Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 20(10): 18368–18385.
DOI: 10.3934/mbe.2023816
Received: 25 June 2023
Revised: 11 September 2023
Accepted: 18 September 2023
Published: 25 September 2023

Research article

A screened predictive model for esophageal squamous cell carcinoma based on salivary flora data

Yunxiang Meng[1], Qihong Duan[1], Kai Jiao[2] and Jiang Xue[1,*]

[1] School of Mathematics and Statistics, Xi'an JiaoTong University, Xi'an, China.

[2] Department of Oral Mucosal Diseases, State Key Laboratory of Military Stomatology & National Clinical Research Center for Oral Diseases & Shaanxi Key Laboratory of Stomatology, School of Stomatology, The Fourth Military Medical University, Xi'an, China.

* **Correspondence:** Email: x.jiang@xjtu.edu.cn; Tel: +862982665732.

Abstract: Esophageal squamous cell carcinoma (ESCC) is a malignant tumor of the digestive system in the esophageal squamous epithelium. Many studies have linked esophageal cancer (EC) to the imbalance of oral microecology. In this work, different machine learning (ML) models including Random Forest (RF), Gaussian mixture model (GMM), K-nearest neighbor (KNN), logistic regression (LR), support vector machine (SVM) and extreme gradient boosting (XGBoost) based on Genetic Algorithm (GA) optimization was developed to predict the relationship between salivary flora and ESCC by combining the relative abundance data of *Bacteroides*, *Firmicutes*, *Proteobacteria*, *Fusobacteria* and *Actinobacteria* in the saliva of patients with ESCC and healthy control. The results showed that the XGBoost model without parameter optimization performed best on the entire dataset for ESCC diagnosis by cross-validation (Accuracy = 73.50%). Accuracy and the other evaluation indicators, including Precision, Recall, F1-score and the area under curve (AUC) of the receiver operating characteristic (ROC), revealed XGBoost optimized by the GA (GA-XGBoost) achieved the best outcome on the testing set (Accuracy = 89.88%, Precision = 89.43%, Recall = 90.75%, F1-score = 90.09%, AUC = 0.97). The predictive ability of GA-XGBoost was validated in phylum-level salivary microbiota data from ESCC patients and controls in an external cohort. The results obtained in this validation (Accuracy = 70.60%, Precision = 46.00%, Recall = 90.55%, F1-score = 61.01%) illustrate the reliability of the predictive performance of the model. The feature importance rankings obtained by XGBoost indicate that *Bacteroides* and *Actinobacteria* are the two most important factors in predicting ESCC. Based on these results, GA-XGBoost can predict and diagnose ESCC according to the relative abundance of salivary flora, providing an effective tool for the non-invasive prediction of esophageal malignancies.

Keywords: esophageal squamous cell carcinoma; oral microecology; machine learning; saliva; flora; XGBoost

1. Introduction

ESCC is a malignant epithelial tumor with squamous cell differentiation and is a type of EC. ESCC mainly occurs in Asian countries such as China [1], the incidence and the mortality rates of ESCC are among the highest for malignant tumors [2]. In China, more than 90% of EC cases are ESCC [3], accounting for approximately half of all ESCC cases worldwide [4]. With the improvement of medical treatment in recent years, the morbidity rate and death rate of EC have decreased. However, due to the lack of early disease screening methods and the fact that patients with EC do not show obvious symptoms until the middle to late stages of the disease, the survival rate of patients within 5 years is only 13–18% [5]. Therefore, it is essential to prevent the disease early and to develop a screening tool for ESCC.

Many factors contribute to the development of ESCC [6], factors such as alcohol consumption, smoking, diet and nutritional deficiencies are the main causes of ESCC [7]. Moreover, poor oral health is also a risk element for ESCC lesions, and this finding has been reported in China, Japan, India and Latin America [8–10]. In addition, there is an association between oral microbiota and systemic diseases [11–13]. Since the composition of the oral microbiota is similar to that of the esophageal microbiota, alterations in the oral microbiome may lead to alterations in the esophageal microbiome [14], which in turn may cause EC [15, 16]. Some researchers have indicated that the oral microbiota of patients with ESCC differs from that of the controls [11, 17–24]. Wang et al. and Li et al. concluded that the relative abundance of *Bacteroides* on ESCC was significantly lower than that of the control [18, 22]. Furthermore, a study from the United States showed that an increase in *Prevotella oral taxon 306* belonging to *Bacteroides* was associated with a low risk of ESCC [24]. For *Proteobacteria*, a study by Chen et al. showed a reduction in the abundance of *Neisseria* in ESCC patients [17]. Several other studies on the relationship between oral flora and ESCC risk also showed a significant decline in *Proteobacteria* among ESCC patients compared to controls [18, 22, 25, 26]. Zhao et al. found that *Veillonella*, which belongs to *Firmicutes*, was enriched in EC patients [21]. Moreover, the finding in Lu's study [27] that the risk of suffering from ESCC was associated with an increased relative abundance of *Fusobacteria* and *Actinobacteria* in saliva samples. Therefore, it is necessary to explore the association between the relative abundance of oral flora and ESCC.

Machine learning is the study of how computers can simulate the implementation of human learning behaviors and thus acquire new knowledge and skills from mountains of data. Machine learning (ML) can help intelligent systems make a variety of judgments based on data. Common ML-supervised learning algorithms include RF, GMM, KNN, LR, SVM and XGBoost. These algorithms are commonly used for classification problems and regression problems. Nowadays, machine learning is widely used in biomedicine as a mainstream data processing method, and this discipline has enabled risk prediction for a variety of diseases such as cancer [28]. Furthermore, ML models were utilized as early as 2005 for survival prediction in patients with esophageal cancer. Mofidi et al. used clinical information, body mass index and pathology datasets of patients with EC to develop models to predict survival in patients with EC and surgical resection of the esophagogastric junction [29]. Hayashida et al. trained models to predict the efficacy of chemotherapy and radiation therapy in patients with esophageal cancer from protein profile data obtained from serum samples [30]. However, the development of predictive models for EC is lagging compared to survival models, it is still an open question of how to use oral flora data to build ML models for prediction and early screening of ESCC.

Saliva samples are easy to collect and widely used to detect and classify microorganisms [31]. Meanwhile, considering that the composition of human flora is susceptible to change due to multiple external factors and human activities, such as residence location and dietary habits [32], data from more sources would help to explore deeply the relationship between oral flora and ESCC. In this study, RF, GMM, KNN, LR, SVM and genetic algorithm-based optimization XGBoost (GA-XGBoost), will be built to distinguish healthy controls from ESCC patients based on the relative abundance data of phylum level bacteria in saliva such as *Bacteroides*, *Firmicutes*, *Proteobacteria*, *Fusobacteria* and *Actinobacteria*, which associated with ESCC [17–19, 22, 25–27, 33, 34]. The ML models also provide early warning for patients at risk of ESCC, allowing sufficient time for intervention and treatment to improve survival.

2. Materials and methods

2.1. Data collection and processing

The data used in this study are from 8 articles published between 2015 and 2021 that explored the relationship between salivary microbiota and ESCC [17–19, 22, 25–27, 33, 34]. This study was conducted under approved guidelines as all included literature is publicly available.

To analyze the relationship between ESCC and salivary microflora, data based on the mean and standard deviation of relative abundance of *Bacteroides*, *Firmicutes*, *Proteobacteria*, *Fusobacteria* and *Actinobacteria* reported in each article for ESCC patients and healthy controls were generated. For the relative abundance data presented in the form of graphs, values were extracted through the Web Plot Digitizer (version 4.6) and the data were represented as box plots, with the mean and standard deviation calculated from the quartiles of the box plots [35]. If the relative abundance data of a particular bacterium were missing in one article, the average value of that bacterium in the rest of the articles was used to fill in the data. For example, the relative abundance data of *Fusobacteria* are missing in the study of Wan et al., and the average of the relative abundance of *Fusobacteria* in the remaining 7 articles is calculated and that value is used as the relative abundance measure of *Fusobacteria* in the study of Wan et al. [35].

A total of 8000 sets of data from 8 sources [17–19, 22, 25–27, 33, 34] were generated, which included the relative abundance data of *Bacteroides*, *Firmicutes*, *Proteobacteria*, *Fusobacteria* and *Actinobacteria* in 4000 ESCC patients and 4000 healthy controls (Table 1).

Table 1. Descriptive statistics of each bacterium in all data.

Relative abundance	ESCC (N = 500)	Controls (N = 500)	Total (N = 1000)
Actinobacteria(%)	2.72 ± 1.09	1.97 ± 0.61	2.35 ± 0.96
Bacteroidetes(%)	29.26 ± 5.63	39.28 ± 4.72	34.27 ± 7.22
Firmicutes(%)	19.21 ± 2.60	15.29 ± 1.95	17.25 ± 3.02
Fusobacteria(%)	13.50 ± 4.09	7.79 ± 2.27	10.65 ± 4.37
Proteobacteria(%)	29.20 ± 11.18	30.33 ± 7.64	29.76 ± 9.59

2.2. Machine learning methods

The analyses were conducted with Python programming language (version 3.9) and the "scikit-learn" library. Using a series of ML methods: RF, GMM, KNN, LR, SVM and XGBoost, a binary classification model was constructed according to oral salivary flora (0: non-ESCC patients, 1: ESCC patients) to predict ESCC [36].

2.2.1. K-fold cross-validation

K-Fold cross-validation [37] is a common method for parameter optimization and model selection in ML. It randomly divides the dataset into k disjoint subsets of the same size, selects one subset each time as the testing set and the remaining $k-1$ subsets as the training set, then repeats k times to obtain k models. Calculate the test error of each model and use the average of k test errors as the performance metric of the model. The performance index of the model E can be calculated as follows:

$$E = \frac{1}{k}\sum_{i=1}^{k} E_i, \tag{2.1}$$

where E_i represents the test error of the i-th model.

2.2.2. Random forest (RF)

Random forest [38] is a classifier that uses multiple decision trees to train and predict samples. In this model, a training set is formed by sampling N times from N samples in the way of resampling, and the unsampled samples are used to predict and evaluate its error. The probability P that each sample is not drawn is calculated as follows:

$$P = (1 - \frac{1}{N})^N. \tag{2.2}$$

For each node of the decision tree, m features are randomly selected and their optimal split is calculated based on these features by the Gini index. The Gini index $G(M)$ can be calculated as follows:

$$G(M) = \sum_{k=1}^{K} p_k(1-p_k) = 1 - \sum_{k=1}^{K} {p_k}^2, \tag{2.3}$$

here M is the training subset and p_k is the probability of a sample belonging to the k-th category. RF can handle both classification and numerical features, and the model is also relatively resistant to over-fitting, making false predictions only when more than half of the base classifiers are in error. However, due to its inherent complexity, it requires more time to train them than other similar algorithms.

2.2.3. Gaussian mixture models (GMM)

The GMM [39] is a widely used clustering algorithm that uses a Gaussian distribution as a parametric model. GMM can be regarded as a model consisting of a combination of K Gaussian models. The data $x = (x_1, x_2, \ldots, x_n)$ is generated from several standard Gaussian distributions, then the probability density $P(x)$ of the Gaussian distribution is computed as follows:

$$P(x|\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\exp(-\frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{2}), \tag{2.4}$$

where μ is the expectation of data and Σ is the covariance. The Gaussian mixture model is defined by the formula:

$$P(x) = \sum_{i=1}^{K}\pi_i P(x|\mu_i,\Sigma_i), \tag{2.5}$$

where the mean μ_i and variance Σ_i are the parameters to be estimated, π_i is the probability of generating data for each sub-model. GMM can be used not only for clustering but also for probability density estimation. However, estimating covariance can be difficult when there are not enough points for each mixture model.

2.2.4. K-nearest neighbor (KNN)

KNN [40] is a basic classification and regression method. In the classification problem, each data in the sample set containing n samples $x_i = (x_{i1}, x_{i2}, \ldots, x_{in})$ have a corresponding relationship with their classification. After inputting a new sample without labels, each feature of the new data is compared with the corresponding feature of the data in the sample set, the algorithm extracts the classification labels of the k most similar data of the sample and finally selects the classification label with the most occurrences among the k most similar data as the label of the new sample. The model generally selects the nearest neighbors by calculating the distance from the test data to each of the training data. The distance d is commonly calculated by the formula:

$$d = \sqrt{\sum_{k=1}^{n}(x_{1k} - x_{2k})^2}. \tag{2.6}$$

KNN is a mature theory with simple ideas, and can be used for both classification and regression. However, the prediction bias is relatively high when the samples are not balanced.

2.2.5. Logistic regression (LR)

LR [41] is a ML method used to solve binary classification problems and is also a generalized linear model. It maps the input set of data $x = (x_1, x_2, \ldots, x_n)$ with n features to a number between 0 and 1 by the sigmoid function. When the function value is greater than 0.5, the classification result is judged as 1 (ESCC patients), otherwise, the classification result is 0 (non-ESCC patients). The predicted result $y(x)$ is computed as followed:

$$y(x) = \frac{1}{1 + e^{-(w^T x+b)}}, \tag{2.7}$$

here T means transpose, w and b are parameters to be evaluated. LR is computationally small and has a low storage footprint, which can be used in big data scenarios and is especially suitable for classification problems. However, it is essentially a linear classifier, so it does not handle the case of correlation between features well.

2.2.6. Support vector machine (SVM)

SVM is a kind of ML model that classifies data in a supervised learning manner, which uses a hinge loss function to compute empirical risk and adds a regularization term to the solution system to optimize structural risk. SVM is one of the common kernel learning methods and the basic concept of SVM is to find the separating hyperplane that can correctly partition the training data set with maximum geometric separation. SVM has high accuracy and provides a good theoretical guarantee to avoid over-fitting. However, the model is memory-consuming and difficult to interpret, and the efficiency is not great when there are many observed samples.

2.2.7. Extreme Gradient Boosting (XGBoost)

XGBoost is an efficient gradient-boosting decision tree algorithm that is composed of multiple classification regression trees, so the algorithm can process problems such as classification regression. As a forward addition model, the core of the XGBoost algorithm is to integrate multiple weak learners into one strong learner by certain methods. The model uses multiple trees to make decisions together. Each tree is the result of the target before and all the trees of the difference in value between predicted results, and the final result is obtained by combining the results of all the trees. Given a dataset containing n samples, there is a corresponding true result $\hat{y}_i$ for any sample $x_i(i\epsilon(1, 2, \ldots, n))$. The predicted probability $\hat{y}_i$ can be calculated as follows:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i),\ f_k \in \mathcal{F}, \tag{2.8}$$

where K denotes the number of decision trees, f_k is the prediction score of a single decision tree and $\mathcal{F}$ is the space composed of all trees. The objective function obj of XGBoost can be expressed as follows:

$$obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{2.9}$$

here l is the loss function and Ω is the penalty term. The expression for the penalty term is:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda||\omega_i||^2, \tag{2.10}$$

where T indicates the number of leaves of a single tree, ω_i is the score of the leaf node i, γ and λ are hyperparameters. XGBoost has very high accuracy as well as greater flexibility, but the model still requires traversal of the dataset during node splitting.

2.2.8. Genetic algorithms (GA)

The GA is a computational model of the biological evolutionary process based on the Darwinian biological evolutionary theory of natural selection and genetic mechanism, which searches for the optimal solution by simulating the natural evolutionary process. The algorithm targets all individuals in the population and uses randomization techniques to perform an efficient search of a parameter space being encoded. In this case, selection, crossover and mutation constitute the genetic operations of the GA. After the initial population is generated, the evolution of generation by generation produces

increasingly better approximate solutions according to the principles of survival of the fittest. In each generation, individuals are selected for their fitness in the problem domain, and new populations are generated by combinatorial crossover and mutation with the help of natural genetic operators.

2.2.9. XGBoost based on genetic algorithm optimization (GA-XGBoost)

Genetic algorithms are often used for model parameter tuning, which can make the model have better fitting or prediction ability. The most important step in the XGBoost model is model tuning, so the genetic algorithm can be combined with XGBoost, using the genetic algorithm to adjust the parameters of the XGBoost model, to obtain the best model.

2.3. Model performance evaluation

For a binary classification model, there are four classification results after the classifier divides the instances into positive and negative classes:

True Positive (TP): True positive example, an instance is positive and is also determined to be positive.

True Negative (TN): The correct negative case, where an instance is a false category and is also judged to be a false category.

False Negative (FN): False negative example, an instance is a positive category but is judged as a false category.

False Positive (FP): Wrong positive example, which is a false category but is judged to be a positive category.

Accuracy is the degree to which a thing is expressed or described correctly. The formula for calculating the Accuracy is:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}. \tag{2.11}$$

Precision is the accuracy of the prediction of positive cases, which is the proportion of the sample predicted as positive cases that are predicted correctly, and it is used to assess the accuracy of predicting positive cases. The formula for calculating the Precision is:

$$Precision = \frac{TP + TN}{TP + FP}. \tag{2.12}$$

Recall is the proportion of positive cases that are predicted correctly to the total sample of actual positive cases in the sample of actual positive cases. The formula for calculating the Recall is:

$$Recall = \frac{TP}{TP + FN}. \tag{2.13}$$

Precision and Recall are a pair of contradictory metrics. In general, Recall values seem to be low when Precision is high, while Recall values tend to be high when Precision is low. In order to be able to consider these two metrics together, F1-score is proposed, which is a weighted summed average of Precision and Recall, calculated by the formula:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \tag{2.14}$$

The ROC curve describes the variation of classifier performance with the classifier threshold [42], and each point on the curve reflects the perceptibility to the same signal stimulus. This curve is a curve plotted with the true positive rate as the vertical coordinate and the false positive rate as the horizontal coordinate. When comparing the performance of two or more models, it is possible to visually identify the strengths and weaknesses of the models. The ROC curve of the model closer to the top left corner means that the model is more accurate. The true positive rate TDR was determined as follows:

$$TDR = \frac{TP}{TP + FN}. \tag{2.15}$$

The expression for the false positive rate FDR is:

$$FDR = \frac{FP}{TN + FP}. \tag{2.16}$$

The AUC is considered as the area under the ROC curve, the larger area under the curve, the better the performance of the corresponding model. AUC values range from 0 to 1 and are generally between 0.5 and 1 [42].

2.4. Experimental design

The dataset (N = 4000 ESCC patients and N = 4000 healthy controls) compares six machine learning models (RF, GNN, KNN, LR, SVM and XGBoost) with default parameters in the "scikit-learn" library for Python utilizing a five-fold cross-validation method: Randomly divide the dataset into five groups, four for training the model and the remaining group for testing. This process is repeated five times so that each group is used as a testing set once. The average performance of all five groups is considered the final performance of the model. Repeat this process five times for each ML algorithm and show the mean value of each model after five training sessions. The model with the highest accuracy (XGBoost) was selected as the final prediction model.

This dataset was divided into a training set (80%) and a testing set (20%). The parameters of the model with the highest accuracy are optimized on the training set using GA, and the performance metrics (Accuracy, Precision, recall, F1-score, ROC and AUC) of the parameter-optimized model are evaluated on the testing set.

2.5. External verification

Since oral flora presents some variation due to multiple factors [32], this work verified the ability of the model by generating 1000 sets of data (N = 500 ESCC patients and N = 500 controls) with means and standard deviations of the relative abundance of *Bacteroides*, *Firmicutes*, *Proteobacteria*, *Fusobacteria* and *Actinobacteria* which reported in Wei [33]. Table 2 illustrates the data characteristics of the external validation queue.

Table 2. Descriptive statistics of the external validation data.

Relative abundance	ESCC(N = 500)	Controls(N = 500)	Total(N = 1000)
Actinobacteria(%)	2.72 ± 1.09	1.97 ± 0.61	2.35 ± 0.96
Bacteroidetes(%)	29.26 ± 5.63	39.28 ± 4.72	34.27 ± 7.22
Firmicutes(%)	19.21 ± 2.60	15.29 ± 1.95	17.25 ± 3.02
Fusobacteria(%)	13.50 ± 4.09	7.79 ± 2.27	10.65 ± 4.37
Proteobacteria(%)	29.20 ± 11.18	30.33 ± 7.64	29.76 ± 9.59

3. Results

3.1. Statistical analysis of salivary flora abundance data

We included data on the relative abundance of salivary flora from 4000 ESCC patients versus 4000 healthy controls, with 6400 data sets in the training set and 1600 data sets in the testing set. In the training set, the number of ESCC patients (N = 3177) was less than the number of controls (N = 3223). Conversely, in the testing set, there were more data from the ESCC group (N = 823) than from the healthy controls (N = 777). Compared to the controls, the oral saliva of the ESCC patients showed a lower number of floras belonging to *Proteobacteria*, *Fusobacteria* and *Actinobacteria*. Meanwhile, the elevated relative abundance of *Bacteroides* and *Firmicutes*, may be associated with ESCC (Table 3).

Table 3. Descriptive statistics for training and testing set.

Variable	Training(N = 6400)		Testing(N = 1600)	
	ESCC(N = 3177)	Controls(N = 3223)	ESCC(N = 823)	Controls(N = 777)
Actinobacteria(%)	26.45 ± 13.95	23.99 ± 13.06	26.30 ± 13.69	23.51 ± 12.80
Bacteroidetes(%)	34.70 ± 11.65	30.95 ± 11.46	34.40 ± 11.26	30.89 ± 11.69
Firmicutes(%)	21.23 ± 14.99	25.07 ± 12.63	21.22 ± 15.18	25.67 ± 12.65
Fusobacteria(%)	5.85 ± 3.06	7.38 ± 3.88	5.92 ± 3.10	7.24 ± 3.75
Proteobacteria(%)	6.42 ± 4.13	8.28 ± 7.45	6.57 ± 4.26	8.09 ± 7.14

3.2. Model construction and evaluation

The data were cross-validated in five folds with the default hyperparameters of ML models and compared the predictive power of six models (RF, GMM, KNN, LR, SVM and XGBoost). According to the average accuracy, XGBoost was most effective in predicting the risk of a sample suffering from ESCC and correctly predicted 73.50% of the samples. The performance of the RF is the second highest after the performance of XGBoost, with an accuracy of 73.34%. The accuracy of the other models is lower compared to the accuracy of the first two models, including KNN (Accuracy = 70.28%), SVM (Accuracy = 68.45%), GMM (Accuracy = 58.16%) and LR (Accuracy = 57.36%) have relatively poor performance (Table 4).

The parameters in the XGBoost model were optimized in the training set with GA, and the parameters n_estimators were cross-validated to obtain the best value of 82. The remaining parameters were iterated through 50 iterations of GA to obtain the parameter combinations that optimize the performance of the XGBoost model. Table 5 shows the parameter combinations and model capabilities of the GA-XGBoost

model and other models (XGBoost, RF, GMM, KNN, LR and SVM) used on the testing set.

Table 4. Default parameters and performance for each machine learning algorithm.

Models	Parameters	Accuracy
XGBoost	n_estimators = 100, colsample_bytree = 1, gamma = 0, learning_rate = 0.3, max_depth = 6, min_child_weight = 1, alpha = 0, lambda = 1, subsample = 1	73.50%
RF	n_estimators = 100, min_samples_split = 2, min_samples_leaf = 1, min_impurity_split = 0	73.34%
KNN	n_neighbors = 5, leaf_size = 30, p = 2	70.28%
SVM	C = 1, tol = 1 $\times 10^{-4}$, max_iter = –1	68.45%
GMM	None	58.16%
LR	Penalty = l2, tol = 1 $\times 10^{-4}$, c = 1, max_iter = 10	57.36%

Table 5. Performance of GA-XGBoost and other machine learning models.

Models	Parameters	Accuracy	Precision	Recall	F1-score	AUC
GA-XGBoost	n_estimators = 82, colsample_bytree = 0.93, gamma = 0.25, learning_rate = 0.32, max_depth = 5, min_child_weigh = 2.17, alpha = 1, lambda = 0.47, subsample = 1	89.88%	89.43%	90.75%	90.09%	0.97
XGBoost	n_estimators=100, colsample_bytree=1, gamma=0, learning_rate=0.3, max_depth=6, min_child_weight=1, alpha=0, lambda=1, subsample=1	88.19%	87.85%	89.04%	88.44%	0.97
RF	n_estimators=100, min_samples_split=2, min_samples_leaf=1, min_impurity_split=0	89.56%	89.06%	90.49%	89.77%	0.97
KNN	n_neighbors=5, leaf_size=30, p=2	85.44%	82.87%	88.11%	85.41%	0.94
SVM	C=1, tol=1e-4, max_iter=-1	79.44%	82.14%	78.33%	80.19%	0.89
GMM	None	60.12%	70.35%	59.51%	64.48%	0.67
LR	Penalty=l2, tol=1e-4, c=1, max_iter=10	59.94%	58.81%	61.58%	60.16%	0.62

The accuracy of GMM, LR and SVM in predicting whether a sample suffered from ESCC through the relative abundance data of salivary flora was less than 80%. Compared to KNN (Accuracy = 85.44%), XGBoost (Accuracy = 88.19%) and RF (Accuracy = 89.56%), GA-XGBoost (Accuracy = 89.88%) has a better accuracy in predicting ESCC. Precision and Recall, as a pair of contradictory evaluation metrics, both show the best performance in the GA-XGBoost model

(Precision = 89.43%, Recall = 90.75%), followed closely by RF (Precision = 89.06%, Recall = 90.49%), XGBoost (Precision = 87.85%, Recall = 89.04%) and KNN(Precision = 82.87%, Recall = 88.11%). The Precision of SVM is not low, but the Recall is below 80%, approximately 78.33%. F1-score is an evaluation metric that balances both Precision and Recall. So for F1-score, GA-XGBoost is also the most effective model among the seven models (F1-score = 90.09%). However, GMM and LR performed negatively in the four evaluation metrics of accuracy, Precision, Recall and F1-score. Except GMM, which had a Precision of about 70.35%, none of the other metrics exceeded 70%. The performance of these models was further evaluated using the AUC of ROC, and GA-XGBoost had similar AUC values to RF as well as XGBoost, both at 0.97, followed by KNN (AUC = 0.94), SVM (AUC = 0.89), GMM (AUC = 0.67) and LR (AUC = 0.62).

Based on a comprehensive evaluation of five evaluation metrics, the GA-XGBoost model showed the best performance in predicting ESCC. Therefore, the XGBoost model optimized based on GA is more effective in exploring the relationship between oral flora and ESCC than other traditional ML models.

3.3. External validation of XGBoost optimized by genetic algorithm

Since GA-XGBoost has the best performance among all models, in order to verify the generalizability of the model, the XGBoost model with the same parameters are used for validation in an external cohort. The external cohort provided 1000 samples to validate the model. GA-XGBoost revealed a better prediction performance in the external cohort. On the one hand, based on the results of the confusion matrix (Figure 1), the model achieves 70.60% Accuracy, 46.00% Precision, 90.55% Recall and 61.01% F1-score (Table 6). On the other hand, the AUC value of the model (AUC = 0.88) achieved similarly positive results (Figure 2).

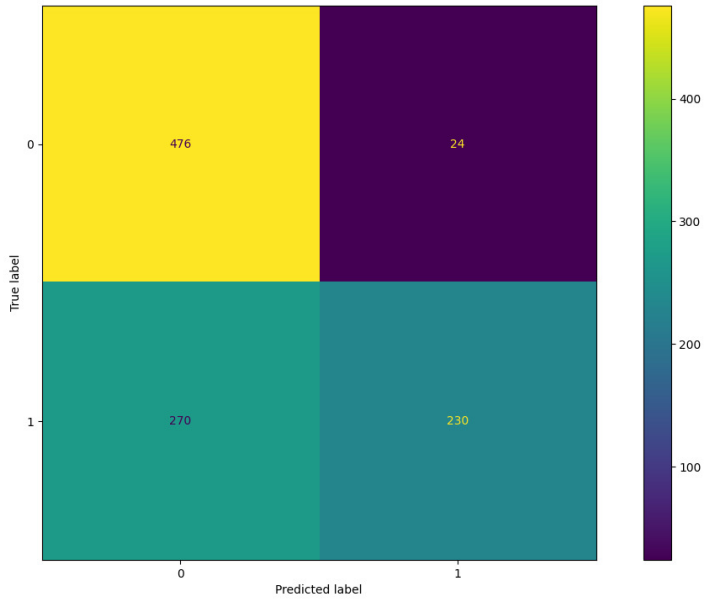

Figure 1. Confusion matrix of external validation.

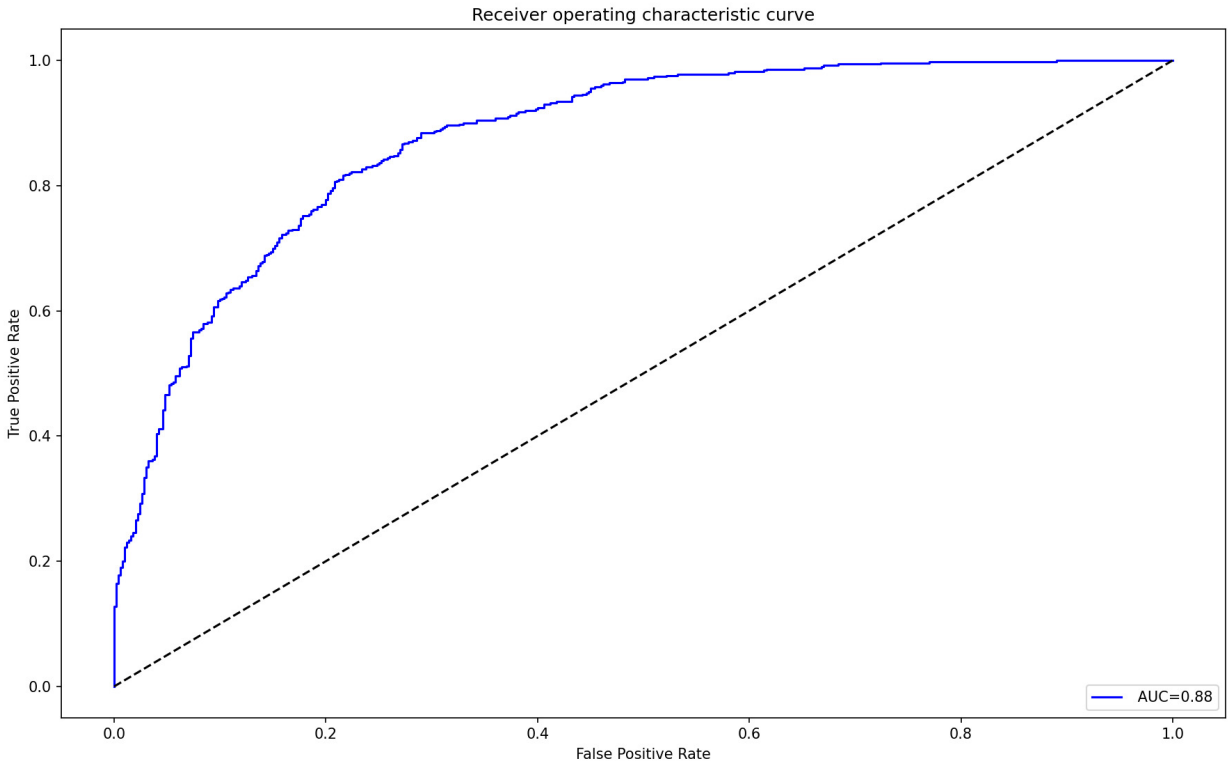

Figure 2. Receiver operating characteristic of external validation.

Table 6. Model performance in external cohort.

Model	Parameters	Accuary	Precision	Recall	F1-score
GA-XGBoost	n_estimators = 82, colsample_bytree = 0.93, gamma = 0.25, learning_rate = 0.32, max_depth = 5, min_child_weight = 2.17, alpha = 1, lambda = 0.47, subsample = 1	70.60%	46.00%	90.55%	61.01%

3.4. Importance ranking of predictor variables

The ranking of feature importance calculated from the XGBoost model after optimizing the parameters showed that Bacteroides was the most important predictor. During the construction of the decision tree in XGBoost, the number of key decisions made conditional on *Bacteroides* was 319, accounting for 24.67% of the total number of decisions. The following predictor was *Actinobacteria* (21.73%). Although the importance index values of *Firmicutes*, *Proteobacteria* and *Fusobacteria* were less than that of *Bacteroides* and *Actinobacteria*, they all ranged from 15 to 20% (Figure 3).

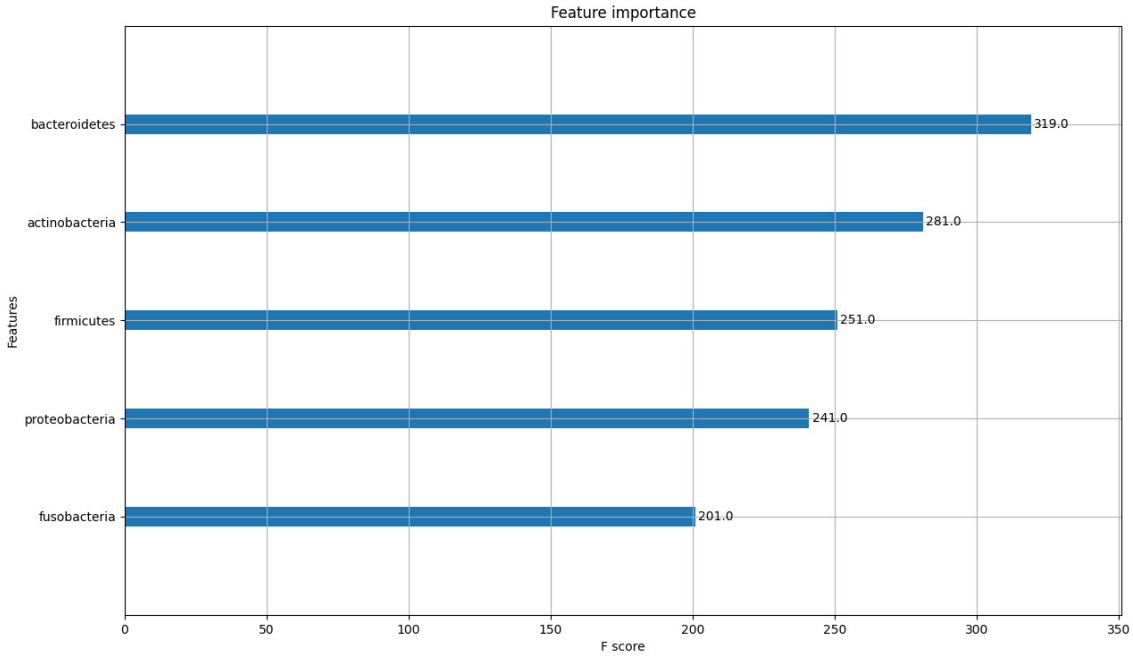

Figure 3. Feature importance ranking of GA-XGBoost.

4. Discussion

Early diagnosis and screening of cancer is an important method to improve the cure rate of cancer, prolong the survival time of patients and reduce the financial and emotional burden of patients. In the past decade, artificial intelligence technologies and ML approaches have been widely used in the early cancer screening, clinical aid diagnosis and mechanistic research. However, EC is not easily detected in the early stage, and ML research in the field of EC is lagging behind compared with other cancers, so building ML models is crucial in EC research. In this study, six ML models were applied to explore the relationship between the relative abundance of oral salivary flora and ESCC. The results indicated that the GA-XGBoost model observed higher performance in predicting whether a sample was suffering from ESCC or not. This is because the XGBoost model, as an integrated algorithm, has an advantage over other models by adding additional canonical terms to avoid model overfitting in the construction of decision trees [43]. According to our knowledge, this is the first study to use the ML methods approach to model and predict the presence of ESCC by the relative abundance of flora at the salivary phylum level.

Previous studies have examined oral microbiota and periodontal characteristics as clinicopathological factors in patients with EC [20, 44]. Moghtadaei et al. constructed a risk prediction model based on common risk factors for EC such as age, residence, smoking history and oral health, which obtained an accuracy of nearly 90% and surpassed the traditional LR model [20]. Kawasaki et al. confirmed that the prevalence of *T. forsythia* and *S. anginosus* in dental plaque as well as the prevalence of *A. actinomycetemcomitans* in saliva and alcohol consumption habits were associated with a high risk of EC [20]. Then he constructed a multiple LR model with the relative abundance of *S. anginosus*d in saliva and sociological factors such as age, body mass index, smoking history and alcohol consumption

habits as variables, which had an AUC value of 0.82. However, these models are often suitable for large sample surveys and are not applicable to individual clinical diagnosis because factors such as patients' residence, smoking history and drinking habits change according to people's habits, making it impossible to obtain comprehensive and accurate data.

Saliva is not only easy to extract samples, but also effective in making predictions and judgments about whether an individual case has ESCC or not. Based on this property of saliva, a GA-XGBoost model with 89.88% Accuracy, 89.43% Precision, 90.75% Recall, 90.09% F1-score and AUC = 0.97 was constructed by using oral saliva flora in this study. Even though the performance of the GA-XGBoost model was slightly reduced in the external validation set, the prediction performance of the model still performed positively (Accuracy = 70.60%, Precision = 46.00%, Recall = 90.55%, F1-score = 61.01%, AUC = 0.88). It is noteworthy that the recall of the model in the external queue reaches 90.55%, but the precision is only 46.00%. Since recall and precision affect each other and are mutually constrained, a high value for both cannot generally be achieved in realistic situations. In the field of disease prediction, it is acceptable to misclassify a disease, but not to undiagnose a disease. Therefore, Recall's value should be more meaningful. These results show the ability of salivary flora and GA-XGBoost models to predict ESCC. Although salivary flora is influenced by a relatively large number of factors, the model is capable of effectively predicting the presence or absence of ESCC.

In this study, we considered the importance of five phylum-level bacteria: *Bacteroides*, *Firmicutes*, *Proteobacteria*, *Fusobacteria* and *Actinobacteria* for ESCC. Our model suggests that *Bacteroides* and *Actinobacteria* are the two most important influencing elements, which have a powerful impact on the prediction of ESCC. A study from China and a study from the United States reported that alterations in *Prevotella* belonging to *Bacteroides* may be a potential predictor of EC [21, 24]. Furthermore, Kawasaki et al. suggested that Actinobacteria could be used for the diagnostic evaluation of esophageal cancer even though it is a rarely detected periodontal disease bacterium [20]. Therefore, the relationship between *Bacteroides*, *Actinobacteria* and ESCC can be further explored.

All the data in this study were sourced from ESCC patients with healthy controls in China. Since the flora in the human body is susceptible to multiple factors, the capabilities and parameters of the model may be altered when the model is in application to other regions or environments. In addition, the present study only considered the association between oral microorganisms and ESCC. In order to further improve the performance of the model, sociological factors such as region, age, gender and drinking habits could be included in the model. Regardless of these limitations of the current study, modeling the relationship between oral flora and ESCC can provide further insight into the role played by oral flora in ESCC, leading to non-invasive methods for the prevention and diagnosis of ESCC.

5. Conclusions

In this study, following comparison and screening, a GA-XGBoost model with parameter optimization of the XGBoost algorithm using the global search capability of the GA was proposed to establish a disease prediction model for early screening of ESCC by the relative abundance of oral salivary flora. The parameters of the XGboost model were not easily determined to affect the performance of the model, and the accuracy of the algorithm was improved while avoiding the occurrence of overfitting. The experimental results show that the genetic algorithm improves the performance of the XGBoost model, and the GA-XGBoost model has better judgment than the traditional machine learning algorithm. The study provides a

new biomarker for the diagnosis of ESCC by the relative abundance of *Bacteroides* and *Actinobacteria*, which also provides a new research direction to investigate the association between oral flora and ESCC.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. M. Arnold, I. Soerjomataram, J. Ferlay, D. Forman, Global incidence of oesophageal cancer by histological subtype in 2012, *Gut*, **64** (2015), 381–387. https://doi.org/10.1136/gutjnl-2014-308124
2. E. J. Snider, G. Compres, D. E. Freedberg, H. Khiabanian, Y. R. Nobel, S. Stump, et al., Alterations to the Esophageal Microbiome Associated with Progression from Barrett's Esophagus to Esophageal Adenocarcinoma, *Cancer Epidem. Biomar. Prev.*, **28** (2019), 1687–1693. https://doi.org/10.1158/1055-9965.EPI-19-0008
3. J. Zhao, Y. T. He, R. S. Zheng, S. W. Zhang, W. Q. Chen, Analysis of esophageal cancer time trends in China, 1989–2008, *Asian Pac. J. Cancer Prev.*, **13** (2012), 4613–4617. https://doi.org/10.7314/apjcp.2012.13.9.4613
4. A. Q. Liu, E. Vogtmann, D. T. Shao, C. C. Abnet, H. Y. Dou, Y. Qin, et al., A Comparison of Biopsy and Mucosal Swab Specimens for Examining the Microbiota of Upper Gastrointestinal Carcinoma, *Cancer Epidem. Biomar. Prev.*, **28** (2019), 2030–2037. https://doi.org/10.1158/1055-9965.EPI-18-1210
5. R. Lozano, M. Naghavi, K. Foreman, S. Lim, K. Shibuya, V. Aboyans, et al., Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010, *Lancet*, **380** (2012), 2095–2128. https://doi.org/10.1016/S0140-6736(12)61728-0
6. C. C. Abnet, M. Arnold, W. Q. Wei, Epidemiology of esophageal squamous cell carcinoma, *Gastroenterol.*, **154** (2018), 360–373. https://doi.org/10.1053/j.gastro.2017.08.023
7. J. Lagergren, E. Smyth, D. Cunningham, P. Lagergren, Oesophageal cancer, *Lancet*, **390** (2017), 2383–2396. https://doi.org/10.1016/S0140-6736(17)31462-9
8. C. C. Abnet, Y. L. Qiao, S. D. Mark, Z. W. Dong, P. R. Taylor, S. M. Dawsey, Prospective study of tooth loss and incident esophageal and gastric cancers in China, *Cancer Causes Control*, **12** (2001), 847–854. https://doi.org/10.1023/a:1012290009545
9. N. A. Dar, F. Islami, G. A. Bhat, I. A. Shah, M. A. Makhdoomi, B. Iqbal, et al., Poor oral hygiene and risk of esophageal squamous cell carcinoma in Kashmir, *Br. J. Cancer*, **109** (2013), 1367–1372. https://doi.org/10.1038/bjc.2013.437

10. N. Guha, P. Boffetta, V. Wünsch Filho, J. Eluf Neto, O. Shangina, D. Zaridze, et al., Oral health and risk of squamous cell carcinoma of the head and neck and esophagus: results of two multicentric case-control studies, *Am. J. Epidemiol.*, **166** (2007), 1159–1173. https://doi.org/10.1093/aje/kwm193

11. S. Kageyama, T. Takeshita, M. Furuta, M. Tomioka, M. Asakawa, S. Suma, et al., Relationships of variations in the tongue microbiota and pneumonia mortality in nursing home residents, *J. Gerontol. A*, **73** (2018), 1097–1102. https://doi.org/10.1093/gerona/glx205

12. K. E. Kholy, R. J. Genco, T. E. Dyke, Oral infections and cardiovascular disease, *Trends Endocrin. Met.*, **26** (2015), 315–321. https://doi.org/10.1016/j.tem.2015.03.001

13. E. Zaura, B. W. Brandt, A. Prodan, M. J. Teixeira de Mattos, S. Imangaliyev, J. Kool, et al., On the ecosystemic network of saliva in healthy young adults, *ISME J.*, **11** (2017), 1218–1231. https://doi.org/10.1038/ismej.2016.199

14. R. Vasapolli, K. Schütte, C. Schulz, M. Vital, D. Schomburg, D. H. Pieper, et al., Analysis of transcriptionally active bacteria throughout the gastrointestinal tract of healthy individuals, *Gastroenterology*, **157** (2019), 1081–1092. https://doi.org/10.1053/j.gastro.2019.05.068

15. X. Cao, Intestinal inflammation induced by oral bacteria, *Science*, **358** (2017), 308–309. https://doi.org/10.1126/science.aap9298

16. B. Corning, A. P. Copland, J. W. Frye, The esophageal microbiome in health and disease, *Curr. Gastroenterol. Rep.*, **20** (2018), 1–7. https://doi.org/10.1007/s11894-018-0642-9

17. X. Chen, B. Winckler, M. Lu, H. Cheng, Z. Yuan, Y. Yang, et al., Oral microbiota and risk for esophageal squamous cell carcinoma in a high-risk area of China, *PloS One*, **10** (2015), e0143603. https://doi.org/10.1371/journal.pone.0143603

18. Z. Li, L. Dou, Y. Zhang, S. He, D. Zhao, C. Hao, et al., Characterization of the oral and esophageal microbiota in esophageal precancerous lesions and squamous cell carcinoma, *Front. Cell. Infect. Microbiol.*, **11** (2021), 714162. https://doi.org/10.3389/fcimb.2021.714162

19. H. Li, Z. Lou, H. Zhang, N. Huang, D. Li, C. Luo, et al., Characteristics of oral microbiota in patients with esophageal cancer in China, *BioMed Res. Int.*, **2021** (2021), 2259093. https://doi.org/10.1155/2021/2259093

20. M. Kawasaki, Y. Ikeda, E. Ikeda, M. Takahashi, D. Tanaka, Y. Nakajima, et al., Oral infectious bacteria in dental plaque and saliva as risk factors in patients with esophageal cancer, *Cancer*, **127** (2021), 512–519. https://doi.org/10.1002/cncr.33316

21. Q. Zhao, T. Yang, Y. Yan, Y. Zhang, Z. Li, Y. Wang, et al., Alterations of Oral microbiota in Chinese patients with esophageal cancer, *Front. Cell. Infect. Microbiol.*, **10** (2020), 541144. https://doi.org/10.3389/fcimb.2020.541144

22. Q. Wang, Y. Rao, X. Guo, N. Liu, S. Liu, P. Wen, et al., Oral microbiome in patients with oesophageal squamous cell carcinoma, *Sci. Rep.*, **9** (2019), 19055. https://doi.org/10.1038/s41598-019-55667-w

23. F. Liu, M. Liu, Y. Liu, C. Guo, Y. Zhou, F. Li, et al., Oral microbiome and risk of malignant esophageal lesions in a high-risk area of China: A nested case-control study, *Chinese J. Cancer Res.*, **32** (2020), 742–754. https://doi.org/10.21147/j.issn.1000-9604.2020.06.07

24. B. A. Peters, J. Wu, Z. Pei, L. Yang, M. P. Purdue, N. D. Freedman, et al., Oral microbiome composition reflects prospective risk for esophageal cancers, *Cancer Res.*, **77** (2017), 6777–6787. https://doi.org/10.1158/0008-5472.CAN-17-1296

25. W. Lv, *Identification of the Microbial Composition of the Patients with Esophageal Squamous Cell Carcinoma and Analysis of the Differences in Microbial Composition from Healthy Subjects*, Master thesis, Hebei Medical University in Shijiazhuang, 2021. https://doi.org/10.27111/d.cnki.ghyku.2021.000887

26. D. Shao, *The Characteristic of Microbial Communities of Oral Cavity, Esophagus and Cardia of Population in High-Risk Regions of Esophageal Cancer in China*, Ph.D thesis, Peking Union Medical College in Beijing, 2021. https://doi.org/10.27648/d.cnki.gzxhu.2021.000407

27. Y. Lu, *Microbiota of the Tumor Tissue and Saliva in Patients with Esophageal Cancer*, Ph.D thesis, Peking Union Medical College in Beijing, 2021.

28. K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, D. I. Fotiadis, Machine learning applications in cancer prognosis and prediction, *Comput. Struct. Biotechnol. J.*, **13** (2014), 8–17. https://doi.org/10.1016/j.csbj.2014.11.005

29. R. Mofidi, C. Deans, M. D. Duff, A. C. de Beaux, S. Paterson Brown, Prediction of survival from carcinoma of oesophagus and oesophago-gastric junction following surgical resection using an artificial neural network, *Eur. J. Surg. Oncol.*, **32** (2006), 533–539. https://doi.org/10.1016/j.ejso.2006.02.020

30. Y. Hayashida, K. Honda, Y. Osaka, T. Hara, T. Umaki, A. Tsuchida, et al., Possible prediction of chemoradiosensitivity of esophageal cancer by serum protein profiling, *Clin. Cancer Res.*, **11** (2005), 8042–8047. https://doi.org/10.1158/1078-0432.CCR-05-0656

31. Z. Xun, Q. Zhang, T. Xu, N. Chen, F. Chen, Dysbiosis and ecotypes of the salivary microbiome associated with inflammatory bowel diseases and the assistance in diagnosis of diseases using oral bacterial profiles, *Front. Microbiol.*, **9** (2018), 1136. https://doi.org/10.3389/fmicb.2018.01136

32. L. A. David, A. C. Materna, J. Friedman, M. I. Campos-Baptista, M. C. Blackburn, A. Perrotta, et al., Host lifestyle affects human microbiota on daily timescales, *Genome Biol.*, **15** (2014), R89. https://doi.org/10.1186/gb-2014-15-7-r89

33. J. Wei, *Analysis of Oral Salivary Microbiota in Patients with Esophageal Squamous Cell Carcinoma and its Clinical Significance*, Master thesis, Southern Medical University in Canton, 2020. https://doi.org/10.27003/d.cnki.gojyu.2020.000723

34. Z. Zhu, *Study on Risk Factors, Serum Biomarkers, and Salivary Microbiota of Upper Gastrointestinal Cancers*, Ph.D thesis, Peking Union Medical College in Beijing, 2021. https://doi.org/10.27648/d.cnki.gzxhu.2021.000132

35. X. Wan, W. Wang, J. Liu, T. Tong, Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range, *BMC Med. Res. Methodol.*, **14** (2014), 1–13. https://doi.org/10.1186/1471-2288-14-135

36. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., Scikit-learn: machine learning in Python, preprint, arXiv:1201.0490.

37. D. Krstajic, L. J. Buturovic, D. E. Leahy, S. Thomas, Cross-validation pitfalls when selecting and assessing regression and classification models, *J. Cheminform.*, **6** (2014), 1–15. https://doi.org/10.1186/1758-2946-6-10

38. G. Biau, E. Scornet, A Random Forest Guided Tour, *Test*, **25** (2016), 197–227. https://doi.org/10.1007/s11749-016-0481-7

39. F. Najar, S. Bourouis, N. Bouguila, S. Belghith, A comparison between different Gaussian-based mixture models, in *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, (2017), 704–708. https://doi.org/10.1109/AICCSA.2017.108

40. I. Saini, D. Singh, A. Khosla, Delineation of ECG wave components using K-nearest neighbor (KNN) algorithm: ECG wave delineation using KNN, in *2013 10th International Conference on Information Technology: New Generations*, (2013), 712–717. https://doi.org/10.1109/ITNG.2013.76

41. K. He, C. He, Housing price analysis using linear regression and logistic regression: A comprehensive explanation using melbourne real estate data, in *2021 IEEE International Conference on Computing (ICOCO)*, (2021), 241–246. https://doi.org/10.1109/ICOCO53166.2021.9673533

42. A. P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, *Pattern Recogn.*, **30** (1997), 1145–1159. https://doi.org/10.1016/s0031-3203(96)00142-2

43. B. Pan, Application of XGBoost algorithm in hourly PM2.5 concentration prediction, *IOP Conf. Ser.: Earth Environ. Sci.*, **113** (2018), 012127. https://doi.org/10.1088/1755-1315/113/1/012127

44. M. Moghtadaei, M. R. Golpayegani, F. Almasganj, A. Etemadi, M. R. Akbari, R. Malekzadeh, Predicting the risk of squamous dysplasia and esophageal squamous cell carcinoma using minimum classification error method, *Comput. Biol. Med.*, **45** (2014), 51–57. https://doi.org/10.1016/j.compbiomed.2013.11.011

AIMS Press