Prediction for the Extent of Cancer based on Multiple Machine Learning Algorithms

Shujun Chen[1, †, *], Yuchen Wang[1, †], Ziang Yang[2, †] and Zezhou Zhang[1, †]

[1] Department of Computer Science and Technology, Tianjin University of Technology, Tianjin, China

[2] Department of Computer Science and Technology, Sun Yat-sen University, Guangzhou, China

* Corresponding author email: aaronchen4396@stud.tjut.edu.cn

†These authors contributed equally.

Abstract. Machine learning is a very important method for predicting the extent of cancer. In recent years, there has been considerable progress in detecting tiny features of the body, such as the specific surface area of cancer cells, the nuclear volume and perimeter of cancer cells. There is not development in measuring macro symptoms, such as Coughing of Blood, Chest Pain. Because in some remote and backward areas, detailed data are difficult to obtain, macro features need to be considered. In this paper, we use some relevant characteristics and environmental factors to predict the degree of cancer. During data preprocessing, some irrelevant contents are deleted, and the training set and the test set are divided. At the same time, some text data are digitized. Then, Naive Bayes, Decision Tree, Random Forest and Support Vector Machine were used in turn to make predictions and record their respective results. As a result, Naive Bayes, Decision Tree, Random Forest and Support Vector Machine achieved a fairly high prediction rate. More relevant features are obtained through the Person correlation coefficient and Feature Importance. On this basis, the prediction will not only maintain a high prediction rate, but also greatly reduce the memory consumption and training time.

Keywords: Cancer; Machine Learning; Decision Tree; Random Forest; Classification.

1. Introduction

Cancer is one of the most serious diseases in the human history. Cancer cells, also known as malignant tumor cells, often grow in an infiltrative or exophytic manner without capsules, and are often poorly demarcated from surrounding tissues. In the process of tumor occurrence and development, recurrence and metastasis often occur. The metastasis of cancer cells is uncontrollable, and radiotherapy and chemotherapy only try to kill cancer cells as much as possible without preventing them from dividing in large quantities. Until now, there is no effective treatment or medicine to cure the cancer, so it is crucial to know why people may suffer from the cancer and which life factors that may decide the extent of cancer. In order to deal with this problem, machine learning can be an effective way to explore the relationship between life factors and cancers.

In previous studies, the degree of cancer based on the physical tests has been fully studied and predicted. For example, the expression levels of miR-21, miR-31 and miR-155 in the serum and tissues of lung cancer patients are high. The serum microRNAs detection can be used as an indicator for the preliminary diagnosis of lung cancer to assist in judging and predicting the degree of deterioration of lung cancer [1]. For simple breast cancer, studies have found that compared with normal cells, the specific surface area of cancer cells is significantly smaller, and the specific surface area of advanced cancer cells is the smallest. Meanwhile, the nuclear volume and perimeter of cancer cells are significantly increased, and the nuclear volume and perimeter of advanced cancer cells are the largest. It indicates that the larger the volume of cancer cells and their nuclei, the longer the nuclear perimeter of cancer cells, the higher the degree of cancer cell deterioration [2]. However, these relevant variables mentioned above need to be measured by high-precision instruments. Without the help of high-precision instruments, in some backward areas of Africa which have poor medical facilities, it is difficult to obtain relevant data. Even if remote consultation can be conducted, relevant experts will not be able to make accurate judgments or even dare to draw conclusions without specific

data. In response to this phenomenon, this study decided to refer to some relevant variables without the assistance of high-precision instruments to preliminarily diagnose the degree of cancer.

To explore the relationship between the degree of cancer and life factors, we first processed and analyzed the data set to obtain the degree of correlation between the data. A variety of machine learning algorithms were used to explore which factors are more closely related to the degree of cancer and can predict the degree of cancer to a certain extent based on the characteristics of the patient. We obtained the open-source dataset and verified it to be authentic and reliable. The dataset divides the level of cancer into low, medium and high and the numbers represent the degree of features. At present, the academic community pays more attention to the incidence of cancer, while ignoring the research on the severity after the occurrence of cancer, and we selected the data set showing the degree of cancer for analysis, aiming to make up for the lack of this aspect. Then four machine learning prediction models called Naïve Bayes [3, 4], Decision Tree [5, 6], Random Forest and Support Vector Machine are employed [7-10]. After our machine learning model is learned and tested on the dataset, cancer patients can accurately predict the severity of their cancer according to their living habits without complex inspection, which is especially suitable for people in remote areas. In the end, our proposed models all have high accuracy. Naïve Bayes model can get an accuracy of 86% in testing, while the other models get a higher accuracy. Decision trees and random forests get 94% and 98%. The average accuracy of different function models of SVM reached 97%.

2. Method

2.1 Dataset Description and Preprocessing

2.1.1 Dataset Description

The dataset is derived from [11], and contains 25 columns and 1000 rows. The first column is the patients' code name. The patient's year appears in the second column. The third column is the gender of the patient, denoted by 1 and 2 respectively. The features listed in columns four through twenty-four include: Air Pollution, Dust Allergy, Occupational Hazards, Genetic Risk, chronic Lung Disease, Balanced Diet, Obesity, Smoking, Passive Smoker, Chest Pain, Coughing of Blood, Fatigue, Weight Loss, Shortness of Breath, Wheezing, Swallowing Difficulty, Clubbing of Fingernails, Frequent Cold, Dry Cough and Snoring. The larger the number, the more severe the degree. The last column is the degree of cancer progression, indicating by Low, Medium, High. The objective of the collected dataset is to carry out the classification task (i.e., predicting the degree of cancer progression).

2.1.2 Preprocessing

Firstly, Heatmap was employed to display the calculation results. As a result, it can find which features are more correlated to the level of cancer in statistics. Then, we displayed the influence of random two features to the level of cancer by histogram. In theory, when more relevant factors change, the degree of cancer will have a corresponding significant change. The dataset uses Low, Medium, High to show the level of cancer. In this situation, it is necessary to make the string object into int or float. 1, 2 and 3 replace Low, Medium and High, respectively. Moreover, the dataset is divided into the training set and test set. The training set has 700 items, and the test set has 300 items. Finally, since patients' code name has no effect on cancer progression, they are removed before training and testing.

2.2 Machine Learning Models

After the data analysis of the dataset, we used various machine learning models to make a determination of the extent of cancer based on life factors due to their satisfactory performance in many tasks [12, 13], and made the learned models predict the extent of cancer. Then four machine learning prediction models called Naïve Bayes, Decision Tree, Random Forest and Support Vector Machine are employed. First, we use decision trees for training the dataset, which is a decision model built using a tree structure based on the attributes of the data. Each internal node corresponds to an

input property, the dataset is designated as the root node, and the child nodes represent the potential values of the parent node's attributes. The potential output values derived from the input attributes are represented by each leaf node. It selects the optimal features when performing tree splitting (dividing the data set) in the creation. We used the first 22 columns of the dataset as feature columns and the degree of cancer expressed in numbers as mentioned above as the classification criteria. Also set the filled parameter to true, which will fill the background color of the decision tree frame nodes.

Then, we employ a random forest model, which is made up of numerous separate decision trees that work together as a whole. Each tree in the random forest predicts a class, and the model will choose the class with the highest votes as its forecast. There are a random selection of 1000 samples to be put back with the 1000 samples (one random sample at a time and then return to continue the selection). As the samples at the decision tree's root node, these 1000 carefully chosen samples are utilized to train the decision tree. Each sample has 22 attributes, and when each node of the decision tree needs to be split, m attributes are randomly selected from these 22 attributes, satisfying the condition m << 22. Then some strategy (e.g., information gain, information gain rate, Gini strategy) is used to select 1 attribute from these m attributes as the splitting attribute for that node. We divide the leaf nodes by setting the maximum number of selected features to 4. In terms of maximum depth, we let the tree expand until all leaf nodes are of the same class of samples, or until the number of min_samples_split is reached. Let the minimum number of samples of leaf nodes be 3. If the number of samples is less than or equal to this value, the current node cannot be further divided.

The third machine learning model we adopt is the linear kernel function in the SVM model for the linearly differentiable case because of its low parameters and fast speed. Moreover, this model usually has high accuracy.

Finally, we use the plain Bayesian model. We use GaussianNB as the algorithmic class of the plain Bayesian. Unlike most of the other classification algorithms, the plain Bayesian directly finds the joint distribution P (X, Y) of feature output Y and feature X, and then derives the relationship using P(Y|X) =P(X,Y)/P(X). The prior probability P(Y=Ck) we use the default value, at this point P(Y=Ck) =mk/m. where m is the total number of training set samples and mk is the number of training set samples whose output is the kth category. After fitting the data using GaussianNB's fit method, we can make predictions.

3. Result and Discussion

3.1 Prediction based on Machine Learning Model

Table 1. Testing set performance

Model name	Naive Bayes	Decision tree	Random forest	SVM linear
Accuracy (%)	91.67	94.67	98.33	100

As shown in Table 1, we used Naive Bayes, Decision Trees, Random Forests, and SVM algorithms. Our data set has a capacity of 700, which come from different channels and include people with cancer of different ages and races. The data are all authentic, reliable and valuable. Based on the experimental results, it can be found that their accuracy varies. In this case, Naive Bayes has the lowest accuracy. Since the Bayesian theorem assumes that the influence of an attribute value on a given class is independent of the values of other attributes, which is often not true in practice, its classification accuracy may decline. This leads to its low accuracy. For the decision tree, we use pre-pruning, which is straightforward in thinking, simple in the algorithm, and high in efficiency. However, there are certain limitations and there is a risk of underfitting. lead to low precision. The reason for the high accuracy of random forest is mainly due to "random" and "forest", one makes it resistant to overfitting, and the other makes it more accurate. In our survey. We found that the accuracy of SVM linear is the highest, nearly 100.These models are very suitable for the data set we studied, indicating that the final prediction level will meet our expectations.

3.2 Feature Importance Analysis

3.2.1 Person Correlation Coefficient Analysis

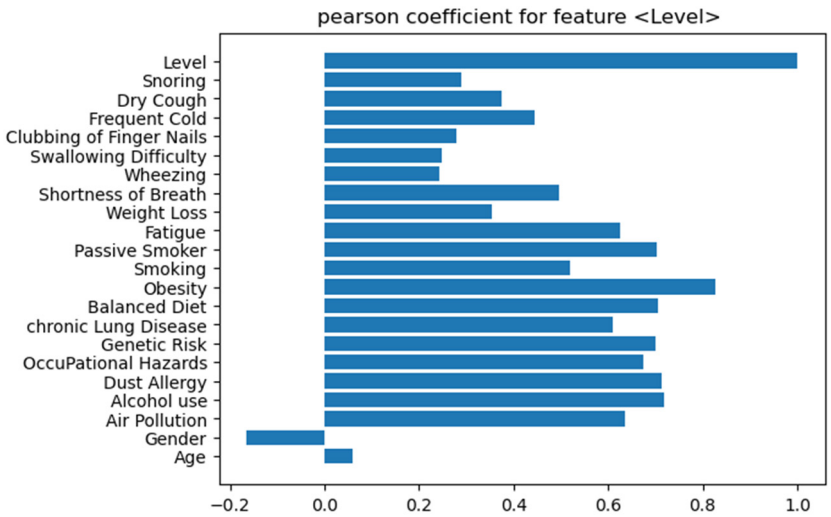

Fig 1. Pearson coefficient for feature

Figure 1 presents the chart of each feature's pearson coefficient with feature cancer level (level feature itself is not considered). We can see that obesity has the biggest value of coefficient, Passive Smoker, Banlanced diet, Genetic Risk, Dust Allergy and Alcohol use are almost equal in second place. So, all of this feature has a positive correlation with cancer Level. Surprisingly, Gender's person coefficient is minus value, which shows a negative correlation. Actually, Gender is a kind of discrete variable. 0 represent female and 1 represent male. Negative correlation show's that female may have a higher risk of getting cancer in a more serious extent.

3.2.2 Decision Tree Framework and Feature Importance

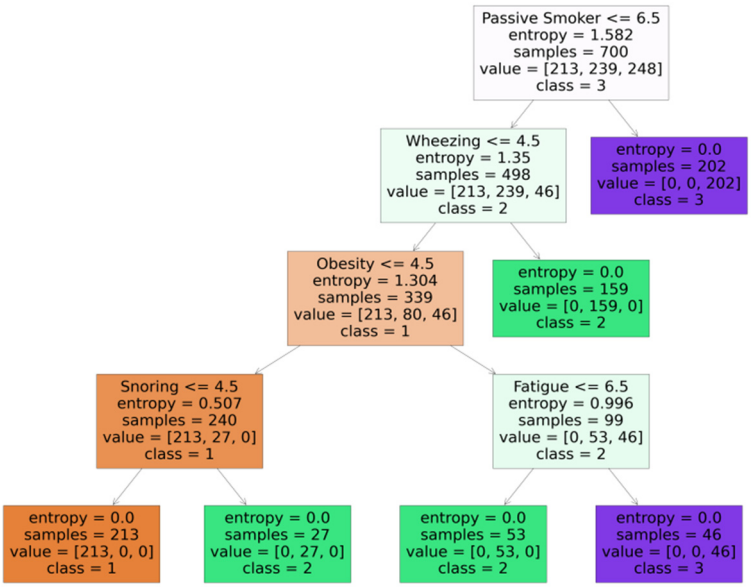

Fig 2. Decision tree visualization

The decision tree framework as shown in Figure 2. can be obtained after we carry out the visualization procedure. To our surprise, the decision tree framework is much simpler than we thought, because we put 21 features into model to fit and train, but we just find only 5 features do enough contribution to be classify, other features did not get the chance to their own contribution under these 5 features' influence.

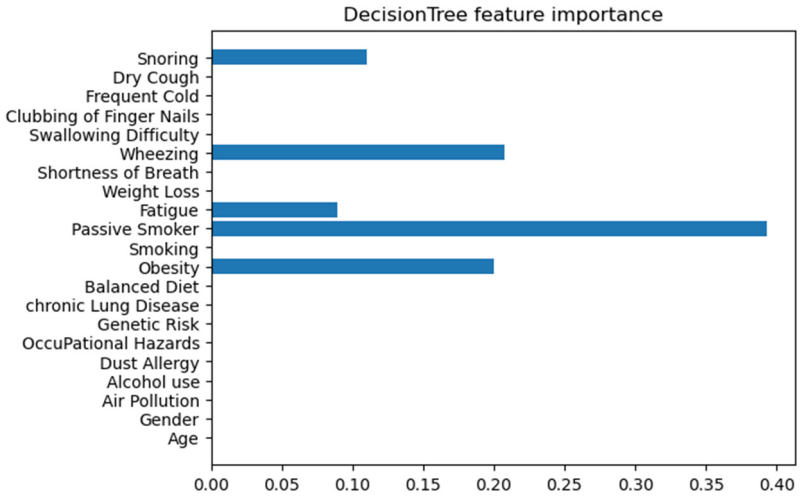

Fig 3. Decision tree feature importance

We use the built-in function to get the feature importance and do some data visualization. These 5 features have a considerable importance value and make other features' importance become useless. Passive Smoker has the biggest value among the whole data.

3.2.3 Random Forest and Feature Importance

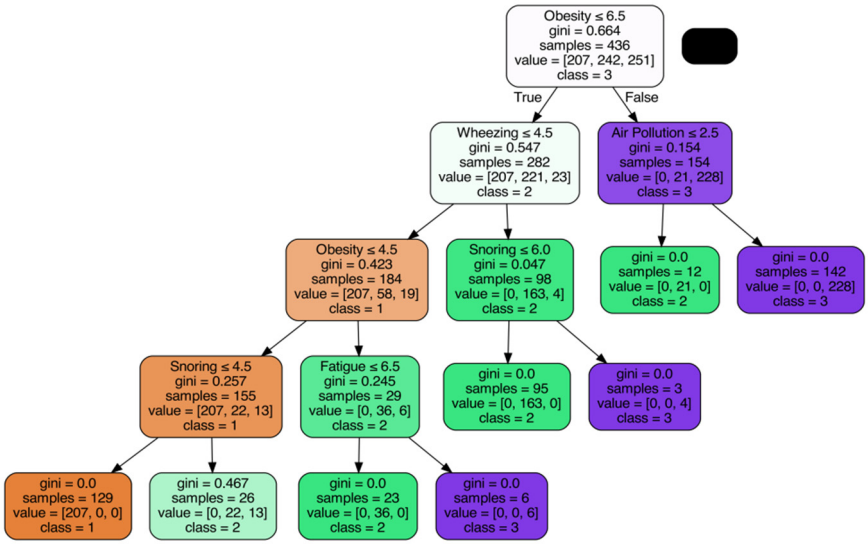

Fig 4. Random forest visualization

Here is one of the trees in the Random Forest, because of random forest use part of the features to train the model, we can see that there are more features show that contribution to the result, like Air Pollution. Because of more and more features' importance can be shown, random forest can reduce the risk of overfit and provide us with a much more objective result.

Here is the RandomForest feature importance chart. We can see this model not only save the prominent features of decision tree model, but also let other feature show their contribution. Gender has 0 importance in random forest means Gender didn't have a power of classification, it is different with the result of pearson coefficient. Actually, pearson coefficient can be affected by other features, so sometimes it is not wisely to use pearson coefficient to judge a correlation without other method.

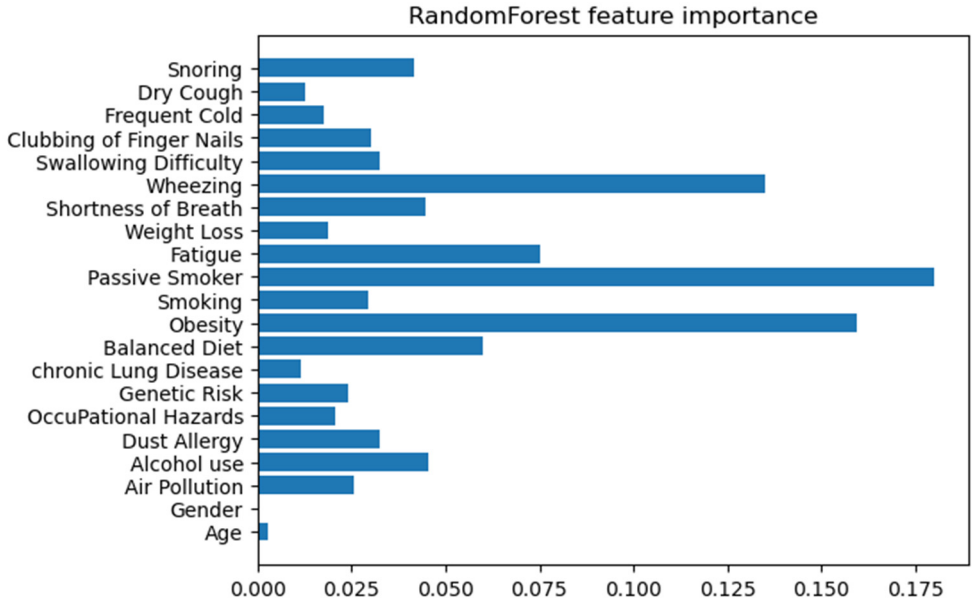

Fig 5. Random forest feature importance

3.2.4 Performance based on the Selected Features

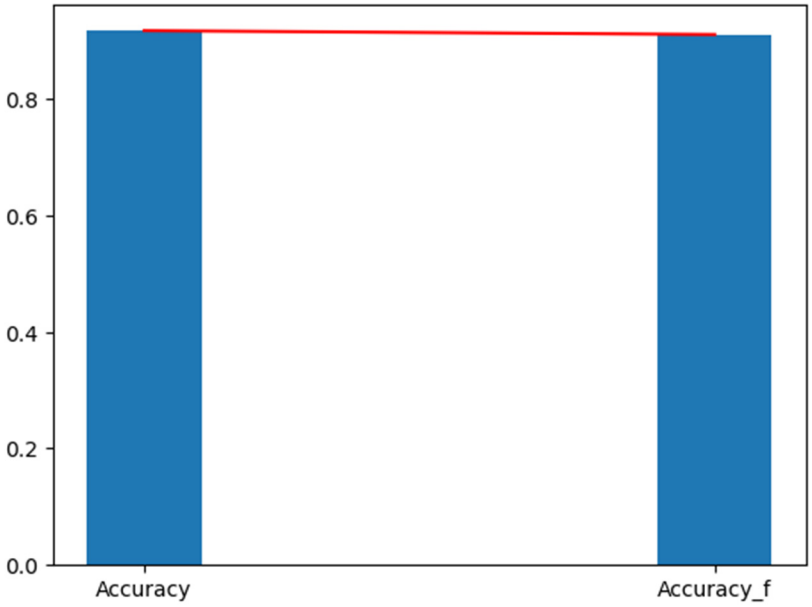

Fig 6. Test result on different features

Accuracy is the result of whole 22 features, and Accuracy_f is the result of 6 main features that we select. We can see the accuracy doesn't change very much and remain a high accuracy finally, which shows that these 6 features are enough to measure one's cancer level. By reducing the number of features, we need to concern, we can speed up inference and training procedure, the efficiency can be promoted in a large extent.

4. Conclusion

To understand the extent to which life factors in cancer patients affect their cancer, we chose to use machine learning to solve this problem. To be more specific, we used machine learning to explore which factors are more closely related to the degree of cancer, so that cancer patients can infer the severity of cancer according to their living habits. After testing, the model we selected has a high accuracy. Then we proceed to the description and processing of the dataset. A brief introduction was

given to the machine learning model we used. We preprocess the data and use machine learning predictive models such as Naive Bayes, Decision Tree and Random Forest. The svm linear model was found to be a good fit for the dataset we studied. Finally, the important analysis of features is carried out, using detailed pictures and data to show the impact of features on cancer, many features are put into the model for fitting and training, and important features are selected, so that people can analyze important features to judge a person's cancer level.

References

[1] GONG Xing. Analysis of changes in serum and miR indexes in lung cancer patients. Journal of Practical Cancer, 2015(5): 645-647.

[2] LIU Meiyu, CHEN Ping, LI Weibo, et al. Stepwise discriminant analysis of morphological parameters of cancer cells with simple breast cancer cells in the middle and advanced stages. Chinese Journal of Stereoscopy & Image Analysis, 2001, 6(1): 26-28.

[3] Webb, Geoffrey I., Eamonn Keogh, and Risto Miikkulainen. Naïve Bayes. Encyclopedia of machine learning 15, 2010, 713-714.

[4] Rish, Irina. An empirical study of the naive Bayes classifier. IJCAI 2001 workshop on empirical methods in artificial intelligence. Vol. 3. No. 22. 2001.

[5] Myles, Anthony J., et al. An introduction to decision tree modeling. Journal of Chemometrics: A Journal of the Chemometrics Society 18.6, 2004, 275-285.

[6] Quinlan, J. Ross. Learning decision tree classifiers. ACM Computing Surveys (CSUR) 28.1, 1996, 71-72.

[7] Biau, Gérard, and Erwan Scornet. A random forest guided tour. Test 25.2, 2016, 197-227.

[8] Rigatti, Steven J. Random Forest. Journal of Insurance Medicine 47.1, 2017, 31-39.

[9] Noble, William S. What is a support vector machine? Nature biotechnology 24.12, 2006, 1565-1567.

[10] Hearst, Marti A., et al. Support vector machines. IEEE Intelligent Systems and their applications 13.4, 1998, 18-28.

[11] Data world. https://data.world, 2022.

[12] Yu, Q et al. Semantic segmentation of intracranial hemorrhages in head CT scans. 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS). IEEE, 2019.

[13] Erickson, Bradley J., et al. Machine learning for medical imaging. Radiographics 37.2, 2017, 505.