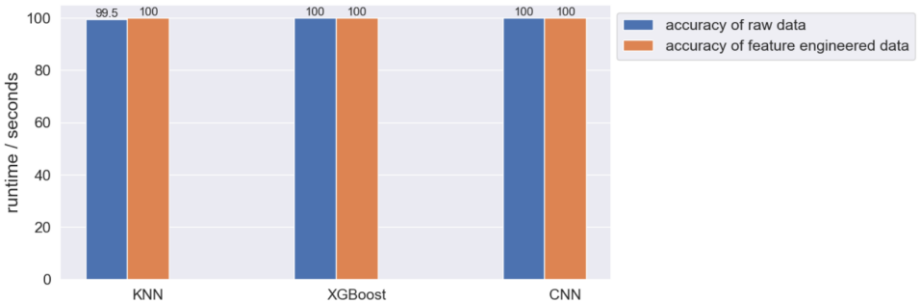

Fig. 5. Classification accuracy comparison

4 Discussion

Due to the simplicity and limited size of the dataset, containing only 1000 samples, the accuracy of each model for the classification prediction is relatively high, even all achieve 100% accuracy after feature engineering. In addition, another factor that leads to high prediction accuracy is that there is no missing data in the final 1000 samples used for prediction, providing rich feature dimensions. Thus, this study is still incomplete in improving model accuracy while focusing more on reducing model runtime. . Similarly, because of the limitations in the dataset, the scalability of feature engineering and classification is poor, not able to add samples with missing items. Based on this projects, models that can deal with datasets with missing items are expected. Also, maintaining efficiency and accuracy when facing a larger dataset is under future consideration.

5 Conclusion

Based on data preprocessing, this paper uses correlation analysis and F-test to perform feature selection and feature merging on the dataset. Then, the runtime and accuracy of three models - KNN, XGBoost, and CNN - are compared using the raw and feature-engineered datasets.

Research has found that all three models can achieve 100% accuracy in this lung cancer classification prediction, while KNN has the shortest runtime - 30.28% of it of XGBoost and 1.33% of it of CNN. Also, through the comparison, it is found that after feature engineering, the accuracy of KNN improves, and the runtime of both KNN and XGBoost decrease to a certain extent, by 13.5% and 16.8%, respectively. However, the runtime of CNN increases by more than two times after feature engineering. From the perspective of feature analysis and the feature importance given by XGBoost, the following conclusions have been drawn. Features like "obesity"," balanced diet", "coughing of blood" and "air pollution" are closely related to the risk of lung cancer and have a significant impact on classifying the risk level of lung cancer, while other features , including "age" and "gender" have little to do with the classification prediction. Therefore, in this study, KNN is the most suitable model for pre-

dicting lung cancer classification, but at the same time, XGBoost has the best interpretability and thus able to provide reference for lung cancer screening and diagnosis. Whereas CNN has the longest runtime and this situation cannot be solved through feature engineering, so it is not suitable for the prediction of this study.

This study contributes to the classification prediction of the likelihood of being diagnosed as lung cancer , which is beneficial for identifying high-risk populations and thus relieving the situation of delayed diagnosis of lung cancer. Meanwhile, this paper provides a features analysis for identifying high-risk lung cancer populations, which can give a reference for clinical diagnosis. However, due to the limitation of the dataset, this study still has shortcomings in improving model accuracy and scalability. Future research will focus on how to make prediction on datasets with missing data and maintain the accuracy of prediction.

References

1. Bray, F., Laversanne, M., Sung, H., et al: Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 1-35 (2024).
2. Nooreldeen, R., Bach, H.: Current and Future Development in Lung Cancer Diagnosis. Int J Mol Sci 22(16), 8661 (2021).
3. Ganti, A.K., Klein, A.B., Cotarla, I., et al: Update of incidence, prevalence, survival, and initial treatment in patients with non-small cell lung cancer in the US. JAMA Oncol 7(12), 1824–1832 (2021).
4. Dong, J.Y., Zhang, L., Zhang, J.: Analysis of influencing factors and construction of prediction model for lung cancer based on random forest algorithm. Chinese Journal of Medicine 58(11), 1188–1193 (2023).
5. Konstantina, K., Themis, P.E., Konstantinos, P.E., Michalis, V.K., Dimitrios, I.F.: Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal 13, 8–17 (2015).
6. Nuhić, J., Kevrić, J.: Lung cancer typology classification based on biochemical markers using machine learning techniques. In: 43rd International Convention on Information, Communication and Electronic Technology , pp. 292–297., Opatija, Croatia (2020).
7. Mustafa , A.D., Mohsin, A.A., Bibo, S.A.: Lung Cancer Prediction and Classification Based on Correlation Selection Method Using Machine Learning Techniques. QAJ 1, 141–149 (2021).
8. Patra, R.: Prediction of Lung Cancer Using Machine Learning Classifier. In: Chaubey, N., Parikh, S., Amin, K. (eds.) Computing Science, Communication and Security 2020, Communications in Computer and Information Science, vol. 1235. Springer, Singapore(2020).
9. Wang, X.: Research on lung cancer risk assessment, diagnosis, and tissue typing system based on data mining technology. Zhengzhou University, PhD dissertation(2019)
10. Li, J.D., Cheng, K.W., Wang, S.H., Fred, M., Robert, P.T., Tang, J.L., Liu, H.: Feature Selection: A Data Perspective. ACM Comput. Surv. 50(6) (2017).
11. Schober, P., Boer, C., Schwarte, L.A.:Correlation Coefficients: Appropriate Use and Interpretation. Anesthesia & Analgesia 126(5), 1763–1768 (2018).
12. Zhang, S..: Challenges in KNN Classification. IEEE Transactions on Knowledge and Data Engineering 34(10), 4663–4675 (2022).

13. Chen, T.Q., Carlos, G.: XGBoost: A Scalable Tree Boosting System. In: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. Association for Computing Machinery, New York, NY, USA (2016).
14. O'shea, K., Nash, R.: An introduction to convolutional neural networks. Computer Science: arXiv preprint, 1–11 (2015).