# Supplementary Materials for paper: FES-RF: A Feature Ensemble Selection Based Random Forest Method For Accurate Cancer Screening

## I. Related Work

### A. The development of RF

With the rapid development of artificial intelligence technology, RF has been widely used in the fields of economic management, medicine, environmental ecology, biological analysis, and so on [20]–[22]. Philipp put forward the super-parameters and the optimization strategy for the RFs. Lee etc used RF to detect lung nodules automatically on CT images and added CAC [23] to RF. By weighting different samples and removing the redundant features, Cuong Nguyen etc applied RF to get high prediction results [24]. Sun etc proposed a framework for cervical cancer diagnosis based on RF along with the aid of Relief feature selection, so as to build robust RF model with compact feature subsets [25].

However, the difficulty of the classification task for medical data mainly lies in the high-dimensional feature space, high feature redundancy, high feature correlation, class imbalance, and the requirement of interpretability in results, traditional data mining algorithms can not well tackle the medical data mining task [26]. Studies had shown that the discriminative ability and stability of diverse learners can be weakened in class imbalance data sets and high dimensional data sets [27]. To well handle these problem, a white-box model with high generalization ability is urgently required.

To solve these problems, this paper proposes an interpretable feature ensemble importance selection RF(FES-RF) method for the multi-class classification problem in the filed of cancer screening. The details are given as follows.

### B. Application of SERS in Cancer Screening

Over the past decades, SERS have been applied to a great deal of analytical systems such as viruses and bacteria detection [28] DNA detection [29], chemical warfare stimulant detection [30], glucose sensing [31], and environmental monitoring [32]. Cancer diagnosis is another type of application for the SERS technique. In the process of canceration, the composition and content of biomolecules such as protein and DNA in human blood will change due to apoptosis and necrosis of cells. Based on this theory, our team has developed a label-free blood measurement experiment based on surface-enhanced Raman spectroscopy for nasopharyngeal [9], gastric [33] and colorectal [34] cancer detection in recent years and have obtained promising preliminary results.

SERS has been applied to distinguish of two or three types of cancers in the previous experiments. Nevertheless, the detection of multiple types of cancers has not been mentioned in the literatures. In this paper, we explore the cancer screening of silver nanoparticles-based SERS for biochemical analysis of blood serum samples, aiming to identify the difference among multiple types of cancers.

## II. Data Set Construction

This study constructs an exclusive SERS spectroscopy dataset by performing SERS measurements and spectral preprocessing on 700 serum samples of patients and healthy volunteers collected from hospitals. In detail, data construction is divided into three steps:

**Synthesis of Ag NPs.** Stable silver (Ag) colloidal solution is prepared by a deoxidation method developed by Leopold and Lendl et al. [35]. Based on this method, 4.5 ml of sodium hydroxide ($0.1mol/L^{-1}$)) and 5 ml of hydroxylamine hydrochloride ($0.06mol/L^{-1}$)) are evenly mixed. With vigorous stirring, 90 mL of silver nitrate aqueous solution ($0.0011mol/L^{-1}$)) is quickly added to the mixture until a uniform milky-white mixture is obtained. The UV-visible absorption spectrum of resulting silver nanocolloidal are shown in Fig.14, with the absorption peak locating at 425 nm. The subfigure shows the TEM micrograph of silver colloidal surface. Then, the silver colloidal solution is concentrated at 10000 RPM using a centrifuge for 10 minutes. The supernatant is discarded and the final concentrate is used for serum samples.

**Preparation of serum samples.** A total of 700 human serum samples are collected from the First Affiliated Hospital of Fujian Medical University in four groups, which are recognized by the local ethics committee. The four subjects include 95 patients diagnosed with HBV, 85 patients diagnosed as leukemia (M5), and 322 patients clinically diagnosed with breast cancer (BC). In addition, 203 volunteers are confirmed to be healthy by health examination, and they serve as the control group. Coagulant is added and centrifuged at 1000 rpm for 10 minutes to remove blood cells, so as to obtain serum.

**SERS measurement.** Aluminum foil is a cheap spectral substrate with almost no background characteristics, which is suitable for Raman spectroscopy measurement of various biological samples. Therefore, the same volume of each serum sample (2.5ul) and silver nano sol (2.5ul) are evenly mixed on aluminum foil and air-dried at room temperature. In this

Identify applicable funding agency here. If none, delete this.

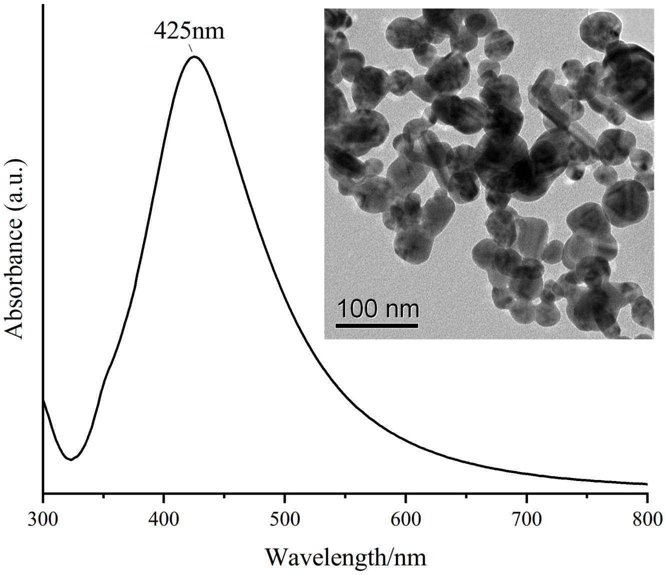

Fig. 1: The UV-visible absorption spectrum of the AgNPs deoxidated by hydroxylamine hydrochloride. The absorption peak is located at 425 nm. The inserted figure show the TEM micrograph of silver colloidal surface.

process, the pipette gun head is used to push and hit to produce as uniform the mixture as possible. The SERS spectra are measured in the range of 400 1800 $cm^{-1}$ using an ultra-fast Raman imaging spectrometer (HORIBA-XploRA plus) equipped with a 785 nm diode laser. All SERS spectra are integrated for 10000 ms at a laser power of 1.0 mW.

### *A. Data Preprocessing*

*1) Spectral correction and normalization:* Since the original SERS signal is disturbed by spontaneous fluorescence signal, we adopt the fifth-order polynomial fitting algorithm [36] to eliminate the spontaneous fluorescence background. In addition, after removing the automatic fluorescence background from the original spectrum, all measured SERS spectra are normalized by the integral area under the curve in the wavenumber range of 400-1800 $cm^{-1}$. In this way, the influence of spectral intensity changes between different spectra on experimental analysis is eliminated and guarantees more accurate spectral shape analysis.

In our experiment, a exclusive dataset composed of 700 instances is obtained through the above process, and the description of dataset are given in Table I. The 700 SERS spectra we measured all consisted of the same 1087 discrete wave numbers corresponding to the intensity. Therefore, it is reasonable to recognize these 1087 discrete wave numbers as the leaning features. It should be noted that there is a class imbalance problem in our dataset, and class 3 contains the less number of samples.

The normalized mean spectra and standard deviation for the same pathological groups in the dataset is given in Fig.3. It illustrates the common SERS spectral peaks at 491, 590, 633, 722, 808, 888, 1004, 1132 and 1201 $cm^{-1}$ are consistently observed in the four serum sample groups. Comparing the mean spectra of four classes, it can be found that the spectral peak intensity at 491, 633, 808, 888, 888, 1132, and 1444 $cm^{-1}$ changes significantly. In addition, some references [9], [37] indicate that various biochemical components in serum samples, such as lipids, proteins, and nucleic acids, are related to the intensity of the corresponding SERS spectral signal. It provides a theoretical basis for our further experimental investigation.

TABLE I: Data description

| Groups | Number of samples | Dimension of features | label |
|---|---|---|---|
| HBV infection | 95 | 1087 | 1 |
| Breast Cancer | 322 | 1087 | 2 |
| M5-leukemia | 80 | 1087 | 3 |
| Normal | 203 | 1087 | 4 |
| Total | 700 | 1087 | |

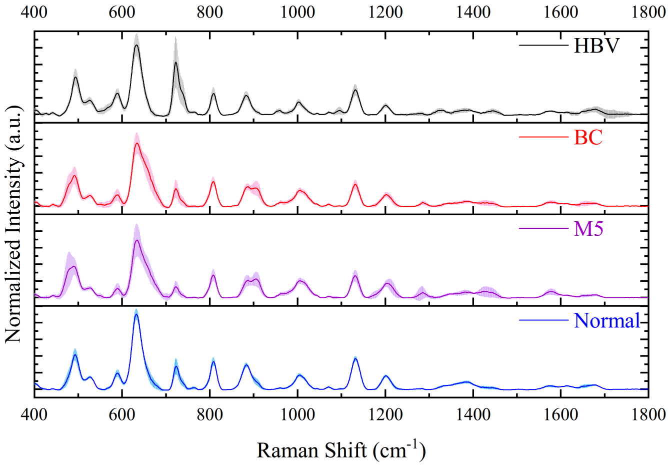

Fig. 2: Comparison of the normalized mean spectra of four groups: HBV patients (black), breast cancer patients (red), M5 leukemia (purple) and Healthy Volunteers (blue) with the standard deviations overlying as shaded color fill.

## III. Ablation Study

In order to verify the effectiveness of feature importance analysis in the model, we conduct ablation experiments and compare the FES-RF Method with two different models, namely: RF and FP-RF. Here, RF is the classical RF model without modification, and FP-RF is a RF model combined with a randomly selected feature selectioin method for the feature importance evaluation, and the top $l$ features are kept for FP-RF. To determine the optimal $l$ for FP-RF, the number of features are changed from 100 to 1000 at the step 100, and the best performance is obtained with setting $l$ to 500.

As shown in table II, FES-RF can promotes the discriminative ability compared with both RF and FP-RF. It is noted that FP-RF and FES-RF both beat RF on the F1-macro and F1-micro metrics, verifying the importance of the feature selection strategy in the classification of SERS data. Therefore, with eliminating the redundant features, the performance of

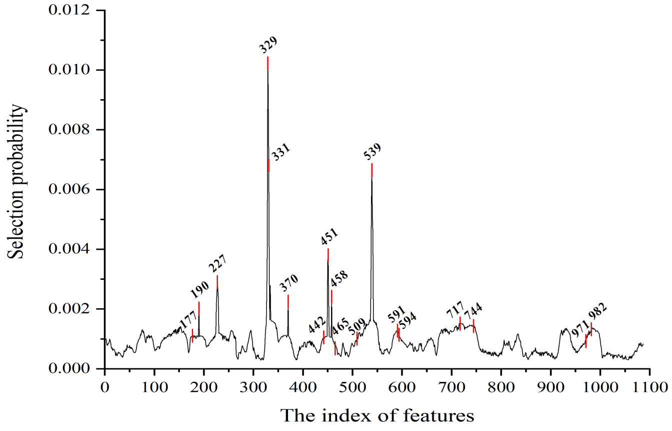

Fig. 3: The feature selection probabilities

classifiers can be much improved. Although FES-RF and FP-RF obtain close scores in the accuracy metric, FES-RF takes advantage over FP-RF in the recall metric, especially on the third class, i.e. the minority class. So, the results reveal that the use of only a type of feature selection method can not well handle the class imbalance problem. And our FFES-RF can correct the bias towards the majority classes through the ensemble selection strategy. In short, the abliation study verifies the importance of each component of our algorithm.

## IV. Correlation analysis and importance ranking with Shapely

According to eight different feature correlation analysis methods, the probability set of all selected features is calculated, and the results are shown in the figure 9. From the figure, it is obvious that the selection probabilities of different features are quite different, and the features with the indices 331, 329, and 538 are much higher than that of other features. To verify whether the frequently selected features are of vital importance, we further calculate the shapely value [48] of these features. Shapely is a common method to quantitatively measure the feature importance. The top 20 features with the highest Shapely values are illustrated in fig.10. It is found that for the selected features, usually higher shapely values correspond to the higher selection probability. We conduct more shapely analysis in Section C in the Appendix.

## V. Related work on feature ranking algorithms

1. Variance Selection method [34]: features are screened through the variance of the feature itself. In order to use the variance selection method, we need to calculate their variance $s$ for each one-dimensional feature $s_i^2$ and then sort from large to small according to the variance value. The variance is calculated as follows:

$$s_i^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\left(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i\right)^2\right] \tag{1}$$

2. Pearson Correlation[35]: Pearson correlation is usually used to measure the degree of linear correlation of interval

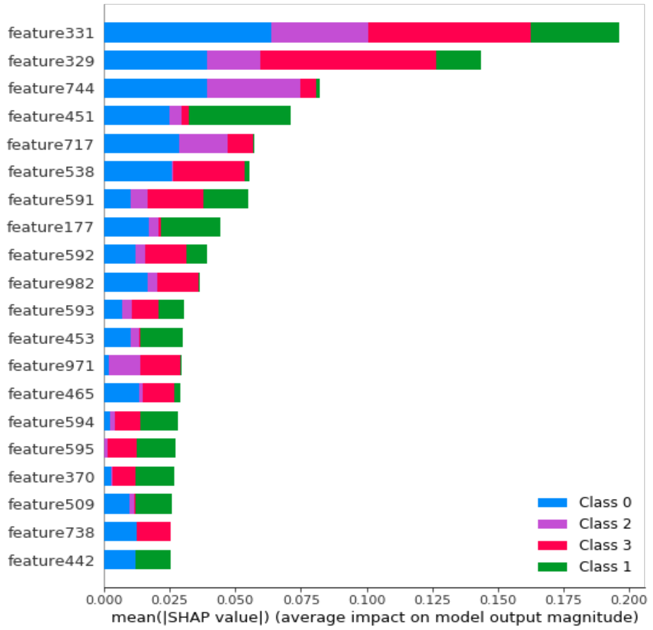

Fig. 4: Top 20 features with the highest impact on the results.

scale data. It is represented by the symbol $r$. the value of $r$ is between - 1 and 1. The greater the absolute value is, the stronger the correlation is. If $r = 0$, there is no linear correlation between $X$ and $y$. We can calculate Pearson correlation $r$ for each one-dimensional feature.

$$r_i = \frac{\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(y_i - \frac{1}{n}\sum_{i=1}^{n} y_i\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i\right)^2}\sqrt{\sum_{i==1}^{n}\left(y_i - \frac{1}{n}\sum_{i=1}^{n} y_i\right)^2}} \tag{2}$$

3. Mutual Info[36]: Indicates whether the two variables X and y are related, and the strength of the relationship. Mutual information is actually a special case of broader relative entropy. If the variables are not independent, we can judge whether they are ”close” to each other by examining the Kullback Leibler divergence between the product of joint probability distribution and edge probability distribution. At this time, Kullback Leibler divergence is mutual information. We have:

$$I\left(X_i; Y\right) = \sum_{x_i \in X_i}\sum_{y_i \in Y} P\left(x_i, y_i\right)\log\frac{P\left(x_i, y_i\right)}{P\left(x_i\right)P\left(y_i\right)} \tag{3}$$

where $P\left(x_i, y_i\right)$ is the joint probability distribution function of $X$ and $Y$, and $P\left(x_i\right)$ And $P\left(y_i\right)$ are the edge probability distribution functions of $X$ and $Y$, respectively.

4. MIC(Maximal Information Coefficient)[37]: It is usually used to detect the nonlinear correlation between variables. The calculation of MIC can be divided into three steps. First, for the two variables to be detected, $X_i$ and $Y$, given $P$, $Q$ values will be determined by $X_i$. The scatter diagram composed of $X_i$ and $Y$ is meshed in $P$ columns and $Q$ rows, and the maximum mutual information is obtained. Through formula (3), then the maximum mutual information value

TABLE II: The Result of Ablation Study.

| Methods | Accuracy | | | | | Recall | | | | | F1-macro | F1-micro |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | class1 | class2 | class3 | class4 | Overall | class1 | class2 | class3 | class4 | Overall | | |
| RF | 89.74 | 89.58 | 88.14 | 93.14 | 90.15 | 93.03 | 90.52 | 65.00 | 93.60 | 85.54 | 87.64 | 90.57 |
| FP-RF | 90.70 | 94.51 | 92.72 | 92.38 | 92.58 | 96.89 | 90.53 | 63.95 | 95.57 | 86.74 | 88.92 | 91.86 |
| FES-RF | 94.41 | 93.68 | 90.00 | 97.04 | 93.78 | 96.20 | 94.68 | 86.74 | 95.17 | 93.20 | 93.48 | 94.57 |

is normalized, and finally the maximum mutual information value at different scales is selected as the MIC value. The formula is:

$$\operatorname{MIC}\left[X_i ; Y\right]=\max _{\left|X_i\right|\left|Y\right|<B} \frac{I\left(X_i ; Y\right)}{\log \left(\min \left(\left|X_i\right|\left|Y\right|\right)\right)} \tag{4}$$

Where $B$ is an empirical value, usually 0.6 or 0.55 power of the total data.

5. Distance Correlation[38]. Distance correlation absorption is proposed to overcome the weakness of the Pearson correlation coefficient. In some cases, even if the value of the Pearson correlation coefficient of two variables $X$ and $Y$ is equal to 0, we can not judge that the two variables are independent of each other (it may also be a nonlinear correlation). If the distance correlation coefficient is 0, we can say that the two variables are completely independent.

Through distance correlation, we can study the independence of two variables X and Y, which are recorded as $dcorr(X, Y)$. When $dcorr(X, Y) = 0$, the two variables are independent of each other. The greater the $dcorr(X, Y)$ value, the greater the correlation between $X$ and $y$. We assume $\{(X_i, Y_i), i = 1, 2, \ldots, n\}$ is the random sample of the population$(X, Y)$. Szekely (2018) (et. c.) defines the DC sample estimates of $X$ and $Y$ of two random variables as:

$$dcorr(X, Y)=\frac{dcov(X, Y)}{\sqrt{dcov(X, X)} \sqrt{dcov(Y, Y)}} \tag{5}$$

6. Kendall correlation [39]: Kendall rank correlation coefficient is a rank statistical parameter with non-parametric characteristics (independent of distribution). It is a correlation coefficient used to measure the strength of the monotonic relationship between two ordered variables. Its value range is $[-1, 1]$. If the two attributes rank the same, the coefficient is 1, and the two attributes are positively correlated. If two attributes rank completely opposite and the coefficient is $-1$, the two attributes are negatively correlated. If the ranking is completely independent, the coefficient is 0. The original Kendall rank correlation coefficient is defined in the concepts of coordinated pairs and uncoordinated pairs. The so-called consistent pair means that the relative relationship between the values of two variables is consistent; Divergent pairs mean that their relative relationship is inconsistent. It is expressed by a formula as the ratio of the difference between concordant and discordant pairs to the total logarithm $(n * (n - 1)/2)$, which is defined as the Kendall coefficient.

$$\tau=\frac{c-d}{n *(n-1) / 2} \tag{6}$$

The $c$,$d$ represent the number of concordant pairs and discordant pairs respectively, and $n$ represents the number of statistical objects of the same kind

7. Spearmen Correlation[40]: Spearman correlation coefficient is a nonparametric index to measure the dependence of two variables. It uses monotone equation to evaluate the correlation direction of two statistical variables. If there are no duplicate values in the data and the two variables are completely monotonically correlated, the Spearman correlation coefficient is $+1$ or $-1$. Spearman correlation coefficient is defined as Pearson correlation coefficient between hierarchical variables. For samples with a sample size of $N$, $N$ original data are converted into hierarchical data, and the correlation coefficient is:

$$\rho=\frac{\sum_i\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\sum_i\left(x_i-\bar{x}\right)^2 \sum_i\left(y_i-\bar{y}\right)^2}} \tag{7}$$

8. Tree-based importance[41]: Gini importance. Tree-based estimators can be used to calculate feature importance, which in turn can be used to discard irrelevant features. In the construction of the tree, we will calculate a criterion (e.g. Gini index or information entropy) for each feature, and the feature importance is the normalized value of the reduction of the criterion. The default implementation in sci-kit learn is what we often call Gini importance.

## VI. ROC Curve and Heat Map

We use ROC curve to show the performance of the data in the test process, in which the error probability is the abscissa and the correct probability is the ordinate. As is shown in Figure 5, the points drawn by the results are connected. Here the AUC score is 0.995, which confirms the excellent performance of EFS-RF.

Figure 6 given below illustrates the confusion matrix, which reveals more details about the classification results. It is found that HBV and M5 are easy classes as they distinguished from other classes with low errors. While classes tend to be wrongly identified as the health classes, and the reason may lie in the sample size of the health class is the largest in the cancer data set. Consequently, classifiers tend to bias towards the majority classes.

## VII. Shapeley Visualization

### A. Summary Analysis

We continue to plot the snap values for each feature for each sample, providing a better way to illustrate the overall patterns. Figure 7 to Figure 10 show the distinguishing ability of the top 20 features that provide the greatest impact on the identification of various classes. Here, each row represents a

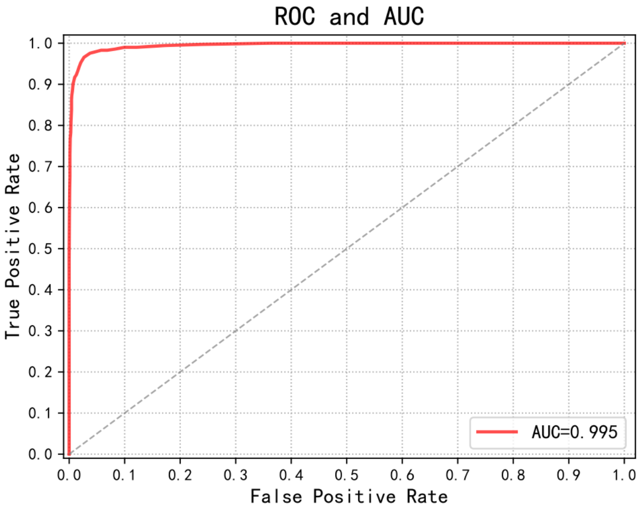

Fig. 5: ROC curve of EFS-RF.

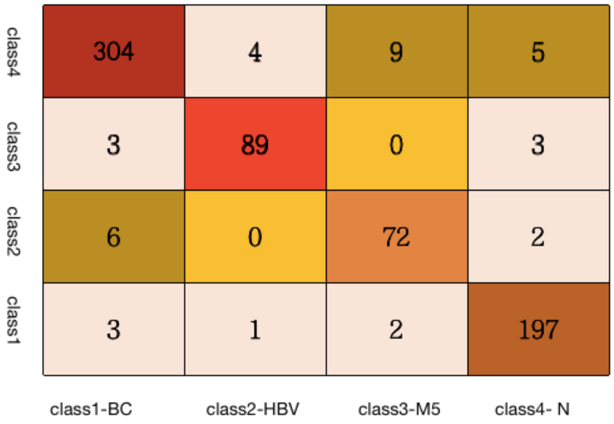

Fig. 6: The confusion matrix of EFS-RF.

feature, and the abscissa is the SHAP value. A point represents a sample, and the color represents the eigenvalue (the red color represents a high value, and the blue color indicates a low value). The abscissa is the variation range of Shapley value. The left ordinate is characterized from high to low, and the right ordinate is characterized by the original value from high to low. The color changes from red to blue. It can be seen that in the classification of class 3, the eigenvalue and Shapley value show obvious correlation, that is, the larger the eigenvalues of features, the larger the Shapley value (e.g. 329,331), and vice versa(e.g. 465,184). It shows that these important features are key to the classification task through the original eigenvalues.

### *B. Dependency Analysis*

This section discusses the influence of the variable on the target value under the interaction of two variables. Figure 11 to Figure 14 show the dependence between the features with the greatest impact in the first two dimensions and its

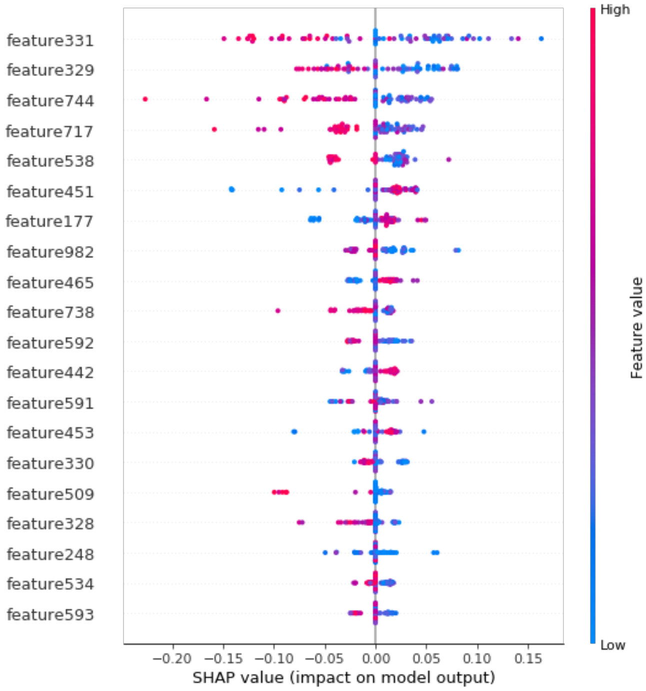

Fig. 7: The distinguishing ability of the first twenty features that have the greatest impact on Class1.

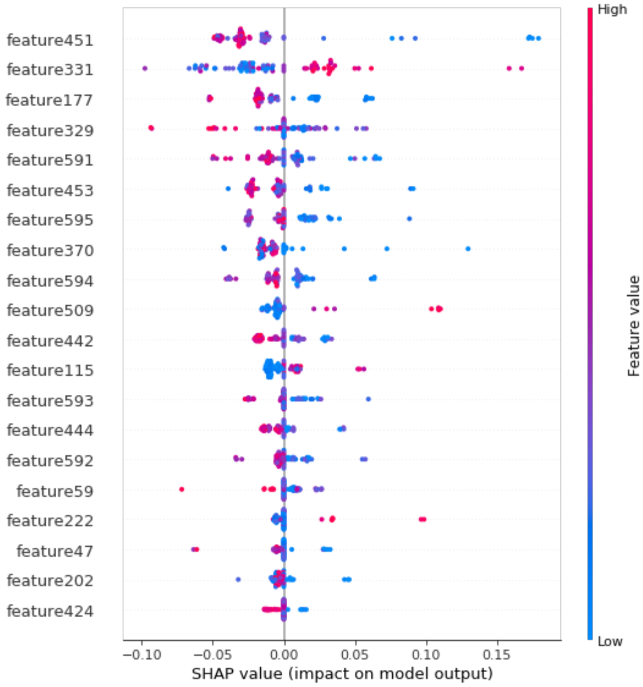

Fig. 8: The distinguishing ability of the first twenty features that have the greatest impact on Class2.

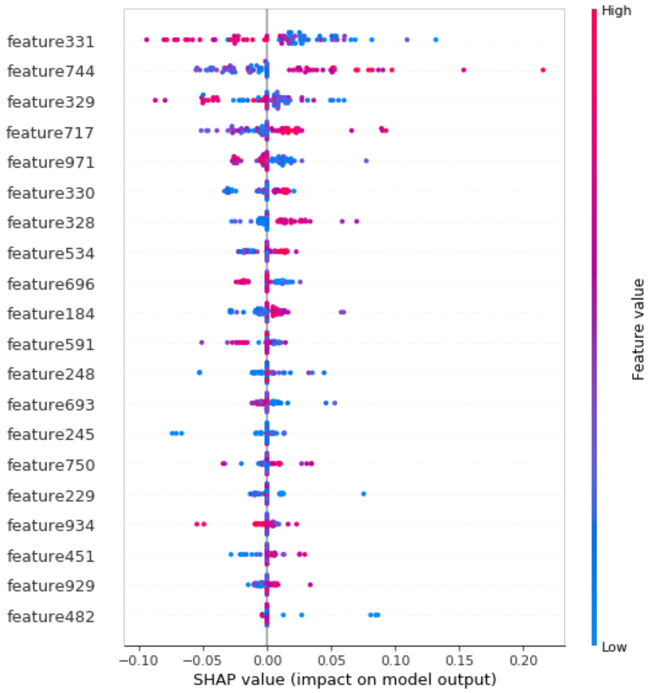

Fig. 9: The distinguishing ability of the first twenty features that have the greatest impact on Class3.

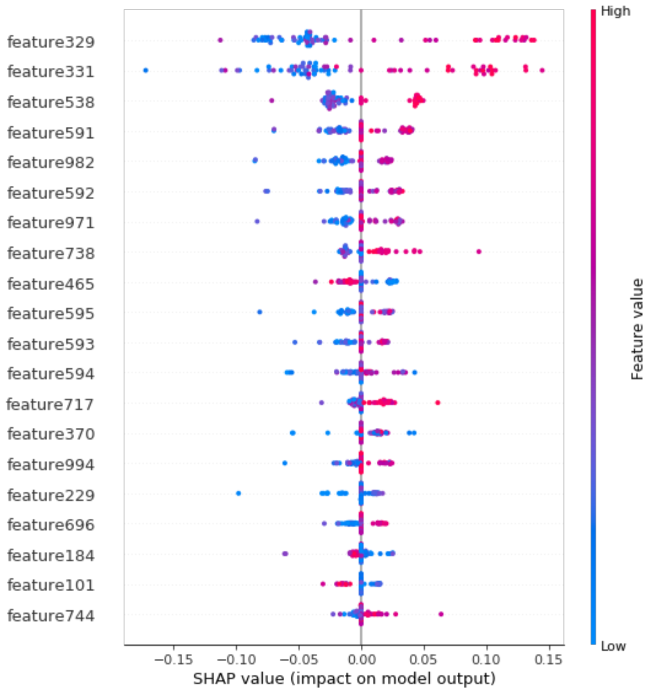

Fig. 10: The distinguishing ability of the first twenty features that have the greatest impact on Class4.

impact on sample classification. The abscissa is the variable age, the left ordinate is the shapely value for feature329, and the right ordinate is the feature 995, As shown in the figure, with the increase of feature 329 (from left to right), feature 995 also increases gradually (the color of sample points gradually changes from blue to red). In addition, with the increase of feature 329 (from left to right), the shape value for feature 329 shows different changes in different classifications (gradually decreasing in i.e. class 1 and increasing in class 4). Moreover, the shape value for feature 329 and feature 995 are sensitive to the of feature 329. Whenever feature 329 fluctuates slightly, they change greatly.

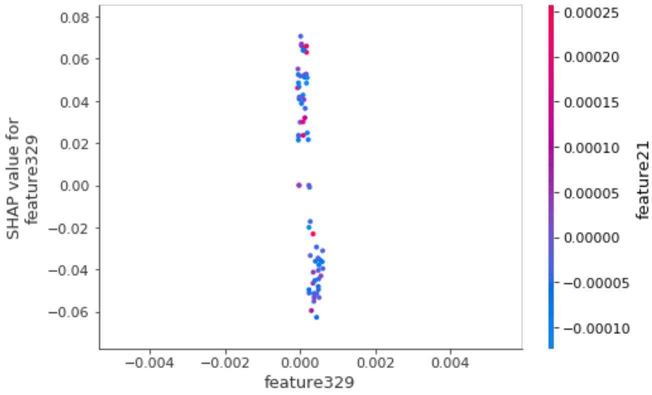

Fig. 11: Influence of the dependence between the first three-dimensional maximum features on Class1.

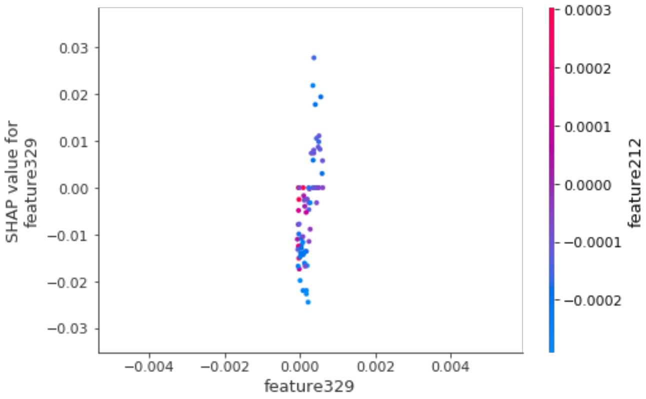

Fig. 12: Influence of the dependence between the first three-dimensional maximum features on Class2.

### C. *Impact on the Single Sample*

Figures 15 to Figures 18 show the impact of the characteristics of a single test sample on the classification effect. The force plot can explain the prediction for a single sample. It visualizes the SHAP values as force. Each eigenvalue is a force that increases or decreases the prediction. The prediction starts from the base value. The baseline is the constant of

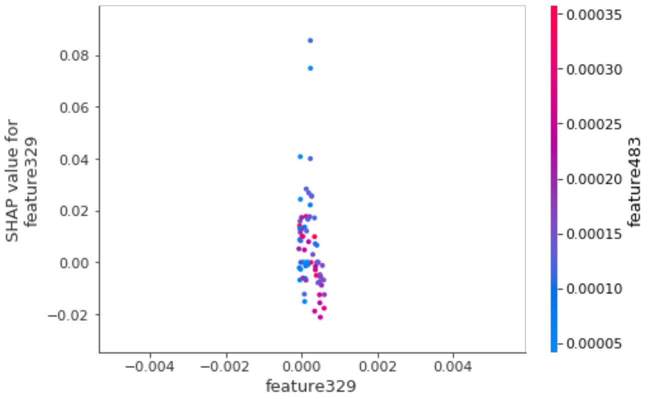

Fig. 13: Influence of the dependence between the first three-dimensional maximum features on Class3.

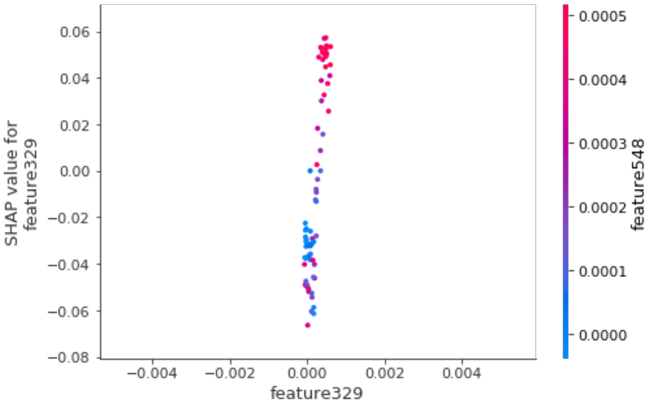

Fig. 14: Influence of the dependence between the first three-dimensional maximum features on Class4.

the interpretation model, and each attribution value is an arrow, leading to increase (positive) or decrease (negative) the forecast. Since tree SHAP explains log odd ratio, negative numbers appear. The characteristics of different dimensions show game interaction for the different classes to which the samples belong. It can be seen that the following figure shows that each feature has its own contribution. The prediction results are obtained from the base value to the final output value. The features that push up the prediction are represented in red, and those features pushing down the prediction are represented in blue. The sample belongs to a class division, and the variables with negative influence are 331, 538, 717, etc. these characteristics show positive influence on other classes.

## References

[1] Stewart, Bernard W., and Paul Kleihues, eds. ”World cancer report.” (2003): 181-188.

[2] Song, M. , et al. ”Cancer prevention: Molecular and epidemiologic consensus.” Science 361.6409(2018):1317-1318.

[3] Gremlich H U , Yan B . Infrared and Raman Spectroscopy of Biological Materials. 2000.

[4] F. Shangyuan et al., ”Gold Nanoparticle Based Surface-Enhanced Raman Scattering Spectroscopy of Cancerous and Normal Nasopharyngeal Tissues under Near-Infrared Laser Excitation,” Applied Spectroscopy, 2009/10/01 2009.

[5] H. Xiaohua, H. E.-S. Ivan, Q. Wei, and A. E.-S. Mostafa, ”Cancer cells assemble and align gold nanorods conjugated to antibodies to produce highly enhanced, sharp, and polarized surface Raman spectra: a potential cancer diagnostic marker,” Nano Letters, 2007/06/01 2007.

[6] L. Xiaozhou, Y. Tianyue, and L. Junxiu, ”Spectral analysis of human saliva for detection of lung cancer using surface-enhanced Raman spectroscopy,” Journal of Biomedical Optics, 2012/04/17 2012.

[7] C. Mustafa, S. David, and V.-D. Tuan, ”Surface-enhanced Raman scattering for cancer diagnostics: detection of the BCL2 gene,” Expert Review of Molecular Diagnostics, 2003/09/27 2003.

[8] Chan, Raymond Javan, et al. ”Relationships between financial toxicity and symptom burden in cancer survivors: a systematic review.” Journal of pain and symptom management 57.3 (2019): 646-660.

[9] S. Feng et al., ”Nasopharyngeal cancer detection based on blood plasma surface-enhanced Raman spectroscopy and multivariate analysis,” Biosens Bioelectron, vol. 25, no. 11, pp. 2414-9, Jul 15 2010.

[10] S. Feng et al., ”Study on gastric cancer blood plasma based on surface-enhanced Raman spectroscopy combined with multivariate analysis,” Science China Life Sciences, vol. 54, no. 9, pp. 828-834, 2011.

[11] Y. Lu et al., ”Label free hepatitis B detection based on serum derivative surface enhanced Raman spectroscopy combined with multivariate analysis,” Biomed Opt Express, vol. 9, no. 10, pp. 4755-4766, Oct 1 2018.

[12] L. Juqiang et al., ”Differentiation of digestive system cancers by using serum protein-based surface-enhanced Raman spectroscopy,” Journal of Raman Spectroscopy, 2016/07/08 2016.

[13] X. Rui et al., ”Non-invasive detection of hepatocellular carcinoma serum metabolic profile through surface-enhanced Raman spectroscopy,” Nanomedicine: Nanotechnology, Biology and Medicine, 2016/08/09 2016.

[14] Biau, Gérard, and Erwan Scornet. ”A RF guided tour.” Test 25.2 (2016): 197-227.

[15] Qi, Yanjun. ”RF for bioinformatics.” Ensemble machine learning. Springer, Boston, MA, 2012. 307-323.

[16] Huljanah, Mia, et al. ”Feature selection using RF classifier for predicting prostate cancer.” IOP Conference Series: Materials Science and Engineering. Vol. 546. No. 5. IOP Publishing, 2019.

[17] Amjad, Arslan, et al. ”Raman spectroscopy based analysis of milk using RF classification.” Vibrational Spectroscopy 99 (2018): 124-129.

[18] Khan, Saranjam, et al. ”RF-based evaluation of Raman spectroscopy for dengue fever analysis.” Applied spectroscopy 71.9 (2017): 2111-2117.

[19] Arend, Natalie, et al. ”Detection and differentiation of bacterial and fungal infection of neutrophils from peripheral blood using raman spectroscopy.” Analytical Chemistry 92.15 (2020): 10560-10568.

[20] Svetnik, Vladimir, et al. ”Application of Breiman’s RF to modeling structure-activity relationships of pharmaceutical molecules.” International Workshop on Multiple Classifier Systems. Springer, Berlin, Heidelberg, 2004.

[21] Naghibi, Seyed Amir, Kourosh Ahmadi, and Alireza Daneshi. ”Application of support vector machine, RF, and genetic algorithm optimized RF models in groundwater potential mapping.” Water Resources Management 31.9 (2017): 2761-2775.

[22] Polishchuk, Pavel G., et al. ”Application of RF approach to QSAR prediction of aquatic toxicity.” Journal of chemical information and modeling 49.11 (2009): 2481-2488.

[23] Lee, Shu Ling Alycia, Abbas Z. Kouzani, and Eric J. Hu. ”RF based lung nodule classification aided by clustering.” Computerized medical imaging and graphics 34.7 (2010): 535-542.

[24] Nguyen, Cuong, Yong Wang, and Ha Nam Nguyen. ”RF classifier combined with feature selection for breast cancer diagnosis and prognostic.” (2013).

[25] Sun, Guanglu, et al. ”Cervical cancer diagnosis based on random

[26] Haghiri, Siavash, Damien Garreau, and Ulrike Luxburg. ”Comparison-based RFs.” International Conference on Machine Learning. PMLR, 2018.

[27] Athey, Susan, Julie Tibshirani, and Stefan Wager. ”Generalized RFs.” The Annals of Statistics 47.2 (2019): 1148-1178.

[28] M. J. Roger and G. Royston, ”Characterisation and identification of bacteria using SERS,” Chemical Society Reviews, 2008/05/01 2008.

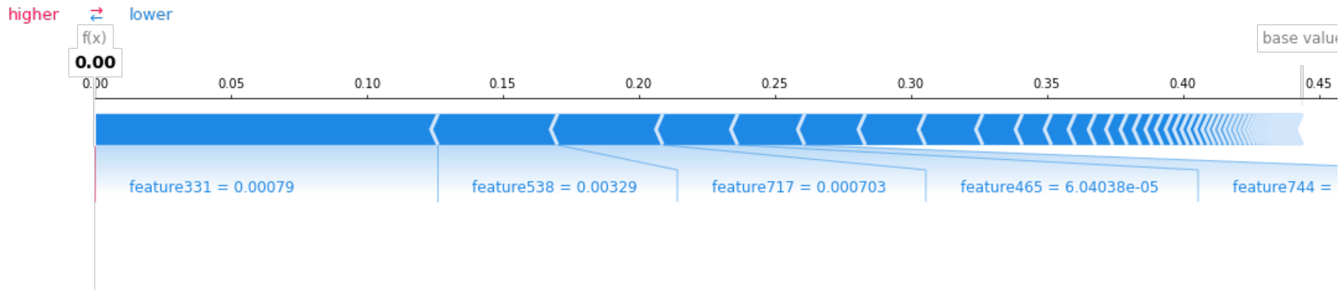

Fig. 15: Impact of a single sample on Class1 classification.

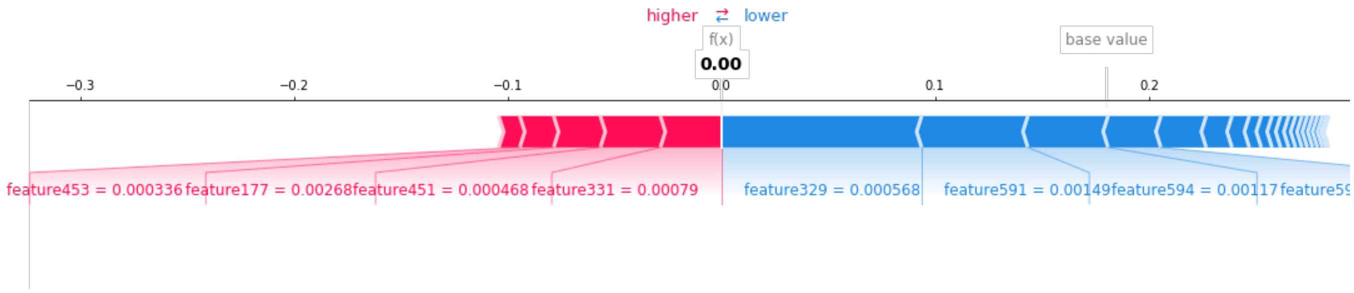

Fig. 16: Impact of a single sample on Class2 classification.

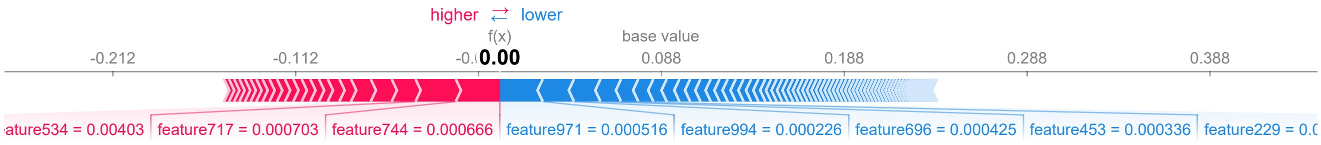

Fig. 17: Impact of a single sample on Class3 classification.

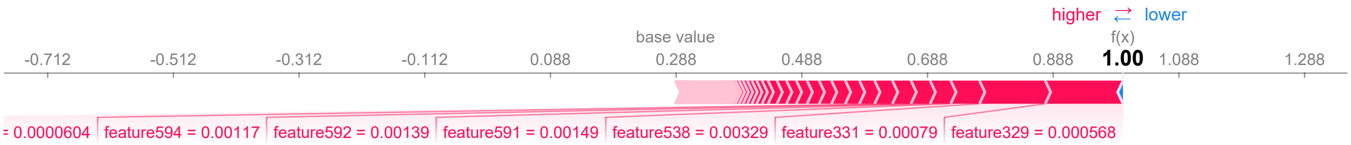

Fig. 18: Impact of a single sample on Class4 classification.

[29] Taton, A. T, Mirkin, A. C, Letsinger, and L. R, ”Scanometric DNA array detection with nanoparticle probes,” Science, 2000/09/08 2000.

[30] A. S. Douglas, B. B. Kevin, and P. V. D. Richard, ”Surface-enhanced Raman spectroscopy of half-mustard agent,” Analyst, 2006/04/01 2006.

[31] A. S. Douglas et al., ”Glucose sensing using near-infrared surface-enhanced Raman spectroscopy: gold surfaces, 10-day stability, and improved accuracy,” Analytical Chemistry, 2005/07/01 2005.

[32] D. L. Stokes, J. P. Alarie, V. Ananthanarayanan, and T. Vo-Dinh, ”Fiber optic SERS sensor for environmental monitoring,” Proceedings of SPIE - The International Society for Optical Engineering, vol. 3534, 1999.

[33] X. Lin et al., ”Label-free liquid biopsy based on urine analysis using surface-enhanced Raman spectroscopy for noninvasive gastric and breast cancer detection,” Journal of Raman Spectroscopy, vol. 51, no. 11, pp. 2245-2254, 2020.

[34] J. Lin et al., ”Rapid and label-free urine test based on surface-enhanced Raman spectroscopy for the non-invasive detection of colorectal cancer at different stages,” Biomedical Optics Express, vol. 11, no. 12, 2020.

[35] N. Leopold and B. Lendl, ”A New Method for Fast Preparation of Highly Surface-Enhanced Raman Scattering (SERS) Active Silver Colloids at Room Temperature by Reduction of Silver Nitrate with Hydroxylamine Hydrochloride,” The Journal of Physical Chemistry B, vol. 107, no. 24, pp. 5723-5727, 2003.

[36] Z. Jianhua, L. Harvey, I. M. David, and Z. Haishan, ”Automated Autofluorescence Background Subtraction Algorithm for Biomedical Raman Spectroscopy,” Applied Spectroscopy, 2007/11/01 2007.

[37] A. Bonifacio, S. D. Marta, R. Spizzo, S. Cervo, and V. Sergo, ”Surface-enhanced Raman spectroscopy of blood plasma and serum using Ag and Au nanoparticles: a systematic study,” Analytical and Bioanalytical Chemistry, vol. 406, no. 9, 2014.

[38] Clark, Linda A., and Daryl Pregibon. ”Tree-based models.” Statistical models in S. Routledge, 2017. 377-419.

[39] Liaw, Andy, and Matthew Wiener. ”Classification and regression by randomForest.” R news 2.3 (2002): 18-22.

[40] Sharaff, Aakanksha, and Harshil Gupta. ”Extra-tree classifier with

metaheuristics approach for email classification." Advances in Computer Communication and Computational Sciences. Springer, Singapore, 2019. 189-197.

[41] Safavian, S. Rasoul, and David Landgrebe. "A survey of decision tree classifier methodology." IEEE transactions on systems, man, and cybernetics 21.3 (1991): 660-674.

[42] C Chen, Tianqi, et al. "Xgboost: extreme gradient boosting." R package version 0.4-2 1.4 (2015): 1-4.

[43] Rao, Haidi, et al. "Feature selection based on artificial bee colony and gradient boosting decision tree." Applied Soft Computing 74 (2019): 634-642.

[44] Ke, Guolin, et al. "Lightgbm: A highly efficient gradient boosting decision tree." Advances in neural information processing systems 30 (2017): 3146-3154.

[45] Peterson, Leif E. "K-nearest neighbor." Scholarpedia 4.2 (2009): 1883.

[46] Steinwart, Ingo, and Andreas Christmann. Support vector machines. Springer Science & Business Media, 2008.

[47] Szegedy, Christian, Alexander Toshev, and Dumitru Erhan. "Deep neural networks for object detection." (2013).

[48] Winter, Eyal. "The shapley value." Handbook of game theory with economic applications 3 (2002): 2025-2054.

[49] Scheffe, Henry. The analysis of variance. Vol. 72. John Wiley and Sons, 1999.

[50] Benesty, Jacob, et al. "Pearson correlation coefficient." Noise reduction in speech processing. Springer, Berlin, Heidelberg, 2009. 1-4.

[51] Shlens, Jonathon. "Notes on Kullback-Leibler divergence and likelihood." arXiv preprint arXiv:1404.2000 (2014).

[52] Wang, Ruping, et al. "MIC-KMeans: a Maximum information coefficient based high-dimensional clustering algorithm." Computer Science Online Conference. Springer, Cham, 2018.

[53] Li, Runze, Wei Zhong, and Liping Zhu. "Feature screening via distance correlation learning." Journal of the American Statistical Association 107.499 (2012): 1129-1139.

[54] Abdi, Hervé. "The Kendall rank correlation coefficient." Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, CA (2007): 508-510.

[55] Myers, Leann, and Maria J. Sirois. "Spearman correlation coefficients, differences between." Encyclopedia of statistical sciences 12 (2004).

[56] Clark, Linda A., and Daryl Pregibon. "Tree-based models." Statistical models in S. Routledge, 2017. 377-419.