Sensors and Materials, Vol. 33, No. 8 (2021) 2639–2654
MYU Tokyo

S & M 2645

Classification of Esophageal Adenocarcinoma, Esophageal Squamous Cell Carcinoma, and Stomach Adenocarcinoma Based on Machine Learning Algorithms

Xiaoping Chen,[1,2] Lihui Zheng,[1] Jianqi Yao,[1*] and Cheng-Fu Yang[3,4**]

[1]College of Mathematics and Statistics & FJKLMAA, Fujian Normal University, Fuzhou 350000, China
[2]Center for Applied Mathematics of Fujian Province, Fujian Normal University, Fuzhou 350000, China
[3]Department of Chemical and Materials Engineering, National University of Kaohsiung, Kaohsiung 811, Taiwan
[4]Department of Aeronautical Engineering, Chaoyang University of Technology, Taichung 413, Taiwan

(Received March 23, 2021; accepted June 17, 2021)

Keywords: EAC, ESCC, SAC, machine learning algorithm, confusion matrix

Esophageal and gastric cancers are common malignant tumors. In medicine, it is difficult to differentiate the sickness symptoms of esophageal adenocarcinoma (EAC), esophageal squamous cell carcinoma (ESCC), and stomach adenocarcinoma (SAC). In particular, the molecular characteristics of EAC and SAC are very similar, which makes them difficult to distinguish. Information collected by sensors can be analyzed by machine learning. In this study, we used cancer data published in Nature in 2017, which were downloaded from cBioPortal, to classify the three types of cancer by five machine learning algorithms, and we compared the classification effects for different models by calculating confusion matrices. According to the research data in this paper, the random forest (RF) model is the best of the five machine learning classification models for the overall classification effect of the three types of cancer. More specifically, the classification effect of this model is the best for EAC, whereas the classification effect for ESCC is not ideal. The classification based on the RF model can effectively enhance the differentiation between the symptoms of EAC, SAC, and ESCC, enabling cancer patients to receive more accurate treatment and have an improved prognosis.

1. Introduction

Esophageal cancer is a common malignant tumor, and its morbidity and mortality rank eighth and fifth out of all malignant tumors, whereas the morbidity and mortality of gastric cancer rank fifth and third out of all malignant tumors, respectively. Esophageal carcinoma is histologically divided into esophageal adenocarcinoma (EAC) and esophageal squamous cell carcinoma (ESCC).[(1)] In recent decades, the morbidity of esophageal cancer in Western countries has increased several times, the five-year survival rate is in the range of 12–20%,[(2,3)] and esophageal cancer has caused more than 400000 deaths worldwide every year.[(4)] Esophageal cancer mainly occurs in the lower esophagus and is associated with obesity, gastric reflux, and Barrett's esophagus. By analyzing the molecular characteristics of patients with esophageal and gastric

*Corresponding author: e-mail: jianqiyao2019@163.com
**Corresponding author: e-mail: cfyang@nuk.edu.tw
https://doi.org/10.18494/SAM.2021.3392

cancers, it was found that EAC and stomach adenocarcinoma (SAC) have very similar unstable chromosomal variations, which indicate that these cancers can be considered a single disease entity.[5] The increases in the morbidities of EAC and proximal stomach cancer are synchronous.[6] The boundaries between SAC and EAC and the classification of adenocarcinoma that crosses the gastroesophageal junction are still indistinct, and there are also many disputes about the practicability of histological features.[7–9]

Given the uncertainties of the boundaries between EAC and SAC, by analyzing the EAC, ESCC, head and neck squamous cell carcinoma (HNSCC), and SAC, it has been found that the symptoms of ESCC are similar to those of HNSCC, and the symptoms of EAC are similar to those of SAC. The distinction between ESCC and EAC has not only known histopathological and epidemiological characteristics but also known molecular characteristics. Many methods in machine learning can provide the importance of independent variables in a classification and their influence on classified dependent variables, and be used to evaluate the relationship between independent variables and classified dependent variables. These results are more objective and reasonable than the logistic regression model in the interpretation of coefficients. Machine learning can also combine different competing models to produce more accurate predictions than a single model. At present, there are many sensors collecting data, and there is useful information in these data. By combining these data processing and model training in machine learning, complex tasks can be solved. Until now, very few studies have investigated the use of machine learning classification to distinguish the symptoms of SAC, ESCC, and EAC. Thus, we investigated the use of machine learning algorithms to classify the different cancer types, and confusion matrices were investigated to measure and compare the classification effects of different models. The effects of important variables on the different cancer types were identified, which could promote better classification of these cancers and the emergence of new therapies.

2. Subjects and Methods

2.1 Study subjects

In this study, all the used cancer data, which were published in Nature in 2017, were downloaded from cBioPortal. The data include the clinicopathological and molecular characteristics of 90 cases of ESCC, 79 cases of EAC, 388 cases of SAC, and two cases of esophageal gastric cancer. These cancer data were obtained after processing fresh frozen tumor samples, which were obtained from multiple countries with informed consent and approval by the local institutional review board.

2.2 Clinical measurements and genetic assessments

Germline deoxyribonucleic acid (DNA) was extracted from blood or nonmalignant esophageal mucosa in these data samples, and complete exon sequencing, analysis of single-nucleotide polymorphism (SNP) array, evaluation of somatic copy-number alterations (SCNAs), analysis of DNA methylation, and mRNA and microRNA sequencing were conducted.

2.3 Statistical analysis

In this paper, there were 559 samples with 103 variables in total, including clinical pathology, histopathology, and molecular characteristics. After deleting the variables with missing rates greater than 60%, the remaining 75 variables were imputed. Many variables existed in a variety of forms and had the same missing information, so five imputation methods were chosen: missForest, *k* neighborhood, center interpolation, classification regression tree, and random forest (RF). Finally, we chose missForest as the base algorithm because it yielded the best results. Although many packages can be used to impute missing values, they usually do not recognize categorical variables, whereas missForest can handle missing values from continuous variables and categorical variables.

All the classifications were performed using the imputed values. Because there were only two cases of esophageal gastric cancer, this category was not suitable for classification and was deleted. The classification of ESCC, EAC, and SAC was performed on the remaining 557 samples. The data were divided into a training set and a test set, with 70% of the data randomly selected as the training set and 30% as the test set. Cancers were classified using a variety of machine learning methods, such as traditional decision trees, conditional inference trees, bagging, AdaBoost, and RF. The misclassification rate, accuracy rate, precision rate, and recall rate of the classification methods were calculated using the confusion matrix to evaluate the different classification methods. All statistical analyses were performed with R software (version 3.5.3).

2.3.1 Decision tree model

The decision tree model is an easy-to-use and nonparametric classifier that classifies instances based on variable characteristics. The structure is tree-shaped, composed of nodes and directed edges, and does not require any priori assumptions on the data. For the decision tree model, its calculation speed is high, its measured results are easy to interpret, and its robustness is strong. Based on the ID3 algorithm and the C4.5 algorithm, the main characteristics of decision tree learning are feature selection, decision tree generation, and branch reduction. When learning the training set samples, the decision tree model is constructed according to the minimum loss function, and a set of test data can be classified with the decision tree model. An important concept in the decision tree algorithm is entropy, which is a measured result of the uncertainty of random variables. If we let X be a discrete random variable with a finite number of values, the probability distribution can be expressed as

$$P(X = x_i) = p_i,\ i = 1,2,...,n\,. \tag{1}$$

Then, the entropy of X is defined as

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i\,. \tag{2}$$

The greater the entropy is, the greater the uncertainty of the random variables. The conditional entropy $H(Y|X)$ represents the uncertainty of random variable Y under the condition that random variable X is known. This is the mathematical expectation of the entropy of the conditional probability distribution given the conditions for X, i.e.,

$$H(Y|X) = \sum_{i=1}^{n} p_i H(Y|X = x_i) . \tag{3}$$

The information gain represents the information of a known feature X, leading to the degree of information uncertainty reduction of Y. The information gain of feature A for the training data set D is $g(D, A)$, which is the difference between the empirical entropy $H(D)$ of D and the empirical conditional entropy $H(D|A)$ for the given condition of feature A, that is,

$$g(D, A) = H(D) - H(D|A) . \tag{4}$$

The decision tree model applies the information gain criteria to select the features, solves the information gain of various schemes under different conditions by means of diagrams, and then makes decisions through comparison processes. Features with large information gains have stronger classification capabilities.

2.3.2 Bagging model

Bagging, which is also known as bootstrap aggregating, is an integration technique that trains classifiers by selecting S new datasets from the original dataset, and the observations in these new datasets are selected without replacement. The trained classifiers are used to classify the new samples, and then the results of the classification of all classifiers are counted, and the most frequent category is the final tag.

The input of the sample can be set as $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$, the number of iterations of the weak classifier can be represented by T, and the output is the final strong classifier $f(x)$.

(1) For $t = 1, 2, ..., T$,
   (a) t are the random samples of m observations, which are collected from the training set to obtain a sampling set D_t containing m samples.
   (b) The tth weak learner is trained with sampling set D_t.

(2) The category with the most votes cast by the T weak classifier is the one we finally choose.

Bagging classification is a particularly effective technology when learning is unstable and tends to overfit, i.e., small changes in training data lead to significant changes in the predicted output. Models prone to overfitting do not generalize well outside the training data. Bagging works well with high-variance models, such as decision trees, and when it is used with low-variance models, such as linear regression, it does not significantly affect the learning process. It effectively reduces the variance by clustering together individuals, which are composed of different statistical attributes (such as different standard deviation means, etc.). The number of

basic learners to be selected depends on the characteristics of the data set. Bagging can be executed in parallel to check for excessive computing resources, which is a major advantage, and it is a common algorithm booster used in various fields.

2.3.3 Adaptive boosting model

Adaptive boosting (abbreviated as AdaBoost) is a common boosting and iterative algorithm, and its basic learner is the classification tree. Each iteration can generate a new classifier on the training set, then the classifier is used to classify all the samples to recognize the importance of each sample. Specifically, the algorithm assigns a weight to each training sample, and each sample is labeled with a new classifier after training. If the focal point of a sample has been classified correctly, its weight will be reduced. If the focal point of a sample has not been classified correctly, its weight will be increased. The larger the weight is, the higher the proportion of samples will be in the next training iteration; that is, these points with high error rates will receive more attention in subsequent training iterations. The iteration process lasts until the error rate is small enough or a certain number of iterations is reached.

Assuming that the sample is (X_1, Y_1), (X_2, Y_2), …,(X_n, Y_n), to simplify the description, the dependent variable is assumed to be a binary variable $Y \in \{0, 1\}$. The concrete steps of the AdaBoost algorithm are as follows. (1) Select the initial self-service sampling weight as $w_i^{[0]} = 1/n$ ($i = 1, \ldots, n$) for the observation point, and set $m = 0$. (2) Increase m by 1, then use the classification tree to fit the weighted sampling, with weight $w^{[m-1]}$ used to generate classifier $\hat{g}^{[m]}(\cdot)$. (3) Calculate the weighted misjudgment rate in the sample as follows:

$$\begin{aligned} err^{[m]} &= \sum_{i=1}^{n} w_i^{[m-1]} I(Y_i \neq \hat{g}^{[m]}(X_i)) / \sum_{i=1}^{n} w_i^{[m-1]}, \\ \alpha^{[m]} &= \ln\left(\frac{1 - err^{[m]}}{err^{[m]}}\right), \end{aligned} \tag{5}$$

then update the weights as

$$\begin{aligned} \tilde{w}_i &= w_i^{[m-1]} \exp(\alpha^{[m]} I(Y_i \neq \hat{g}^{[m]}(X_i))), \\ w_i^{[m]} &= \tilde{w}_i / \sum_{j=1}^{n} \tilde{w}_j. \end{aligned} \tag{6}$$

Steps (2) and (3) are repeated until the predetermined number of iterations is reached. Thus, a combinatorial classifier classified by weighted votes is established as

$$\hat{f}_{adaboost}(x) = \arg\min_{y \in \{0,1\}} \sum_{m=1}^{m_{stop}} \alpha^{[m]} I(\hat{g}^{[m]}(x) = y). \tag{7}$$

AdaBoost can be used to improve the performance of any machine learning algorithm, and it is most suitable for algorithms with poor learning ability. To improve the detection accuracy,

AdaBoost requires a large set of training samples, each training of a weak classifier requires a sample, and each sample has many characteristics. Therefore, the number of calculations required to obtain optimal weak classifiers from a large number of features through training is huge.

2.3.4 RF model

An RF classifier contains multiple decision trees, and the output category is determined by the mode of the category output by an individual tree. The RF is composed of multiple decision trees, and there is no correlation between each decision tree in the forest. The final output of the model is determined by each decision tree in the forest. When classification problems are handled, the final category of each decision tree in the forest is given for the test samples. Finally, the output category of each decision tree in the forest is comprehensively considered, and the categories of test samples are determined by voting.

To evaluate the role of each variable in the classification model, an RF classifier gives the importance score of each variable. In an RF classifier, each node is segmented using the best node in a randomly selected set of sub-predictors for that node. Compared with other classifiers, this somewhat counterintuitive strategy performs very well in each form and is robust to overfitting. In addition, an RF classifier is very friendly because it has only two parameters (the number of variables in the random children of each node and the number of trees in the forest) and is usually not very sensitive to their values.

3. Results

3.1 Classification results for different classification models

In total, 34 clinical pathological variables and molecular variables were selected for the exploration of cancer classification models, and the cancer classification results for different classification models were analyzed. The data were divided into training and test sets; 70% of samples were randomly selected as the training set and the remaining 30% of samples were used as the test set. The classification results obtained by the training set for different models were analyzed.

3.1.1 Classification results for decision tree model

3.1.1.1 Classic decision tree classification results

The classic decision tree algorithm usually involves an oversized tree, which leads to overfitting and poor classification performance for units outside the training set. Therefore, ten-fold cross-validation is used to select the tree with the smallest prediction error. Table 1 shows the complexity parameter (CP) values, which are used to help set the size of the final tree by imposing a penalty on an oversized tree. The size of the tree is the number of branches (*nsplit*), and a tree with n branches will have $n + 1$ terminal nodes. *rel error* is the error corresponding to

various trees in the training set, the cross-validation error (*xerror*) is based on the tenfold cross-validation error from the training sample, and *xstd* is the standard deviation of the cross-validation error.

Figure 1 shows the relationship between the cross-validation error and the CP value. For all trees where the cross-validation error is within one standard deviation of the minimum cross-validation error, the smallest tree will be the best tree. Figure 1 shows that the optimal tree corresponds to three partitions. Table 1 shows that the CP value corresponding to the three partitions was 0.0714, and the most important branch was cut off according to the CP value by using the prune function.

Figure 2 shows the pruned classic decision tree with the ideal size used to predict cancer types. When Mutation_Count was larger than 158, the type of cancer was EAC, indicating that Mutation_Count can be used as a significant indicator to distinguish between SAC and EAC.

Table 1
CP values.

	CP	*nsplit*	*rel error*	*xerror*	*xstd*
1	0.11160714	0	1.0000000	1.0000000	1.0000000
2	0.08035714	2	0.7767857	0.9464286	0.07840673
3	0.07142857	5	0.5357143	0.7321429	0.07182627
4	0.06250000	6	0.4642857	0.6964286	0.07050740
5	0.02678571	7	0.4017857	0.6696429	0.06947085
6	0.01785714	9	0.3482143	0.6696429	0.06947085
7	0.01000000	12	0.2946429	0.6607143	0.06911596

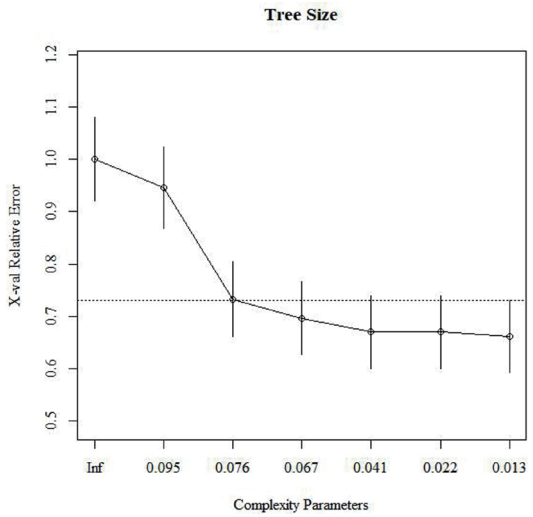

Fig. 1. CP and cross-validation errors.

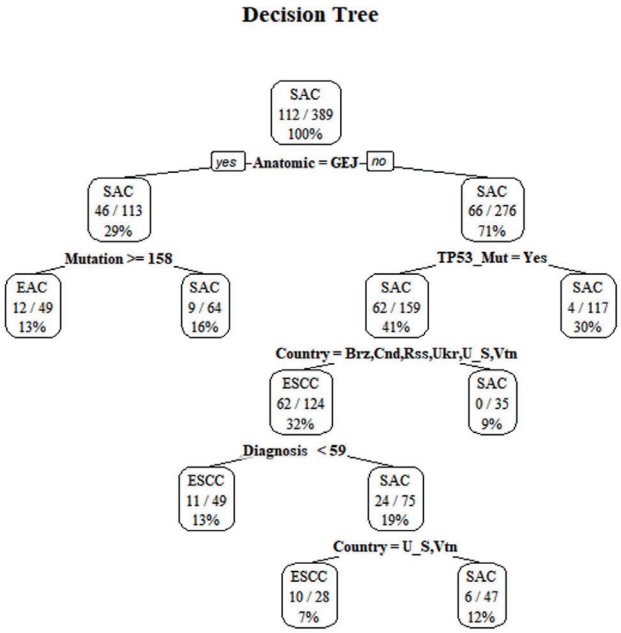

Fig. 2. Pruned classic decision tree used to predict cancer types.

However, Country and Diagnosis_Age can be used as indicators for distinguishing between ESCC and SAC, which could help in the prediction of cancer types and facilitate subsequent targeted treatments.

3.1.1.2 Classification results for conditional inference tree

A variant of the traditional decision tree is a conditional inference tree, which is similar to a traditional decision tree, but the selections of variables and partitions are based on significance testing, pruning is not necessary, and the generation process is more automated. Figure 3 shows a conditional inference tree in which the shaded area in each node represents (from left to right) the proportions of ESSC, EAC, and SAC. The object attributes in the conditional inference tree were Country, Anatomic_Site, TP53_Mutate, Lymphocyte_Infiltration, Diagnosis_Age, and Histologic_Grade.

3.1.2 Classification results for bagging model

It can be seen from Fig. 4 that Anatomic_Site, Country, Genome_Altered, Diagnosis_Age, Mutation_Count, and Histologic_Grade were the most important variables in the bagging model, similar to the results obtained from the decision tree (Anatomic_Site, Country, Diagnosis_Age, and Histologic_Grade).

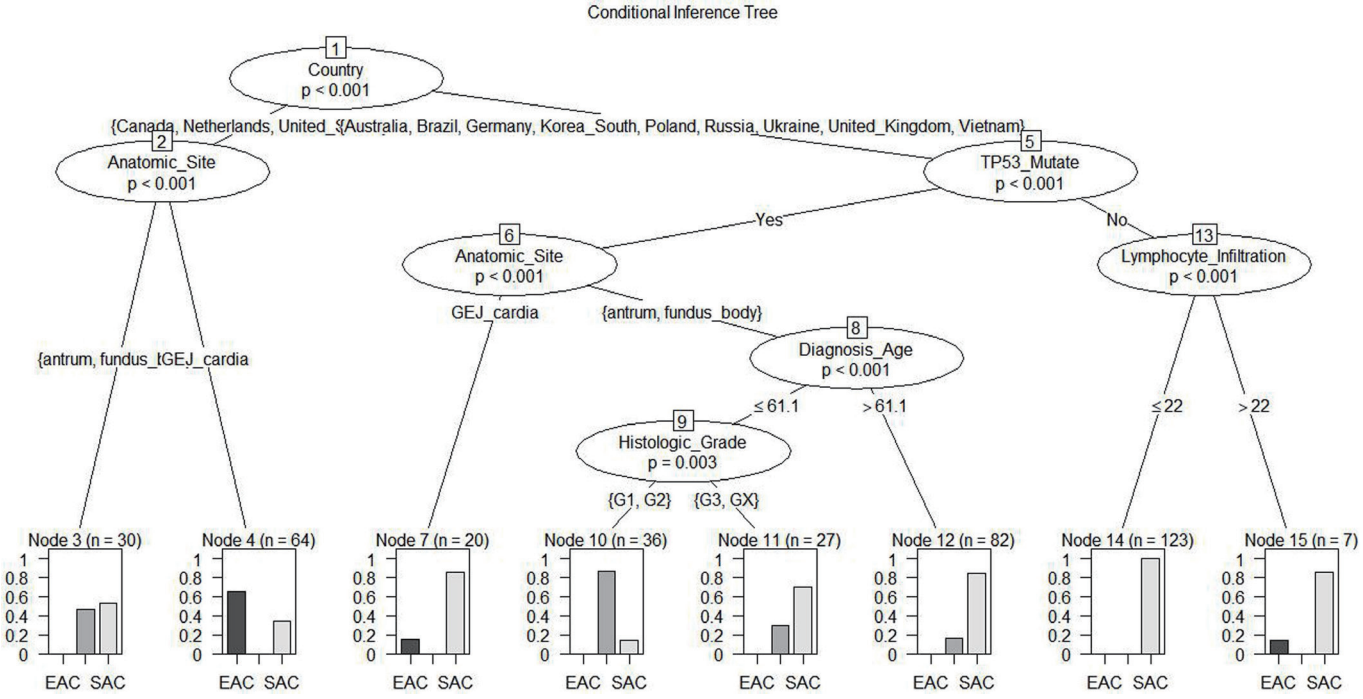

Fig. 3. Conditional inference tree classification of cancer types.

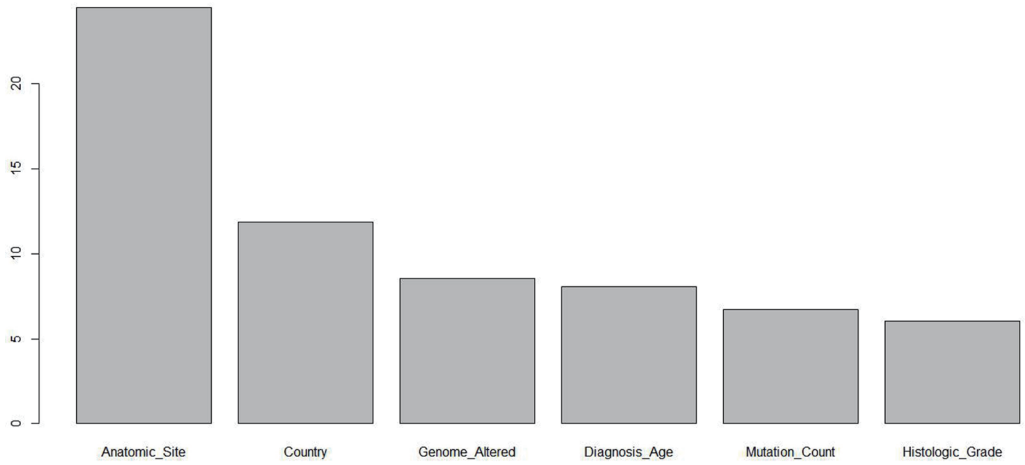

Fig. 4. Variable importance map for the classification with the bagging model.

3.1.3 Classification results for an AdaBoost model

As shown in Fig. 5, Anatomic_Site, Country, Diagnosis_Age, Genome_Altered, Lymphocyte_Infiltration, and Mutation_Count were the most important variables for the classification of the AdaBoost model, similar to the results obtained for the decision tree model (Anatomic_Site, Country, Diagnosis_Age, and Lymphocyte_Infiltration).

3.1.4 Classification results for RF model

It can be seen from Fig. 6 that Anatomic_Site, Country, Histologic_Grade, Lymphocyte_Infiltration, Genome_Altered, and Mutation_Rate were important variables in the RF model, which were the same as the most important variables in the decision tree model and the bagging and AdaBoost models. This finding shows that Anatomic_Site, Country, and Genome_Altered were the most important indicators in distinguishing classification models and are an important basis for classifying cancer types.

3.2 Computation of confusion matrix and comparison of classification results for different classification models

There are several methods for evaluating classification models: confusion matrices, which include gain charts, lift charts, KS charts, and receiver operating characteristic curves. The data were divided into a training set and a test set, with 70% of the data randomly selected as the training set and 30% selected as the test set for comparison. We used the confusion matrix of the training set to determine the best classification method, and the accuracy, precision, and recall rates were calculated using the obfuscation matrix. The accuracy rate is the proportion of all correct predictions (positive and negative), the precision rate is the proportion of correct

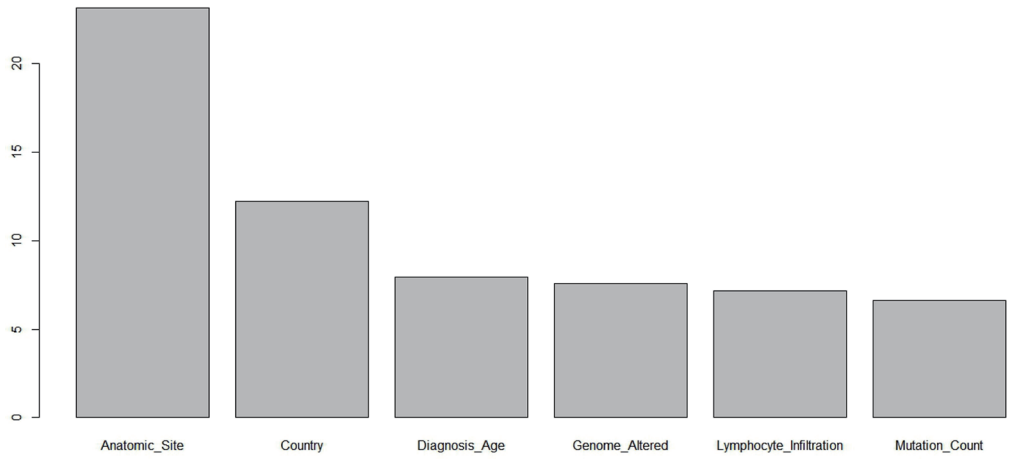

Fig. 5. Variable importance map for the classification with the AdaBoost model.

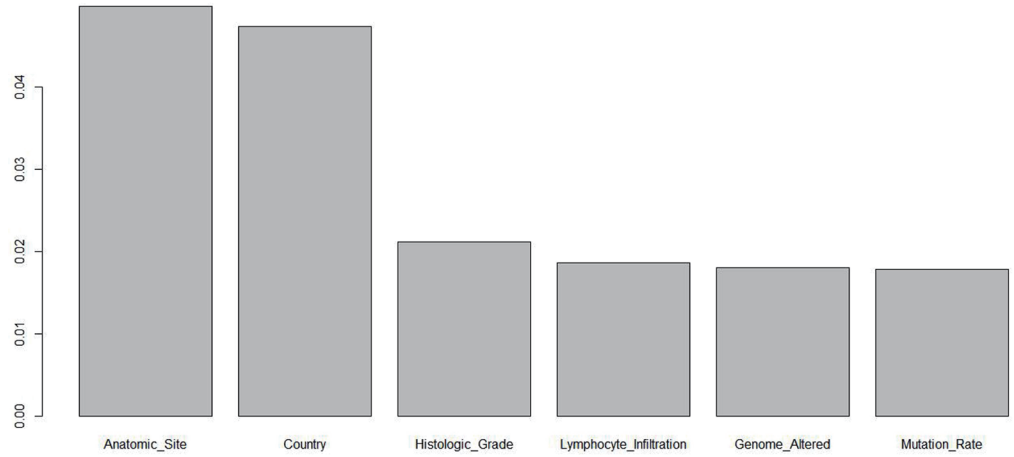

Fig. 6. Variable importance map for the classification with the RF model.

predictions that are positive to the total number of predictions that are positive, and the recall rate is the percentage of all positive predictions that are correct.

The confusion matrices for the classical decision tree model, conditional inference tree model, bagging model, AdaBoost model, and RF model for the test set are shown in Table 2. From each confusion matrix, we can calculate the accuracy rates, precision rates, and recall rates for classification models for the test set, and the results are shown in Table 3.

According to Table 2, the error rates of the classical decision tree model, conditional inference tree model, bagging model, AdaBoost model, and RF model were 21.40, 19.00, 15.40, 12.5, and 11.30%, respectively. From these results, the RF model has the best classification effect, followed by the AdaBoost model, bagging model, and conditional inference tree model, while the decision tree model has the lowest classification effect. As can be seen from Table 3, both the precision and recall rates of all classification models for ESCC were low, and the classification effect of

Table 2
Confusion matrices for the classification models for the test set.

	Classical decision tree model			Conditional inference tree model		
Actual \ Predicted	EAC	ESCC	SAC	EAC	ESCC	SAC
EAC	27	0	6	32	0	1
ESCC	0	18	6	0	8	16
SAC	5	19	87	8	7	96
	Bagging model			AdaBoost model		
Actual \ Predicted	EAC	ESCC	SAC	EAC	ESCC	SAC
EAC	29	0	4	30	0	3
ESCC	0	17	7	0	18	6
SAC	4	11	96	5	7	99
	RF model					
Actual \ Predicted	EAC	ESCC	SAC			
EAC	26	0	7			
ESCC	0	15	9			
SAC	1	2	108			

Table 3
Evaluation index for classification models for the test set.

	Classical decision tree model			Conditional inference tree model		
Evaluation index	EAC (%)	ESCC (%)	SAC (%)	EAC (%)	ESCC (%)	SAC (%)
Accuracy rate	93.45	85.12	78.57	94.64	86.31	80.95
Precision rate	84.38	48.65	87.88	80.00	53.30	85.00
Recall rate	81.82	75.00	78.38	96.97	33.33	86.49
	Bagging model			AdaBoost model		
Evaluation index	EAC (%)	ESCC (%)	SAC (%)	EAC (%)	ESCC (%)	SAC (%)
Accuracy rate	95.24	89.29	84.52	95.24	92.23	87.50
Precision rate	87.88	70.83	89.72	85.71	72	91.67
Recall rate	87.88	70.83	86.49	90.91	75	89.12
	RF model					
Evaluation index	EAC (%)	ESCC (%)	SAC (%)			
Accuracy rate	95.24	93.45	88.69			
Precision rate	96.30	88.24	87.10			
Recall rate	78.79	62.5	97.30			

the model for ESCC was inadequate, whereas the classification effect for EAC was the best. However, for the RF model, the classification effect for SAC was the best, and the recall rates for ESCC and EAC were the lowest at 78.79 and 62.5%, respectively. The comprehensive comparison showed that the RF model had the highest effect.

For the data set we studied, by analyzing the confusion matrix for different classification methods, we found that the RF model has the best effect in predicting EAC because it has the highest results for the classification of accuracy, precision, and recall rates. Other classification methods have the best correct discrimination for EAC, indicating that it is easily distinguished from the other two cancers.

In this paper, the categorical variables of cancer include clinicopathological characteristics, demographic characteristics, and molecular characteristics, and the variables are mainly

discrete, although some are continuous. The above models can deal with both continuous and discrete data effectively. Overfitting easily occurs in the decision tree model, but we avoid this problem by pruning. However, the classification accuracy of the decision tree model is inferior to those of the other models. AdaBoost improved the performance through boosting, in which it is unnecessary to screen features and overfitting occurs, making AdaBoost suitable for cases with more complex data types in this paper. RF is modified by bagging and is not prone to overfitting, because the training samples do not account for all the samples. When dealing with classification imbalances, RF can also provide an effective method to balance the error of the data set, giving it a better classification effect than the other classification algorithms.

3.3 Classification of cancers for RF model

By analyzing the confusion matrix, it is concluded that the RF model can classify the cancer data better than the other models. Next, we use the RF model to classify cancer types and analyze the classification results, and the confusion matrix and error for the RF model are presented in Table 4. For the RF classification, the error rate in classifying the three cancers is very low, especially for SAC.

The Gini index (Gini inequality) indicates the probability that a randomly selected sample will be split in the sample set. The smaller the Gini index is, the smaller the probability that the selected sample in the set will be split, that is, the higher the purity of the set, and the higher the Gini index is, the less pure the set. The Gini index is equal to the probability of a sample being selected multiplied by the probability of the sample being misclassified:

$$Gini(p) = \sum_{k=1}^{K} p_k (1 - p_k) = 1 - \sum_{k=1}^{K} {p_k}^2. \tag{8}$$

Figure 7 shows accurate measurements of the importance of each variable obtained by using the three levels of the dependent variables (cancer types) and the effect of the variables on the prediction accuracies of all cancer types and the Gini index. The larger the number of chaotic categories contained in the population, the larger the Gini index will be (similar to the concept of entropy). For a certain node, the lower the entropy is, the purer it will be, and the smaller the Gini index is, the purer the Gini index will be. Thus, the purer the node is, the more it can determine which type it belongs to, and the more ideal the result is. As shown in Fig. 7, for the RF model, the variables with the highest importance for EAC were Anatomic_Site, Country,

Table 4
Confusion matrix and error for the RF model for the test set.

Actual	Predicted			Class error
	EAC	ESCC	SAC	
EAC	64	0	15	0.189
ESCC	0	55	35	0.389
SAC	6	6	376	0.031

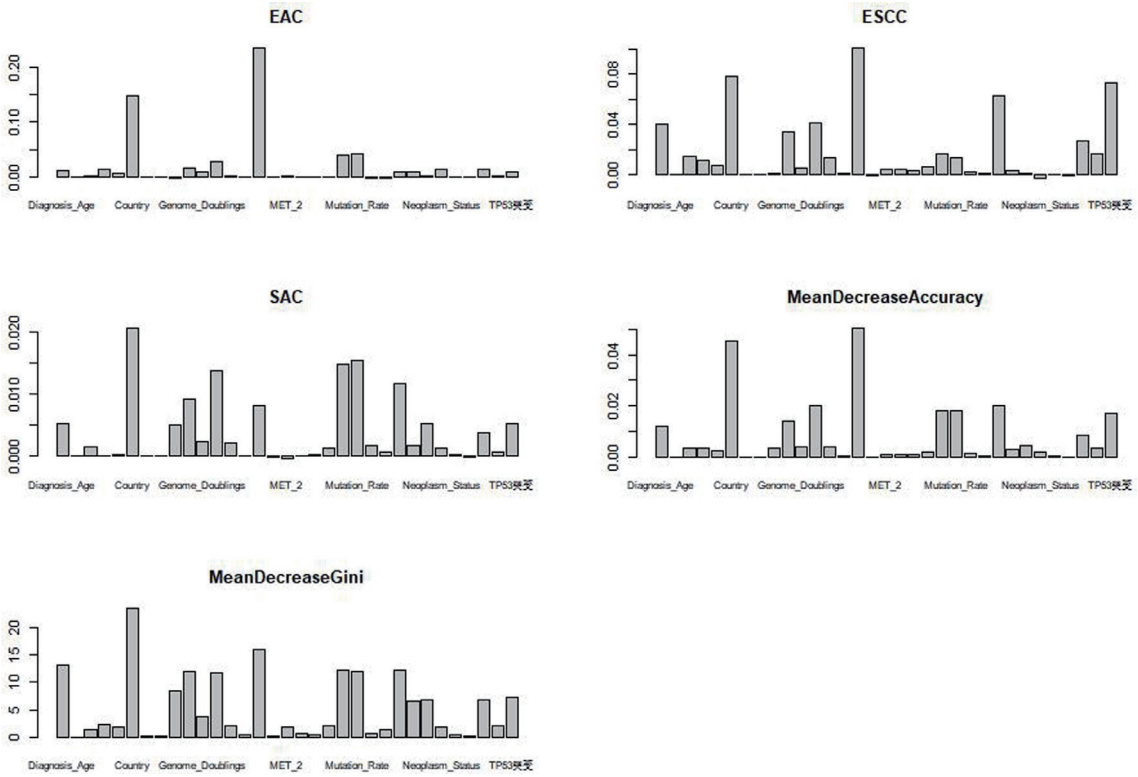

Fig. 7. Variable importance maps of different measures for the classification with the RF model.

Mutation_Rate, Mutation_Count, Genome_Altered, and Histologic_Grade. The variables that were most important for ESCC were Anatomic_Site, Country, TP53_Mutate, Lymphocyte_Infiltration, Diagnosis_Age, and Genome_Altered.

The variables that were most important for SAC were Mutation_Count, Mutation_Rate, Country, Histologic_Grade, Genome_Altered, and Lymphocyte_Infiltration. The variables with the most precise measures that affect the accuracy of prediction of all cancer types were Anatomic_Site, Country, Mutation_Count, Mutation_Rate, Histologic_Grade, and Genome_Altered. The variables with the highest prediction accuracy for all cancer types based on the Gini index were Country, Anatomic_Site, Mutation_Count, Mutation_Rate, Genome_Altered, and Diagnosis_Age. In summary, for the RF classification method, the most important target attributes for the classification of EAC, ESCC, and SAC were Country and Genome Altered.

4. Discussion

In a previous study, the molecular characteristics of the histological subtypes of EAC and ESCC were different across all detection platforms.[5] The similarity between ESCC and HNSCC is greater than that between ESCC and EAC. Therefore, in classification by machine learning, ESCC and EAC are the easiest to distinguish. Previous studies found that the similarity between EAC and SAC is higher than that between EAC and ESCC. However, according to the

model in which EAC is derived from Barrett's esophagus rather than the stomach, Barrett's esophagus and EAC may be derived from proximal gastric cells or the embryonic residual cell population of the gastroesophageal junction, and EAC is considered to be separate from SAC.[10,11] However, since the molecular characteristics of EAC and chromosomal instability (CIN) gastric cancer are similar, we may not be able to completely distinguish them from CIN gastric cancer by relying solely on molecular analysis. Therefore, it is necessary to analyze both the clinical and pathological characteristics and the molecular characteristics and to use a machine learning method to classify them.

5. Conclusions

The information collected by sensors can be analyzed by machine learning. Machine learning has the advantages of high classification accuracy, fast calculation, and strong learning ability. By analyzing the data, reliable conclusions can be obtained. The combination of sensors and machine learning has certain practicability in many fields; the more data collected by the sensors in the future, the more machine learning algorithms can be optimized, thereby improving the accuracy of the analysis results.

In this paper, the decision tree, bagging, AdaBoost, and RF machine learning classification algorithms were used to classify three types of cancer, and the importance of variables for the different classification models was analyzed. The classification results showed that all models were the least effective and had the lowest precision in the classification of ESCC. Country and Anatomic_Site were the most important variables for the different classification methods, indicating that they are very important for differentiating between cancer types. To verify the classification effects of the different classification models, confusion matrices were used to evaluate the models, and the classification results of the RF model were found to be the best. In this paper, we studied an unbalanced data set, for which the RF model can provide an effective method to balance errors in the data set. If a large part of a feature is lost, the RF algorithm can still maintain accuracy. The RF algorithm has strong anti-interference ability and anti-overfitting ability, resulting in its high classification performance in this study. However, in this paper, we only classified EAC, ESCC, and SAC, and no studies have yet focused on differentiating between other cancer types using machine learning classification algorithms. Because ESCC is not easy to distinguish from other squamous cell carcinomas, such as those of the head and neck, which have high similarity to ESCC, through general medical means, further studies are necessary. Classification based on the RF model can effectively improve the differentiation between EAC and SAC, enabling cancer patients to receive more accurate treatments and have an improved prognosis.

Acknowledgments

The authors sincerely thank the participants for their help and willingness to participate in this study. They also thank the reviewers for their helpful comments. This study was supported by the National Natural Science Foundation of China (Nos. 11601083 and 12001105), the

Program for Probability and Statistics: Theory and Application (IRTL1704), and Innovative Research Team in Science and Technology in Fujian Province University (IRTSTFJ). This work was also supported by projects under Nos. MOST 108-2221-E-390-005 and MOST 109-2221-E-390-023.

References

1 J. R. Siewert and K. Ott: Semin. Radiat. Oncol. **17** (2007) 38. https://doi.org/10.1016/j.semradonc.2006.09.007
2 R. D. Angelis, M. Sant, M. P. Coleman, S. Francisci, P. Baili, D. Pierannunzio, A. Trama, O. Visser, H. Brenner, and E. Ardanaz: Lancet Oncol. **15** (2014) 23. https://doi.org/10.1016/S1470-2045(13)70546-1
3 K. D. Miller, R. L. Siegel, C. C. Lin, A. B. Mariotto, J. L. Kramer, J. H. Rowl, K. D. Stein, R. Alteri, and A. J. Dvm: CA. Cancer J. Clin. **66** (2016) 271. https://doi.org/10.3322/caac.21349
4 L. A. Torre: CA: Cancer J. Clin. **65** (2015) 87. https://doi.org/10.3322/caac.21262
5 The Cancer Genome Atlas Research Network: Nature **541** (2017) 169. https://doi.org/10.1038/nature20805
6 S. S. Devesa and J. F. Fraumeni: JNCI **91** (1999) 747. https://doi.org/10.1093/jnci/91.9.747
7 T. W. Rice, E. H. Blackstone, and V. W. Rusch: Ann. Surg. Oncol. **17** (2010) 1721. https://doi.org/10.1245/s10434-010-1024-1
8 Y. S. Suh, D. S. Han, S. H. Kong, H. J. Lee, Y. T. Kim, W. H. Kim, K. U. Lee, and H. K. Yang: Ann. Surg. **255** (2012) 908. https://doi.org/10.1097/SLA.0b013e31824beb95
9 J. M. Leers, S. R. DeMeester, N. Chan, S. Ayazi, A. Oezcelik, E. Abate, F. Banki, J. C. Lipham, J. A. Hagen, and T. R. Demeester: J. Thorac. Cardiovasc. Surg. **138** (2009) 594. https://doi.org/10.1016/j.jtcvs.2009.05.039
10 X. Wang, H. Ouyang, Y. Yamamoto, P. Kumar, T. Wei, R. Dagher, M. Vincent, X. Lu, A. Bellizzi, and K. Ho: Cell **145** (2011) 1023. https://doi.org/10.1016/j.cell.2011.05.026
11 M. Quante, G. Bhagat, J. Abrams, F. Marache, P. Good, M. Lee, Y. Lee, R. Friedman, S. Asfaha, and Z. Dubeykovskaya: Cancer Cell **21** (2012) 36. https://doi.org/10.1016/j.ccr.2011.12.004

About the Authors

Xiaoping Chen received his B.S. and M.S. degrees from Fujian Normal University, Fuzhou, in 2004 and 2007, respectively, and his Ph.D. degree from Shanghai University of Finance and Economics, Shanghai, in 2015. Since 2007, he has been working at Fujian Normal University. He was a research assistant in the Department of Management Sciences at the City University of Hong Kong in 2014. From 2017 to 2018, he was a visiting scholar in the Department of Statistics and Actuarial Science at the University of Hong Kong. His research interests are in survival analysis, nonparametric statistics, biomedical statistics, and economic statistics. (xpchen@fjnu.edu.cn)

Lihui Zheng received her B.S. degree from the Civil Aviation University of China, Tianjin, in 2015. She is a graduate student at Fujian Normal University, Fuzhou. Her research interests are in survival analysis, nonparametric statistics, and biomedical statistics. (lhzheng_11@163.com)

Jianqi Yao received her B.S. degree from Jiangxi University of Science and Technology, Ganzhou, in 2017, and her M.S. degree from Fujian Normal University, Fuzhou, in 2020. Her research interests are in survival analysis, economic statistics, and biomedical statistics. (jianqiyao2019@163.com)

Cheng-Fu Yang earned his B.S., M.S., and Ph.D. degrees in 1986, 1988, and 1993, respectively, from the Department of Electrical Engineering of Cheng Kung University. In 2004, he became a professor of chemical and materials engineering at the National University of Kaohsiung (NUK). He obtained the Outstanding Contribution Award of the Chinese Ceramic Society in 2009. In 2010, he was the first (and only) person to become a distinguished professor of NUK. He became a fellow of the Taiwanese Institute of Knowledge Innovation (TIKI) in 2014 and a fellow of the Institution of Engineering and Technology (IET) in 2015. He became a Mingjiang scholar and invited chair professor of Jimei University, Xiamen, Fujian, China, and an honorary chair professor of Chaoyang University of Technology in 2020. (cfyang@nuk.edu.tw)