NISTIR 8383

Standard Errors and Significance Testing in Data Analysis for Testing Classifiers

Jin Chu Wu
Raghu N. Kacker

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8383

National Institute of
Standards and Technology
U.S. Department of Commerce

NISTIR 8383

Standard Errors and Significance Testing in Data Analysis for Testing Classifiers

Jin Chu Wu
Raghu N. Kacker

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8383

July 2021

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Institute of Standards and Technology
James K. Olthoff, Performing the Non-Exclusive Functions and Duties of the Under Secretary of Commerce for Standards and Technology & Director, National Institute of Standards and Technology

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

National Institute of Standards and Technology Interagency or Internal Report 8383
Natl. Inst. Stand. Technol. Interag. Intern. Rep. 8383, 25 pages (July 2021)

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8383

Standard Errors and Significance Testing in Data Analysis for Testing Classifiers

Jin Chu Wu and Raghu N. Kacker
National Institute of Standards and Technology, Gaithersburg, MD 20899

Abstract – The one-classifier and two-classifier significance testing for evaluation and comparison of classifiers are conducted to investigate the statistical significance of differences and provide quantitative information in terms of the significance level, i.e., *p*-value, in a new ROC analysis where three score distributions and two decision thresholds are employed, and data dependency caused by multiple use of the same subjects is involved. To analyze the performance of classifiers, the standard error of the cost function is estimated using the nonparametric three-sample two-layer bootstrap algorithm on a two-layer data structure constructed after dataset optimization, based on our prior rigorous statistical research in ROC analysis on large datasets with data dependency. In comparison, the positive correlation coefficient must be taken into consideration, which is computed using a synchronized resampling algorithm; otherwise, the likelihood of detecting the statistical significance of difference between the performance levels of two classifiers can be wrongly reduced.

Keywords: ROC analysis, data dependency, standard error, bootstrap, statistical significance, significance testing

1 Introduction

The conventional ROC analysis involves two score distributions, i.e., the distributions of target scores and non-target scores, as depicted in Figure 1 (A) [1-6]. Target scores are created by comparing two different objects (e.g., images, speech segments, etc.) of the same subject (e.g., face, speaker, etc. correspondingly), and non-target scores are generated by matching two objects of two different subjects. It is assumed that the random samples drawn from a population are independent and identically distributed (i.i.d.). The statistics of interest may be the true accept rate at a specified false accept rate [3], or a weighted sum of the probabilities of type I (miss) and type II (false alarm) errors determined at a given decision threshold t as shown in Figure 1 (A) [4], or other measures [5].

In a new ROC analysis that was adopted in the speaker recognition evaluation (SRE) [7-9], as depicted in Figure 1 (B), three score distributions are generated, i.e., the distributions of target scores, known non-target scores (generated by known speakers), and unknown non-target scores (created by unknown speakers), and two decision thresholds t_1 and t_2 are involved. Hence, two type I errors $\alpha_T(t_1)$ and $\alpha_T(t_2)$, and four type II errors $\beta_K(t_1)$, $\beta_K(t_2)$, $\beta_U(t_1)$ and $\beta_U(t_2)$ are created accordingly. To analyze the performance of classifiers, the statistic of interest, i.e., the cost function (CF) is defined to be an average of the two weighted sums of the probabilities of type I and type II errors at the two thresholds correspondingly (see Section 3).

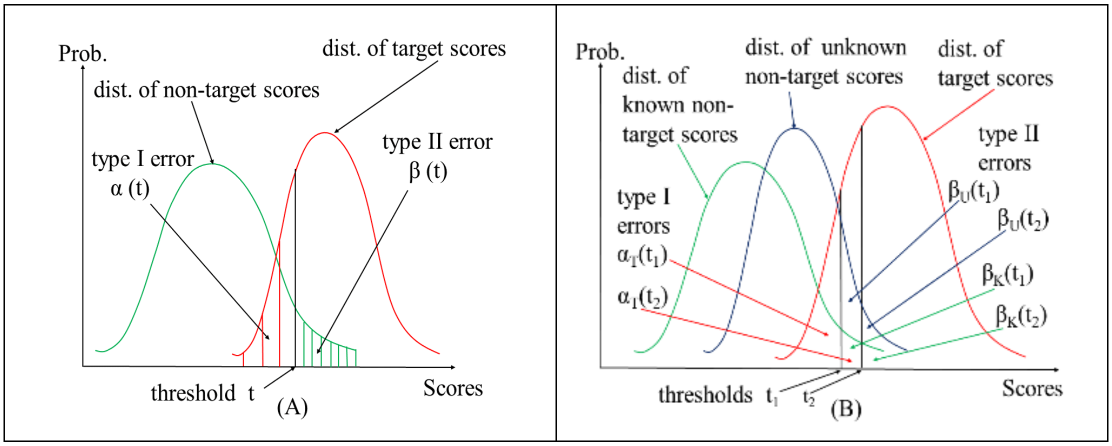

Figure 1 (A): The conventional ROC analysis formed by two score distributions, one threshold t, and two error rates. (B): A new ROC analysis formed by three score distributions, two thresholds t_1 and t_2, and six corresponding error rates.

Moreover, data dependency caused by multiple use of the same subjects exists in order to generate more samples because of limited resources. Particularly in SRE, data dependency was determined by whether the same training speakers were used multiple times while generating datasets [7]. The calls from the same speaker are dependent on this speaker.

In this new ROC analysis, first, the standard error (SE) of CF is estimated using the nonparametric bootstrap algorithm on a two-layer data structure (see below) [3-5, 10-11]. Second, based on such SEs, to investigate the statistical significance of differences for evaluation and comparison of classifiers, the one-classifier and two-classifier significance testing in statistics are carried out to provide quantitative information in terms of the significance level, i.e., *p*-value [3, 5, 11-14]. In comparison, the positive correlation coefficient must be taken into consideration, which is computed using a synchronized resampling algorithm in this article; otherwise, the likelihood of detecting the statistical significance of difference between the performance levels of two classifiers can be wrongly reduced.

In the scenario described above, the traditional analytical approach cannot be applied to estimate the SEs [4-5]. This is because it cannot take account of the distributions of scores, cannot deal with data dependency, and cannot solve the covariance resulting from a weighted sum of the probabilities of type I error and type II error at a decision threshold, which are traded off and thus negatively correlated. As a result, the analytical approach generally underestimates the SEs of measures as opposed to the bootstrap method [4-5, 10-11].

In this article, due to data dependency, the datasets are reorganized into a two-layer data structure after dataset optimization [4, 11, 15]. Target scores are grouped into target sets, when they are created using the same subjects (e.g., the same training speakers in SRE), with equal number of scores in each set, and likewise for known non-target scores and unknown non-target scores (see Section 2-C). Hence, sets are on the first layer and scores within sets are on the second layer. Further, scores in the same set are assumed to be conditionally independent, because they are generated by two sets of objects (e.g., speech segments in SRE) and objects in at least one of the two sets are different. Indeed, this is how the test is designed. Thereafter, to compute the SE of a measure, the nonparametric three-sample two-layer bootstrap resampling takes place randomly *with replacement* (WR) on the three nonparametric score distributions, that is not only on sets but subsequently on scores within the sets.

Our data structure and the variation of bootstrap approach are designed especially for ROC analysis with data dependency [10-11]. For the applications of ROC analysis, scores are not additive, meaning that the addition of two scores does not make any sense. Scores are generated by a classifier while making decision, and different classifiers may employ different types of scoring systems, such as integers, or real numbers, etc. [3, 6]. Thus, no statistic can be used to describe any characteristics related to sets that are created by different subjects, so the traditional statistical methods via computing variances cannot be applied to analyze the stated data structure and the corresponding bootstrap approach.

Hence, the challenges addressed in this article are categorically different from the issues that arise from hierarchical data [16], which are for the applications such as repeated-measures experiments and the classical split-plot experiment, etc. In those cases, to analyze the related data and method, different statistical variances were computed.

In ROC analysis on large datasets, our prior rigorous statistical research was conducted, such as the validation study to provide a sound foundation for using the bootstrap method [17], the extensive bootstrap variability studies to determine the appropriate number of bootstrap replications without and with data dependency in order to reduce the bootstrap variance and ensure the computation accuracy, which took months of CPU time [11, 18-22], and so on. Our data structure and the corresponding bootstrap approach has sound scientific basis.

Here is an important outcome of our research, which can save lots of resources while designing datasets for test related to ROC analysis. If data dependency is involved and the data size is over tens of thousands, datasets should be optimized at the very beginning in such a way that the numbers of scores in sets for each type of scores are equal, respectively (see Section 2-B). The dataset optimization can reduce the size of dataset greatly but changes the performance levels little (see Sections 2-C and 6).

Further, if data dependency is involved but ignored, and even if all the raw data is employed, the SEs of the CF can be underestimated by using the nonparametric three-sample bootstrap method with the i.i.d. assumption (see Section 6).

While evaluating and comparing the performances of classifiers, it may be of interest to determine whether the difference between the performance level of a classifier and a hypothesized performance criterion is real or has happened by chance, or to investigate whether the difference between the performance levels of two classifiers is statistically significant [3, 5, 12-14].

To tackle these challenges, it is insufficient to only estimate the SEs and thus the 95% confidence intervals (CI) of these measures. If a 95% CI contains the hypothesized performance criterion for evaluation or if two 95% CIs overlap for comparison, then it is difficult to draw any quantitative conclusion (e.g., using *p*-value) on the statistical significance of differences. Thus, the significance testing in statistics for detecting differences, including the one-classifier *Z*-test for evaluation and the two-classifier *Z*-test for comparison, may be needed [12-14, 23].

While designing tests, usually all scores of different systems were generated on a common set of speakers and speech segments. All systems tend to assign higher or lower scores to the same trials. Hence, all corresponding scores and thus the resulting CFs are positively correlated. Then, if the positive correlation coefficient of two CFs is not considered, the likelihood of detecting the statistical significance of a difference between the performance levels of the two systems can be wrongly reduced (see Section 5-B). To do so, in this article, a synchronized resampling algorithm is proposed (see Section 5-C).

The optimization of dataset with dependency into a two-layer data structure related to selection probability is shown in Section 2. The CF in the new ROC analysis is presented in Section 3. The two bootstrap algorithms for datasets with dependency and under i.i.d. assumption along with the number of bootstrap replications are explored in Section 4. The significance testing and a synchronized resampling algorithm for computing the correlation coefficient are presented in Section 5. The results with data dependency after dataset optimization of 24 speaker recognition systems[1, 2] (i.e., classifiers) used in SRE, the results assuming data are i.i.d., and their comparisons are all shown in Section 6. The results of significance testing for evaluation and comparison are shown in Section 7. Finally, the conclusions and discussion can be found in Section 8.

2 Optimize the dataset with dependency into a two-layer data structure

A. The notation for the two-layer data structure

Those target scores, known non-target scores, and unknown non-target scores generated using the same training speakers, i.e., the same subjects are dependent and thus are grouped into a target set, a known non-target set, and an unknown non-target set, respectively. Thus, the data structure has two layers: the first layer consists of sets, and the second layer consists of scores. In the following, let $\boldsymbol{X}$ and $\boldsymbol{S}$ denote sets, α score, and μ the number of scores in a set. The first subscript shows whether it is target (T), or known non-target (K), or unknown non-target (U); and the second and third numerate sets and scores in a set, respectively.

Suppose that there are m_T target sets, m_K known non-target sets, and m_U unknown non-target sets. Thus, the set $\boldsymbol{X}_T$ of all target sets, the set $\boldsymbol{X}_K$ of all known non-target sets, and the set $\boldsymbol{X}_U$ of all unknown non-target sets are expressed by

$$\boldsymbol{X}_i = \{\boldsymbol{S}_{i\,j} \mid j = 1, \ldots, m_i\},\ i \in \{T, K, U\}, \qquad (1)$$

where $\boldsymbol{S}_{T\,j}$, $\boldsymbol{S}_{K\,j}$, and $\boldsymbol{S}_{U\,j}$ are target sets, known non-target sets, and unknown non-target sets. In terms of its scores, each of them is expressed by

$$\boldsymbol{S}_{i\,j} = \{\alpha_{i\,j\,k} \mid k = 1, \ldots, \mu_{i\,j}\},\ j = 1, \ldots, m_i \text{ and } i \in \{T, K, U\}, \qquad (2)$$

where $\alpha_{T\,j\,k}$, $\alpha_{K\,j\,k}$, and $\alpha_{U\,j\,k}$ are target scores, known non-target scores, and unknown non-target scores, and $\mu_{i\,j}$ stands for the number of scores in the corresponding set.

[1] Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

[2] The speaker recognition systems are in general proprietary. It is NIST policy not to publicly associate speaker evaluation participant site names with system performance results, as we are evaluating the technology not participants.

Thus, in the set $\boldsymbol{X}_{\mathrm{T}}$, the m_{T} target sets $\boldsymbol{S}_{\mathrm{T}\,1}, \boldsymbol{S}_{\mathrm{T}\,2}, \ldots, \boldsymbol{S}_{\mathrm{T}\,m_{\mathrm{T}}}$ contain $\mu_{\mathrm{T}\,1}, \mu_{\mathrm{T}\,2}, \ldots, \mu_{\mathrm{T}\,m_{\mathrm{T}}}$ target scores $\{\alpha_{\mathrm{T}\,1\,1}, \alpha_{\mathrm{T}\,1\,2}, \ldots, \alpha_{\mathrm{T}\,1\,\mu_{\mathrm{T}1}}\}$, $\{\alpha_{\mathrm{T}\,2\,1}, \alpha_{\mathrm{T}\,2\,2}, \ldots, \alpha_{\mathrm{T}\,2\,\mu_{\mathrm{T}2}}\}$, …, $\{\alpha_{\mathrm{T}\,m_{\mathrm{T}}\,1}, \alpha_{\mathrm{T}\,m_{\mathrm{T}}\,2}, \ldots, \alpha_{\mathrm{T}\,m_{\mathrm{T}}\,\mu_{\mathrm{T}m_{\mathrm{T}}}}\}$, respectively. The analogous notations and descriptions can be applied to the known non-target sets and scores in sets, and the unknown non-target sets and scores in sets, respectively.

Hence, a two-layer data structure is constructed based on the data dependency. Scores in the same set are assumed to be conditionally independent as stated in Section 1. This structure preserves the data dependency while the bootstrap resampling takes place (see Section 4-A) [4, 11, 15]. Further, the two scores of two different speaker recognition systems with the same ordinal number of sets and the same ordinal number of scores in the sets were generated by matching the same two speech segments (i.e., objects) of the same two speakers (i.e., subjects), respectively. Usually this is how the test is designed. Thus, these two scores are correlated.

The set $\boldsymbol{T}$ (i.e., $\boldsymbol{X}_{\mathrm{T}}$) of all target scores, the set $\boldsymbol{K}$ (i.e., $\boldsymbol{X}_{\mathrm{K}}$) of all known non-target scores, and the set $\boldsymbol{U}$ (i.e., $\boldsymbol{X}_{\mathrm{U}}$) of all unknown non-target scores can also be denoted directly in terms of scores,

$$\begin{aligned} \boldsymbol{T} &= \{\alpha_{\mathrm{T}\,j\,k} \mid k = 1, \ldots, \mu_{\mathrm{T}\,j} \text{ and } j = 1, \ldots, m_{\mathrm{T}}\}, \\ \boldsymbol{K} &= \{\alpha_{\mathrm{K}\,j\,k} \mid k = 1, \ldots, \mu_{\mathrm{K}\,j} \text{ and } j = 1, \ldots, m_{\mathrm{K}}\}, \\ \boldsymbol{U} &= \{\alpha_{\mathrm{U}\,j\,k} \mid k = 1, \ldots, \mu_{\mathrm{U}\,j} \text{ and } j = 1, \ldots, m_{\mathrm{U}}\}. \end{aligned} \tag{3}$$

The sets $\boldsymbol{S}_{i\,j}$, $\boldsymbol{T}$, $\boldsymbol{K}$, and $\boldsymbol{U}$ are all viewed in the sense of a multiset, in which members may appear more than once. The empirical distribution is assumed for each of the observed scores in sets [11]. The total numbers of target scores, known non-target scores, and unknown non-target scores, i.e., N_{T}, N_{K}, and N_{U}, satisfy

$$N_i = \sum_{j=1}^{m_i} \mu_{i\,j}\,, \qquad \text{where } i \in \{\mathrm{T}, \mathrm{K}, \mathrm{U}\}. \tag{4}$$

B. The related selection probability

The nonparametric three-sample two-layer bootstrap method will be employed to estimate the SEs of measures in the new ROC analysis with data dependency. The two-layer resampling takes place randomly WR not only on the first layer of the data, i.e., the sets, but also on the second layer of the data, i.e., the scores in the sets. The scores in the same set are assumed to be conditionally independent as stated in Section 1.

Then, the probability for a score $\alpha_{i\,j\,k}$ in a set $\boldsymbol{S}_{i\,j}$ being selected in the two-layer data resampling is

$$\mathrm{P}_{2-\text{layer}}\left(\alpha_{i\,j\,k}\right) = \mathrm{P}\left(\boldsymbol{S}_{i\,j}\right) \times \mathrm{P}\left(\alpha_{i\,j\,k} \mid \boldsymbol{S}_{i\,j}\right) = \frac{1}{m_i} \times \frac{1}{\mu_{i\,j}}, \tag{5}$$
$$k = 1, \ldots, \mu_{i\,j},\ j = 1, \ldots, m_i \text{ and } i \in \{\mathrm{T}, \mathrm{K}, \mathrm{U}\}.$$

For target, known non-target, and unknown non-target scores, respectively, because different sets may contain different numbers of scores denoted by $\mu_{i\,j}$, this selection probability remains the same for all scores within a set, but varies from set to set. However, the nonparametric bootstrap method demands that the objects have equal probabilities to be selected in the random resampling for each type of scores [11].

If the numbers of scores in the target sets, the known non-target sets and the unknown non-target sets, i.e., $\mu_{T\,j}$ ($\mu_{K\,j}$, $\mu_{U\,j}$), j = 1, …, m_T (m_K, m_U), are all set to be equal to μ_T (μ_K, μ_U), respectively, then the probability for each target (known non-target, unknown non-target) score being selected will be 1 / N_T (1 / N_K, 1 / N_U) due to Eq. (4). Thus, the probability of selection for each target score, each known non-target score, and each unknown non-target score will be the same, respectively, in the two-layer resampling.

Moreover, this dataset optimization can ensure that the same numbers of target scores, known non-target scores, and unknown non-target scores, respectively, are resampled at different iterations while using the bootstrap method (see Section 4-A). Hence, such a structure of datasets can reduce the variance of the computation.

C. Optimize datasets

In the raw datasets, there were 41,897 target scores, 1,291,587 known non-target scores, and 407,827 unknown non-target scores. They were grouped into 394 target sets, 1,918 known non-target sets, and 1,918 unknown non-target sets, respectively. Different sets contain quite different numbers of scores, indicating that different training speakers were used different numbers of times.

Sets with equal number of scores can be easily constructed from sets with numbers of scores larger than or equal to a specified number, say, n_s. As stated in Section 2-A, the two scores of two different systems with the same ordinal number of sets and the same ordinal number of scores in the sets are correlated. Hence, if the score number of a set is greater than n_s, based on the randomness in samples of data, the first n_s scores in the set are selected instead of selecting n_s scores randomly without replacement from the set. This way can preserve the order of scores, and thus the correlation of two scores created by two different systems. This approach is useful for computing the correlation coefficient between two systems (see Section 5-C).

The above specified numbers for target sets, known non-target sets, and unknown non-target sets, respectively, are determined by the trial-and-error optimization so that the total numbers of target scores, known non-target scores, and unknown non-target scores can be as large as possible.

Thus, the raw datasets shown above were refined into 95 target score sets, 1,192 known non-target score sets, and 146 unknown non-target score sets, each of which contained 194 target scores, 511 known non-target scores, and 1,967 unknown non-target scores, respectively. So, the total numbers of resulting target scores, known non-target scores, and unknown non-target scores were 18,430, 609,112, and 287,182. In these new datasets, there are still tens or hundreds of thousands of scores [24].

This dataset optimization approximately halves the total number of scores in each category as shown above. However, because of the randomness in samples of data, such dataset optimization has little impact on the performance levels of classifiers (see Section 6) [6].

3 The CF in the new ROC analysis

The CF used in SRE is taken as an example in the new ROC analysis, which is defined to be an average of the two weighted sums of the probabilities of type I and type II errors at two thresholds correspondingly [7]. To make computation easier, the scores for a classifier are all converted to integers if they are not, and then expressed inclusively using the set **s** = {s_{min}, …, s_{max}}. While converting, as many decimal places of scores as possible are kept so that such a conversion does not result in loss of precision. Let f_i (s), s ∈ **s** and i ∈ {T, K, U}, denote the continuous probability density functions of target scores, known non-target scores, and unknown non-target scores. The three corresponding discrete probability distribution functions, denoted by P_i (s), s ∈ **s** and i ∈ {T, K, U}, are expressed as

$$\boldsymbol{P}_{\mathrm{i}} = \{\mathrm{P_i(s)} \mid \forall \mathrm{s} \in \mathbf{s} \text{ and } \sum_{\mathrm{s}\,=\,\mathrm{S_{min}}}^{\mathrm{S_{max}}} \mathrm{P_i(s)} = 1\},\ \ \mathrm{i} \in \{\mathrm{T,K,U}\}. \tag{6}$$

Assume that the two thresholds are $s_{min} < t_i < s_{max}$, i = 1, 2, where $t_1 < t_2$. The probabilities of type I errors for target scores at the two thresholds t_i are expressed,

$$\alpha_{\mathrm{T}}(t_{\mathrm{i}}) = \int_{-\infty}^{t_{\mathrm{i}}} \mathrm{f_T(s)\,ds} = \sum_{\mathrm{s}\,=\,\mathrm{S_{min}}}^{t_{\mathrm{i}}} \mathrm{P_T(s)} = 1 - \sum_{\mathrm{s}\,=\,t_{\mathrm{i}}+1}^{\mathrm{S_{max}}} \mathrm{P_T(s)},\ \ \mathrm{i} = 1,2, \tag{7}$$

and the probabilities of type II errors for known and unknown non-target scores at the two thresholds t_i are expressed,

$$\beta_{\mathrm{j}}(t_{\mathrm{i}}) = \int_{t_{\mathrm{i}}}^{+\infty} \mathrm{f_j\,(s)\,ds} = \sum_{\mathrm{s}=t_{\mathrm{i}}}^{\mathrm{S_{max}}} \mathrm{P_j}\,(s), \qquad \mathrm{j} \in \{\mathrm{K,U}\} \text{ and } \mathrm{i} = 1,2, \tag{8}$$

where P_T (S_{max} + 1) = 0 is assumed and the normalization in Eq. (6) is applied [3, 6]. Indeed, these error rates can be obtained by moving the score from the highest score s_{max} down to the thresholds t_i one score at a time to cumulate the probabilities of the three different scores, respectively. The probabilities of the corresponding scores at the thresholds t_i must be counted [23].

The two weighted sums of the probabilities of type I and type II errors at thresholds t_1 and t_2, respectively, are defined as [7],

$$\begin{gathered} W(t_{\mathrm{i}}) = \mathrm{C_{Miss}} \times \mathrm{P_{Target\,i}} \times [1 - \sum_{\mathrm{s}\,=\,t_{\mathrm{i}}+1}^{\mathrm{S_{max}}} \mathrm{P_T(s)}] + \\ \mathrm{C_{FalseAlarm}} \times (1 - \mathrm{P_{Target\,i}}) \times [\mathrm{P_{Known}} \times \sum_{\mathrm{s}\,=\,t_{\mathrm{i}}}^{\mathrm{S_{max}}} \mathrm{P_K(s)} + (1 - \mathrm{P_{Known}}) \times \sum_{\mathrm{s}\,=\,t_{\mathrm{i}}}^{\mathrm{S_{max}}} \mathrm{P_U(s)}]. \\ \mathrm{i} = 1,2. \end{gathered} \tag{9}$$

The parameters C_{Miss}, $C_{FalseAlarm}$, $P_{Target\ 1}$, $P_{Target\ 2}$, and P_{known} were set to be 1.0, 1.0, 0.01, 0.001, and 0.5, and the thresholds t_1 and t_2 were fixed values t_1 = ln (99) and t_2 = ln (999) [7].

Finally, the CF is defined as the average of these two weighted sums [7],

$$CF = \frac{W(t_1) + W(t_2)}{2}\ . \tag{10}$$

How to design the CF, how to set these parameters, and how to choose these thresholds are all out of the scope of this article. However, these issues would have no impact on how to statistically estimate the SEs of any measures in the new ROC analysis with data dependency using the data structure and the bootstrap algorithms employed in this article.

4 The two bootstrap algorithms and the number of bootstrap replications

A. The nonparametric three-sample two-layer bootstrap algorithm for datasets with dependency

The SE of the CF is statistically estimated using the nonparametric three-sample two-layer bootstrap method based on the prior extensive studies of bootstrap algorithms in ROC analysis on large datasets with data dependency (see Section 1) [3-4, 10-11, 17-22]. From here on, the superscript indices are used to numerate the resampling iterations.

Here is a function Random_WR_Resampling_Set that will be frequently called later,

```
function Random_WR_Resampling_Set (N, $\boldsymbol{\Gamma}$, $\boldsymbol{\Theta}$)
1: for i = 1 to N do
2:     select randomly WR an index $j \in \{1, \ldots, N\}$
3:     $\theta_i = \gamma_j$
4: end for
```

where **Γ** stands for a set of sets or a set of scores, N is its cardinality, **Θ** represents a new set of sets or scores with the same cardinality accordingly, and γ_j and θ_i are members of the sets **Γ** and **Θ**. This function can be applied to either a set of sets or a set of scores. It runs N iterations as shown from Step 1 to Step 4. In the i-th iteration, a member of the set **Γ** is randomly selected WR to be a member of a new set **Θ**, as indicated by Steps 2 and 3. Thus, N members (sets or scores) are randomly selected WR from the set **Γ** to form a new set **Θ**.

The nonparametric three-sample two-layer bootstrap method is conducted not only on the first layer of the new data structure where the resampling units are sets, but also subsequently on the second layer of the data in which the resampling units are scores within sets. Here is the algorithm.

Algorithm I (the nonparametric three-sample two-layer bootstrap)

```
1: for i = 1 to B do
2:     Random_WR_Resampling_Set ($m_T$, $\boldsymbol{X}_T$, $\boldsymbol{X}'^{\,i}_{T} = \{\boldsymbol{S}'^{\,i}_{Tj} \mid j = 1, \ldots, m_T\}$)
3:     for k = 1 to $m_T$ do
4:         Random_WR_Resampling_Set ($\mu_T$, $\boldsymbol{S}'^{\,i}_{Tk}$, $\boldsymbol{S}''^{\,i}_{Tk}$)
5:     end for

6:     Random_WR_Resampling_Set ($m_K$, $\boldsymbol{X}_K$, $\boldsymbol{X}'^{\,i}_{K} = \{\boldsymbol{S}'^{\,i}_{Kj} \mid j = 1, \ldots, m_K\}$)
7:     for k = 1 to $m_K$ do
8:         Random_WR_Resampling_Set ($\mu_K$, $\boldsymbol{S}'^{\,i}_{Kk}$, $\boldsymbol{S}''^{\,i}_{Kk}$)
9:     end for

10:    Random_WR_Resampling_Set ($m_U$, $\boldsymbol{X}_U$, $\boldsymbol{X}'^{\,i}_{U} = \{\boldsymbol{S}'^{\,i}_{Uj} \mid j = 1, \ldots, m_U\}$)
```

```
11:   for k = 1 to $m_U$ do
12:       Random_WR_Resampling_Set ($\mu_U$, $\boldsymbol{S}'^{\,i}_{U k}$, $\boldsymbol{S}''^{\,i}_{U k}$)
13:   end for

14:   $\boldsymbol{X}''^{\,i}_{T} = \{\boldsymbol{S}''^{\,i}_{T j} \mid j = 1, \ldots, m_T\}$ and $\boldsymbol{X}''^{\,i}_{K} = \{\boldsymbol{S}''^{\,i}_{K j} \mid j = 1, \ldots, m_K\}$ and
      $\boldsymbol{X}''^{\,i}_{U} = \{\boldsymbol{S}''^{\,i}_{U j} \mid j = 1, \ldots, m_U\}$ => $\widehat{W}^{i}(t_1)$ and $\widehat{W}^{i}(t_2)$ => $\widehat{\mathrm{CF}}^{i}$
15: end for
16: $\{\widehat{\mathrm{CF}}^{i} \mid i = 1, \ldots, B\}$ => $\widehat{\mathrm{SE}}$ and $(\widehat{Q}\,(\alpha / 2), \widehat{Q}\,(1 - \alpha / 2))$
17: end
```

where B is the number of bootstrap replications, the set $\boldsymbol{X}_T$ of all target sets, the set $\boldsymbol{X}_K$ of all known non-target sets, and the set $\boldsymbol{X}_U$ of all unknown non-target sets are expressed in Eq. (1), and m_T, m_K, and m_U are the corresponding cardinalities.

This algorithm runs B times, as shown from Steps 1 to 15. In the i-th iteration, as indicated by Steps 2, 6, and 10, the function Random_WR_Resampling_Set is applied to the first layer of the data, i.e., a set of sets. That is, m_T target sets, m_K known non-target sets, and m_U unknown non-target sets are randomly selected WR from the set $\boldsymbol{X}_T$, $\boldsymbol{X}_K$, and $\boldsymbol{X}_U$ to form a new set $\boldsymbol{X}'^{\,i}_{T} = \{\boldsymbol{S}'^{\,i}_{T j} \mid j = 1, \ldots, m_T\}$, $\boldsymbol{X}'^{\,i}_{K} = \{\boldsymbol{S}'^{\,i}_{K j} \mid j = 1, \ldots, m_K\}$, and $\boldsymbol{X}'^{\,i}_{U} = \{\boldsymbol{S}'^{\,i}_{U j} \mid j = 1, \ldots, m_U\}$, respectively.

After the first-layer resampling, the same function is applied to the second layer of the data, i.e., the scores in the sets. From Steps 3 to 5, m_T iterations take place. In the k-th iteration, μ_T target scores are randomly selected WR from the target set $\boldsymbol{S}'^{\,i}_{T k}$, which is the k-th new target set from the first-layer resampling, to form the k-th new target set $\boldsymbol{S}''^{\,i}_{T k}$ of the second-layer resampling. The analogous interpretation is applied to known non-target scores in the known non-target set $\boldsymbol{S}'^{\,i}_{K k}$ from Steps 7 to 9, and to unknown non-target scores in the unknown non-target set $\boldsymbol{S}'^{\,i}_{U k}$ from Steps 11 to 13.

At Step 14, all target scores in the new set $\boldsymbol{X}''^{\,i}_{T} = \{\boldsymbol{S}''^{\,i}_{T j} \mid j = 1, \ldots, m_T\}$, all known non-target scores in the new set $\boldsymbol{X}''^{\,i}_{K} = \{\boldsymbol{S}''^{\,i}_{K j} \mid j = 1, \ldots, m_K\}$, and all unknown non-target scores in the new set $\boldsymbol{X}''^{\,i}_{U} = \{\boldsymbol{S}''^{\,i}_{U j} \mid j = 1, \ldots, m_U\}$ are employed to generate the i-th estimates $\widehat{W}^{i}(t_1)$ and $\widehat{W}^{i}(t_2)$ in Eq. (9), by moving one integer score at a time from the highest score s_{max} down to the larger threshold t_2 and then to the smaller threshold t_1 to calculate the corresponding cumulative probabilities in Eqs. (7) and (8). Thereafter, the *i*-th bootstrap replication of the CF, i.e., $\widehat{\mathrm{CF}}^{i}$ is computed using Eq. (10).

Finally, at Step 16, from the bootstrap distribution $\{\widehat{\mathrm{CF}}^{i} \mid i = 1, \ldots, B\}$, the $\widehat{\mathrm{SE}}$ of the CF is estimated by its sample standard deviation, and the $(1 - \alpha) \times 100\%$ CI, i.e., $(\widehat{Q}\,(\alpha/2), \widehat{Q}\,(1 - \alpha/2))$ at the significance level α is estimated by its $\alpha/2 \times 100\%$ and $(1 - \alpha/2) \times 100\%$ quantiles [11]. The sample quantile is obtained by inverting the empirical distribution function with averaging at discontinuities [25]. If the 95% CI is of interest, α is set to be 0.05.

With the optimized data structure shown in Section 2-C, the same numbers of target scores, known non-target scores, and unknown non-target scores, respectively, are obtained in Step 14 at different iterations of the algorithm while computing the bootstrap replication of the CF. This can reduce the variance of the computation.

B. The nonparametric three-sample bootstrap algorithm for datasets with the i.i.d. assumption

Algorithm I can be modified to be the nonparametric three-sample bootstrap algorithm so that it can be applied to datasets with the i.i.d. assumption, if Steps 2 to 13 are replaced by

Random_WR_Resampling_Set ($\mathrm{N_T}$, $\boldsymbol{T}$, $\mathbf{\Theta^i}$)
Random_WR_Resampling_Set ($\mathrm{N_K}$, $\boldsymbol{K}$, $\mathbf{\Xi^i}$)
Random_WR_Resampling_Set ($\mathrm{N_U}$, $\boldsymbol{U}$, $\mathbf{\Psi^i}$)

where $\boldsymbol{T}$, $\boldsymbol{K}$ and $\boldsymbol{U}$ are the original score sets as shown in Eq. (3). That is, in the *i*-th iteration by calling the function Random_WR_Resampling_Set three times, $\mathrm{N_T}$ target scores, $\mathrm{N_K}$ known non-target scores, and $\mathrm{N_U}$ unknown non-target scores are randomly selected WR from the sets $\boldsymbol{T}$, $\boldsymbol{K}$, and $\boldsymbol{U}$, respectively, to constitute new sets $\mathbf{\Theta^i}$, $\mathbf{\Xi^i}$, and $\mathbf{\Psi^i}$. Thereafter, Step 14 is replaced by

$\mathbf{\Theta^i}$, $\mathbf{\Xi^i}$, and $\mathbf{\Psi^i}$ => $\widehat{W}^{\mathrm{i}}(t_1)$ and $\widehat{W}^{\mathrm{i}}(t_2)$ => $\widehat{\mathrm{CF}}^{\mathrm{i}}$

meaning that all scores in these three new sets $\mathbf{\Theta^i}$, $\mathbf{\Xi^i}$, and $\mathbf{\Psi^i}$ are employed to generate the *i*-th bootstrap replication of the CF, i.e., $\widehat{\mathrm{CF}}^{\mathrm{i}}$, using Eqs. (7) through (10).

C. The number of bootstrap replications

The remaining issue is to determine the appropriate number of bootstrap replications B. In our ROC analysis for decision making of classifiers, data samples of scores are over tens of thousands and have no parametric model to fit, the statistics of interest are mostly probabilities or a weighted sum of probabilities, and data dependency may be involved. Thus, to reduce the bootstrap variance and ensure the computation accuracy, the bootstrap variabilities were re-studied, which took months of CPU time, and the appropriate number of bootstrap replications B under the above circumstances was determined to be 2,000 [11, 18-22].

5 Significance testing and an algorithm for computing the correlation coefficient

Based on the observations in Section 6-A, it can be assumed that the CF is normally distributed. Hence, while evaluating the performance level of a classifier against a hypothesized performance criterion, the one-classifier two-tailed significance testing is employed. While comparing the performance levels of two classifiers, the two-classifier two-tailed significance testing is used, and the correlation coefficient of two classifiers' CFs must be precalculated [3, 5, 12-14].

In this article, the two-tailed rather than the one-tailed significance testing is adopted, because there is no reason to believe *a priori* that one is better than the other; moreover, the two-tailed test is generally more conservative than the one-tailed test in the sense that the former is more difficult to reject the null hypothesis for a given significance level [26]. For all tests, the significance level was conventionally chosen to be 5% [11].

A. One-classifier two-tailed significance testing for evaluation

Let D denote the CF of a speaker recognition system and μ_o represent the hypothesized performance criterion. Then, the null and alternative hypotheses are

$$H_o : D = \mu_o$$
$$H_a : D \neq \mu_o \;. \tag{11}$$

Based on the normality assumption, the Z statistic is

$$Z = \frac{\widehat{D} \; - \; \mu_o}{\mathrm{SE}\;(\widehat{D})} \tag{12}$$

where $\widehat{D}$ is the estimator of CF and SE $(\widehat{D})$ stands for its SE. The Z statistic is subject to the standard normal distribution with zero expectation and a variance of one [23]. The two-tailed p-value can be obtained from the Z score using R [27].

For the one-classifier two-tailed significance testing, if the 95% CI of a classifier's CF contains μ_o, i.e., $\widehat{D} - Z_{\alpha/2}\mathrm{SE}(\widehat{D}) < \mu_o < \widehat{D} + Z_{\alpha/2}\mathrm{SE}(\widehat{D})$, where $Z_{\alpha/2} = 1.96$ for $\alpha = 5\%$, then Z in Eq. (12) satisfies $-Z_{\alpha/2} < Z < Z_{\alpha/2}$, meaning that the null hypothesis H_o cannot be rejected with at least 5% significance level, i.e., the difference between the performance level of the classifier and the value μ_o is not statistically significant [23]. Otherwise, H_o is rejected with at most 5% significance level.

Hence, for such an evaluation, if dealt with is the case in which the smaller the CF, the more accurate the classifier, then a classifier passes the test only if its 95% CI is below the horizontal line at the performance criterion μ_o and fails the test otherwise. Certainly, the approach of merely using the relative position between 95% CI and the line at μ_o cannot provide any quantitative information concerning the statistical significance of differences.

B. Two-classifier two-tailed significance testing for comparison

Let D_1 and D_2 denote the CFs of two speaker recognition systems. Then, the null and alternative hypothesis are

$$H_o : D_1 = D_2$$
$$H_a : D_1 \neq D_2 \;. \tag{13}$$

Based on the normality assumption, the general Z statistic for the two-classifier significance testing is expressed as

$$Z = \frac{\widehat{D}_1 \; - \; \widehat{D}_2}{\sqrt{\mathrm{SE}^2(\widehat{D}_1) \; + \; \mathrm{SE}^2(\widehat{D}_2) \; - \; 2\,r\,\mathrm{SE}(\widehat{D}_1)\,\mathrm{SE}(\widehat{D}_2)}} \tag{14}$$

where $\widehat{D}_1$ and $\widehat{D}_2$ are two estimators of CFs of two systems, $\mathrm{SE}(\widehat{D}_1)$ and $\mathrm{SE}(\widehat{D}_2)$ stand for their SEs, and r is the correlation coefficient between $\widehat{D}_1$ and $\widehat{D}_2$. Also, the Z statistic is subject to the standard normal distribution with zero expectation and a variance of one [14, 23]. The two-tailed p-value can be obtained from the Z score using R [27].

Here is an important issue. If the two CFs are positively correlated and the correlation coefficient r is not taken into consideration, it will make the denominator of Eq. (14) larger than it should be and the Z score correspondingly smaller than it should be. This will then wrongly reduce the likelihood of detecting the statistical significance of difference between the performance levels of two systems. As a result, the correlation coefficient *r* must be calculated.

C. An algorithm for computing the correlation coefficient

As indicated in Sections 1, 2-A and 2-C, the two scores of two different systems with the same ordinal number of sets and the same ordinal number of scores in the sets were created by matching the speakers and speech segments with the same ID numbers, for all target scores, known non-target scores and unknown non-target scores. All systems tend to assign higher or lower scores to the same trials. Thus, these two scores of the two systems co-vary positively. Indeed, this is usually the way that test is designed. Hence, the following synchronized resampling algorithm is valid if a test is planned in this way. Further, the CF of a system is computed from the three distributions of these scores. Thus, the two CFs of any two speaker recognition systems are positively correlated. Here is a synchronized resampling function to pick up the correlated scores.

```
function Synchronized_WR_Random_Resampling_Set (M, $\boldsymbol{\Gamma}^A$, $\boldsymbol{\Theta}^A$, $\boldsymbol{\Gamma}^B$, $\boldsymbol{\Theta}^B$)
for j = 1 to M do
    select randomly WR an index k $\in$ {1, …, M}
    $\theta^A_j = \gamma^A_k$
    $\theta^B_j = \gamma^B_k$
end for
```

where $\boldsymbol{\Gamma}^A$, $\boldsymbol{\Theta}^A$, $\boldsymbol{\Gamma}^B$, and $\boldsymbol{\Theta}^B$ represent a set of sets or a set of scores created by Systems A and B, respectively. In other words, this function can be applied to either two sets of sets or two sets of scores $\boldsymbol{\Gamma}^A$ and $\boldsymbol{\Gamma}^B$ to generate two corresponding sets $\boldsymbol{\Theta}^A$ and $\boldsymbol{\Theta}^B$. M is their cardinalities. γ^A_k, θ^A_j, γ^B_k, and θ^B_j are their members.

The two sets or scores γ^A_k and γ^B_k of two Systems A and B have the same ordinal number k. Hence, this function can synchronize the random resampling WR both in a set $\boldsymbol{\Gamma}^A$ of System A and in a set $\boldsymbol{\Gamma}^B$ of System B so that the kth sets or scores in $\boldsymbol{\Gamma}^A$ and $\boldsymbol{\Gamma}^B$, respectively, are chosen to be the jth sets or scores in $\boldsymbol{\Theta}^A$ and $\boldsymbol{\Theta}^B$. Hence, eventually the correlated scores generated by these two systems A and B are selected.

Here is a synchronized resampling algorithm that is applied to two Systems A and B simultaneously to pick up the sets with the same ordinal number and eventually the correlated scores by calling the above function, and then computes the correlation coefficient of the two CFs.

Algorithm II (Compute correlation coefficient)

```
for i = 1 to N do
    Synchronized_WR_Random_Resampling_Set ($m_T$, $\boldsymbol{X}^A_T$, $\boldsymbol{X}^{A\,\prime\,i}_{T} = \{\boldsymbol{S}^{A\,\prime\,i}_{Tj} \mid j = 1, \ldots, m_T\}$,
                                           $\boldsymbol{X}^B_T$, $\boldsymbol{X}^{B\,\prime\,i}_{T} = \{\boldsymbol{S}^{B\,\prime\,i}_{Tj} \mid j = 1, \ldots, m_T\}$)
    for k = 1 to $m_T$ do
        Synchronized_WR_Random_Resampling_Set ($\mu_T$, $\boldsymbol{S}^{A\,\prime\,i}_{Tk}$, $\boldsymbol{S}^{A\,\prime\prime\,i}_{Tk}$, $\boldsymbol{S}^{B\,\prime\,i}_{Tk}$, $\boldsymbol{S}^{B\,\prime\prime\,i}_{Tk}$)
    end for
```

```
6:     Synchronized_WR_Random_Resampling_Set ($m_K$, $\boldsymbol{X}^{A}_{K}$, $\boldsymbol{X}^{A}{}'_{K}{}^{i} = \{\boldsymbol{S}^{A}{}'_{Kj}{}^{i} \mid j = 1, \ldots, m_K\}$,
                                           $\boldsymbol{X}^{B}_{K}$, $\boldsymbol{X}^{B}{}'_{K}{}^{i} = \{\boldsymbol{S}^{B}{}'_{Kj}{}^{i} \mid j = 1, \ldots, m_K\}$)
7:     for k = 1 to $m_K$ do
8:         Synchronized_WR_Random_Resampling_Set ($\mu_K$, $\boldsymbol{S}^{A}{}'_{Kk}{}^{i}$, $\boldsymbol{S}^{A}{}''_{Kk}{}^{i}$, $\boldsymbol{S}^{B}{}'_{Kk}{}^{i}$, $\boldsymbol{S}^{B}{}''_{Kk}{}^{i}$)
9:     end for

10:    Synchronized_WR_Random_Resampling_Set ($m_U$, $\boldsymbol{X}^{A}_{U}$, $\boldsymbol{X}^{A}{}'_{U}{}^{i} = \{\boldsymbol{S}^{A}{}'_{Uj}{}^{i} \mid j = 1, \ldots, m_U\}$,
                                           $\boldsymbol{X}^{B}_{U}$, $\boldsymbol{X}^{B}{}'_{U}{}^{i} = \{\boldsymbol{S}^{B}{}'_{Uj}{}^{i} \mid j = 1, \ldots, m_U\}$)
11:    for k = 1 to $m_U$ do
12:        Synchronized_WR_Random_Resampling_Set ($\mu_U$, $\boldsymbol{S}^{A}{}'_{Uk}{}^{i}$, $\boldsymbol{S}^{A}{}''_{Uk}{}^{i}$, $\boldsymbol{S}^{B}{}'_{Uk}{}^{i}$, $\boldsymbol{S}^{B}{}''_{Uk}{}^{i}$)
13:    end for

14:    $\boldsymbol{X}^{A}{}''_{T}{}^{i} = \{\boldsymbol{S}^{A}{}''_{Tj}{}^{i} \mid j = 1, \ldots, m_T\}$ and $\boldsymbol{X}^{A}{}''_{K}{}^{i} = \{\boldsymbol{S}^{A}{}''_{Kj}{}^{i} \mid j = 1, \ldots, m_K\}$ and
       $\boldsymbol{X}^{A}{}''_{U}{}^{i} = \{\boldsymbol{S}^{A}{}''_{Uj}{}^{i} \mid j = 1, \ldots, m_U\}$ => $\widehat{W}^{A\,i}(t_1)$ and $\widehat{W}^{A\,i}(t_2)$ => $\widehat{\mathrm{CF}}^{A\,i}$
15:    $\boldsymbol{X}^{B}{}''_{T}{}^{i} = \{\boldsymbol{S}^{B}{}''_{Tj}{}^{i} \mid j = 1, \ldots, m_T\}$ and $\boldsymbol{X}^{B}{}''_{K}{}^{i} = \{\boldsymbol{S}^{B}{}''_{Kj}{}^{i} \mid j = 1, \ldots, m_K\}$ and
       $\boldsymbol{X}^{B}{}''_{U}{}^{i} = \{\boldsymbol{S}^{B}{}''_{Uj}{}^{i} \mid j = 1, \ldots, m_U\}$ => $\widehat{W}^{B\,i}(t_1)$ and $\widehat{W}^{B\,i}(t_2)$ => $\widehat{\mathrm{CF}}^{B\,i}$
16: end for

17: $\{\widehat{\mathrm{CF}}^{A\,i} \mid i = 1, \ldots, N\}$ and $\{\widehat{\mathrm{CF}}^{B\,i} \mid i = 1, \ldots, N\}$ => the correlation coefficient $\hat{\gamma}_{C}^{AB}$
18: end
```

where the number of iterations N is set to be 2,000, based on our extensive studies of bootstrap variability in ROC analysis with or without data dependency for large datasets [20-22].

From Step 1 to 16, this algorithm runs N iterations. In the i-th iteration, in Step 2, the above function is applied simultaneously to the first layers of the two systems' data, i.e., the set $\boldsymbol{X}^{A}_{T}$ of all target sets of System A and the set $\boldsymbol{X}^{B}_{T}$ of all target sets of System B (see Eq. (1)), so that the two target sets with the same ordinal number in the two systems' data, respectively, are randomly selected WR to form two new sets $\boldsymbol{X}^{A}{}'_{T}{}^{i}$ and $\boldsymbol{X}^{B}{}'_{T}{}^{i}$.

Then, from Step 3 to 5, this function is applied m_T times simultaneously to the second layer of the data created by Systems A and B. Hence, the target scores in set $\boldsymbol{S}^{A}{}'_{Tk}{}^{i}$ of System A and the target scores in set $\boldsymbol{S}^{B}{}'_{Tk}{}^{i}$ of System B with the same ordinal number in the two systems' data are randomly chosen WR. Thereafter, these correlated scores constitute two new sets of target scores $\boldsymbol{S}^{A}{}''_{Tk}{}^{i}$ and $\boldsymbol{S}^{B}{}''_{Tk}{}^{i}$, respectively.

The analogous interpretation can be applied to known non-target sets and known non-target scores in sets from Step 6 to 9, and unknown non-target sets and unknown non-target scores in sets from Step 10 to 13.

In Step 14, for System A, the target scores in set $\boldsymbol{X}^{A}{}''_{T}{}^{i}$, the known non-target scores in set $\boldsymbol{X}^{A}{}''_{K}{}^{i}$, and the unknown non-target scores in set $\boldsymbol{X}^{A}{}''_{U}{}^{i}$ produce the i-th bootstrap replication of the CF, i.e., $\widehat{\mathrm{CF}}^{A\,i}$. In Step 15, the same explanation is applied to System B. Thus, the correlated pairs $\widehat{\mathrm{CF}}^{A\,i}$ and $\widehat{\mathrm{CF}}^{B\,i}$ are calculated from the correlated scores.

Finally, in Step 17, after N iterations, from these N pairs of correlated bootstrap replications of CFs, a correlation coefficient $\hat{\gamma}_{C}^{AB}$ of CFs is estimated [23]. Because this is a stochastic process, an average of such 20 correlated coefficients will be used in Eq. (14) [3, 5].

6 Results with data dependency, results with i.i.d. assumption, and comparisons

A. Results with data dependency after dataset optimization

sys	CF SE (relative error) 95% CI	sys	CF SE (relative error) 95% CI	sys	CF SE (relative error) 95% CI	sys	CF SE (relative error) 95% CI
1	0.002113 0.000184 (17.07%) (0.001776, 0.002488)	7	0.002924 0.000230 (15.42%) (0.002470, 0.003376)	13	0.003392 0.000233 (13.45%) (0.002942, 0.003846)	19	0.004174 0.000244 (11.47%) (0.003707, 0.004661)
2	0.002164 0.000198 (17.93%) (0.001785, 0.002565)	8	0.002960 0.000244 (16.15%) (0.002503, 0.003442)	14	0.003761 0.000223 (11.64%) (0.003330, 0.004201)	20	0.004417 0.000132 (5.84%) (0.004168, 0.004670)
3	0.002505 0.000214 (16.74%) (0.002089, 0.002924)	9	0.003074 0.000184 (11.72%) (0.002733, 0.003442)	15	0.003812 0.000246 (12.64%) (0.003338, 0.004284)	21	0.004660 0.000125 (5.25%) (0.004407, 0.004897)
4	0.002624 0.000182 (13.63%) (0.002269, 0.002988)	10	0.003110 0.000196 (12.34%) (0.002735, 0.003502)	16	0.004088 0.000209 (10.01%) (0.003710, 0.004533)	22	0.004826 0.000090 (3.64%) (0.004649, 0.004994)
5	0.002802 0.000214 (14.98%) (0.002384, 0.003228)	11	0.003149 0.000204 (12.68%) (0.002741, 0.003530)	17	0.004131 0.000243 (11.53%) (0.003655, 0.004595)	23	0.004854 0.000096 (3.86%) (0.004664, 0.005037)
6	0.002862 0.000252 (17.28%) (0.002370, 0.003368)	12	0.003358 0.000222 (12.93%) (0.002916, 0.003801)	18	0.004135 0.000189 (8.95%) (0.003743, 0.004495)	24	0.004934 0.000176 (6.98%) (0.004602, 0.005283)

Table 1 The estimated CFs, SEs (relative errors), and 95% CIs of 24 speaker recognition systems listed in the descending order of their performance levels. The SE of CF was estimated using the nonparametric three-sample two-layer bootstrap algorithm by taking account of data dependency.

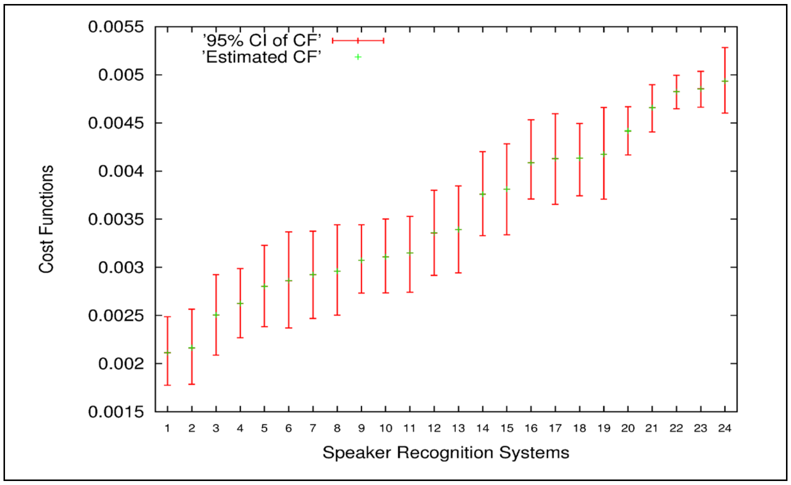

Figure 2 The CFs and 95% CIs of 24 speaker recognition systems where the SEs of CFs were those in Table 1.

Twenty-four speaker recognition systems were investigated. Their estimated CFs, SEs (relative errors), and 95% CIs are shown in Table 1. The smaller the CF, the more accurate the system.

After dataset optimization by taking account of the data dependency, the CF was calculated using Eqs. (7) through (10), the SE was computed using the nonparametric three-sample two-layer bootstrap algorithm, and the relative error was estimated by (1.96 x SE / CF).

Figure 2 shows the CFs with 95% CIs for 24 systems. Some 95% CIs do overlap, in which the statistical significance testing using the SEs estimated here and the correlation coefficients computed later must be carried out to differentiate the performance levels of the systems (see Section 7) [3-5].

The relative errors of 24 systems vary between 3.64% and 17.93%. In general, the more accurate the systems, i.e., the smaller the CFs, the larger the relative errors. For 3/4 of the systems, their relative errors are greater than 10.00%. All these demonstrate that it is inappropriate to ignore the SEs while dealing with the performance levels of classifiers.

Further, as stated in Section 4-A, the 95% CIs of the CFs in Table 1 were all computed using the quantiles of the bootstrap distributions. If the distribution of 2,000 bootstrap replications of CF was assumed to be normal, they could also be calculated using [CF – 1.96 x SE, CF + 1.96 x SE]. It was found that these two sets of 95% CIs matched very well. If the relative error is defined to be $|P_N – P_Q| / P_Q$, where P_N is one of the two end points of the 95% CI calculated by assuming normal distribution and P_Q is the corresponding end point of the 95% CI computed using the quantile definition, then among 48 relative errors for 24 systems, only one is 1.33% and all others vary between 0.01% and 0.85%.

In addition, the Shapiro-Wilk normality test [27] was conducted on the 2,000 bootstrap replications of CF for each system. It was observed that the *p*-values of 20 out of 24 systems (i.e., 83%) were greater than 5%. So, it is suggested that the CF be regarded as approximately normally distributed.

B. Results with i.i.d. assumption without dataset optimization

sys	CF / SE 95% CI	sys	CF / SE 95% CI	sys	CF / SE 95% CI	sys	CF / SE 95% CI
1	0.001710 / 0.000015 (0.001681, 0.001740)	7	0.002240 / 0.000017 (0.002206, 0.002271)	13	0.002623 / 0.000016 (0.002590, 0.002653)	19	0.003499 / 0.000023 (0.003456, 0.003543)
2	0.001724 / 0.000018 (0.001688, 0.001760)	8	0.002225 / 0.000021 (0.002183, 0.002267)	14	0.003196 / 0.000025 (0.003146, 0.003246)	20	0.004136 / 0.000015 (0.004106, 0.004164)
3	0.002066 / 0.000018 (0.002031, 0.002101)	9	0.002667 / 0.000018 (0.002631, 0.002703)	15	0.002990 / 0.000020 (0.002951, 0.003030)	21	0.004460 / 0.000012 (0.004436, 0.004483)
4	0.002411 / 0.000016 (0.002380, 0.002444)	10	0.002635 / 0.000019 (0.002597, 0.002674)	16	0.003369 / 0.000023 (0.003322, 0.003416)	22	0.004458 / 0.000010 (0.004437, 0.004477)
5	0.002432 / 0.000016 (0.002401, 0.002463)	11	0.002396 / 0.000015 (0.002366, 0.002427)	17	0.003476 / 0.000021 (0.003436, 0.003518)	23	0.004665 / 0.000011 (0.004644, 0.004686)
6	0.002297 / 0.000024 (0.002252, 0.002346)	12	0.002925 / 0.000022 (0.002882, 0.002969)	18	0.003816 / 0.000012 (0.003791, 0.003841)	24	0.004311 / 0.000026 (0.004259, 0.004360)

Table 2 The CFs, SEs, and 95% CI of 24 speaker recognition systems numbered in the same order as in Table 1. The SE of CF was computed by applying the nonparametric three-sample bootstrap method on all the raw data assuming that data were i.i.d.

Table 2 shows the CFs, SEs, and 95% CIs of 24 speaker recognition systems, in which the CFs were calculated using Eqs. (7) through (10) applied on all the raw data (see Section 2-C), and the

SEs were estimated using the nonparametric three-sample bootstrap method with the assumption that all the raw data were i.i.d. The systems were numbered in the same order as in Table 1.

C. Comparisons of the two results in terms of both CFs and SEs

C.1. The performance level (i.e., CF)

In Table 2, the total numbers of target scores, known non-target scores, and unknown non-target scores were 41,897, 1,291,587, and 407,827. In Table 1, the total numbers of the corresponding scores used were 18,430, 609,112, and 287,182, which were almost cut in half (see Section 2).

If the absolute relative change in performance level due to dataset optimization is defined as $|CF_{after} – CF_{before}| / CF_{before}$, where CF_{before} and CF_{after} are the CFs computed before and after optimization, then there are 25%, 50% and 25% of all cases, in which the changes in CFs are less than 10%, between 10% and 25%, and between 25% and 33%, respectively. For the best case with 4% absolute relative change, the CF was 0.004665 before dataset optimization and 0.004854 afterwards. For the worst case with 33% change, the corresponding CFs were 0.002225 and 0.002960.

All these indicate that if the data size is tens of thousands or larger in ROC analysis, reducing data size even by half due to dataset optimization has little impact on the performance levels, which is consistent with the conclusion reached by means of the Chebyshev's inequality [24]. Such dataset optimization does not alter the relative positions among different distributions of scores, which has impact on the performance levels in ROC analysis [6].

While creating datasets that involve data dependency, if datasets are constructed according to the rules stated in Section 2-B at the very beginning rather than using the optimization later as described in Section 2-C, then not only can all computational advantages described in Section 2-B be achieved, but resources can also be saved.

C.2. The SE of measure

For all systems, the SEs in Table 2, which were estimated using the nonparametric three-sample bootstrap algorithm on all raw data under the i.i.d. assumption, are correspondingly much smaller (about 10 times) than those in Table 1, which were computed using the nonparametric three-sample two-layer bootstrap algorithm on about half of the raw data grouped into score sets after taking account of data dependency. For instance, for System 1, the estimated SE is 0.000015 in Table 2, but 0.000184 in Table 1.

The estimated SE is inversely proportional to the square root of the data size n, i.e., SE ~ $1 / \sqrt{n}$. The data size used in Table 1 is about half of the data size employed in Table 2 as shown in Section 6-C.1. Even taking account of this effect of data size, i.e., multiplying the estimated SEs in Table 2 by a factor of $\sqrt{2}$, the resultant SEs are still far smaller than those in Table 1. For instance, for System 1, the resultant SE is 0.000015 x $\sqrt{2}$ = 0.000021, which is still about 8.67 times smaller than 0.000184 in Table 1.

Hence, the SEs of CF are underestimated without taking account of the data dependency. This is consistent with the conclusions reached in our previous studies, in which the rationale regarding the different impacts of different bootstrap methods on the estimated SEs was investigated [4]. As a conclusion, if the same subjects are used multiple times while creating datasets, to estimate SEs, the data dependency must be taken into consideration.

7 Results of significance testing for evaluation and comparison

sys	CF	SE	95% CI
A	0.002113	0.000184	(0.001776, 0.002488)
B	0.002164	0.000198	(0.001785, 0.002565)
C	0.002802	0.000214	(0.002384, 0.003228)
D	0.002960	0.000244	(0.002503, 0.003442)
E	0.003761	0.000223	(0.003330, 0.004201)

Table 3 The CFs, SEs, and 95% CIs of Systems A through E selected from Table 1.

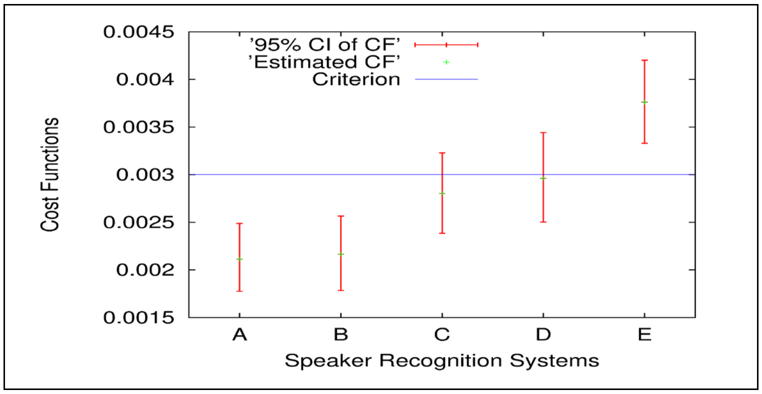

Figure 3 The CFs and 95% CIs of Systems A through E along with the hypothesized performance criterion set to be 0.003.

To demonstrate clearly how the significance testing works, five speaker recognition systems 1, 2, 5, 8, and 14 in Section 6 were chosen and renamed to be Systems A through E. Their estimated CFs, SEs, and 95% CIs can be found in Table 1, and are also shown in Table 3 for convenience. Figure 3 depicts their CFs and 95% CIs along with the hypothesized performance criterion set to be 0.003, which is for evaluation.

The issues of evaluation and comparison of classifiers can be dealt with intuitively to some extent by just using 95% CIs. However, it cannot provide any quantitative information as far as the statistical significance of differences is concerned. In Figure 3, some 95% CIs overlap considerably, but some overlap a little. To this end, the statistical significance testing can be utilized. Certainly, to conduct the two-classifier significance testing, the correlation coefficient must be computed.

A. One-system evaluation

system	A	B	C	D	E
p-value	0.0000	0.0000	0.3558	0.8703	0.0007

Table 4 The one-classifier two-tailed *p*-values for Systems A through E.

For illustration, the performance criterion μ_o was set to be 0.003 as depicted in Figure 3. After applying Eq. (12) and using R [27], the one-classifier two-tailed *p*-values were computed as shown in Table 4. For Systems A, B, and E, the two-tailed *p*-values are much less than 5%, some of which are too small to be shown. It indicates that the null hypothesis $H_o : D = \mu_o$ is rejected in favor of the alternative hypothesis $H_a : D \neq \mu_o$. Hence, the CFs of Systems A and B (System E) are significantly smaller (greater) than 0.003. This is consistent with the observations in Figure 3, where the 95% CIs of Systems A and B (System E) are below (above) the line at 0.003. As stated in Section 6-A, the smaller the CF, the more accurate the speaker recognition system. Thus, Systems A and B pass the test, but System E fails the test.

For Systems C and D, the two-tailed *p*-values are 0.3558 and 0.8703, which are both greater than 5%. It suggests that the null hypothesis $H_o : D = \mu_o$ is accepted at the achieved significance level 36% and 87%, respectively [11]. This is also consistent with the observations in Figure 3, where both of their 95% CIs contain 0.003. Hence, the difference between the performance level of a system and the performance criterion μ_o is not statistically significant, and both systems fail the test. Indeed, all conclusions reached here are consistent with the statements made in Section 5-A.

B. Two-system comparison

sys	B	C	D	E
A	0.839104	0.817586	0.807100	0.733167
B		0.824137	0.826598	0.750948
C			0.820434	0.804367
D				0.848460

Table 5 The average correlation coefficients of CFs of two systems out of 20 runs among Systems A through E.

To determine whether the difference between the performance levels of two classifiers is statistically significant, the two-classifier significance testing must be carried out. To do so, the correlation coefficient of CFs of two systems must be computed using the synchronized resampling Algorithm II (see Section 5-C). To reduce the computational fluctuation derived from the stochastic nature of such resampling, Algorithm II was run 20 times and the average out of 20 runs was taken to be the resultant correlation coefficient listed in Table 5.

In this table, all correlation coefficients of CFs of two systems are positive as expected (see Section 5-C) between 0.73 and 0.85. It indicates that all systems tend to assign higher or lower scores to the corresponding trials. These results provide evidence that the synchronized resampling algorithm is quite reasonable for computing the correlation coefficient.

sys	B	C	D	E
A	0.6398	0.0000	0.0000	0.0000
B		0.0000	0.0000	0.0000
C			0.2598	0.0000
D				0.0000

Table 6 The two-classifier two-tailed *p*-values among Systems A through E, in which the correlation coefficients were taken into consideration.

By employing Eq. (14) and using R [27], the two-classifier two-tailed *p*-values among Systems A through E after taking account of the positive correlation coefficients are presented in Table 6. In this table, only two *p*-values are greater than 5%, which are 63.98% for System A versus System B, and 25.98% for System C versus System D. It indicates that the null hypothesis $H_o : D_1 = D_2$ is accepted at the achieved significance level 64% and 26%, respectively [11]. That is, the differences of the performance levels between Systems A and B and Systems C and D are not statistically significant, even though the observed CFs of Systems A and C are smaller than those of Systems B and D, respectively. This conclusion is consistent with the observation in Figure 3, where the two 95% CIs of CFs for Systems A and B overlap considerably, and the same is true for Systems C and D.

All other *p*-values in Table 6 are much less than 5%, which are too small to be shown. It suggests that the null hypotheses $H_o : D_1 = D_2$ in those cases should be very strongly rejected in favor of the alternative hypotheses $H_a : D_1 \neq D_2$ for those pairs of systems. In other words, the differences of the performance levels between the two corresponding systems are statistically significant. This is supported by the observation in Figure 3, where either the corresponding 95% CIs do not overlap, or System B's 95% CI overlaps System C's 95% CI a little and the same holds true for Systems D and E. The latter two cases show that even though two 95% CIs overlap to some extent, the performance levels of two classifiers could be significantly different.

Here is a very important observation. Regarding Systems B and C and Systems D and E, respectively, if the positive correlation coefficient is not considered, the corresponding two *p*-values are 0.0286 and 0.0154. In other words, the two achieved significance levels are 3% and 2%, which are a little less than 5% (see Section 5), meaning that the alternative hypothesis is accepted with borderline evidence, i.e., the performance difference is not statistically significant [11]. Certainly, this contradicts the observation in Figure 3 as stated above. Thus, failing to take account of the positive correlation coefficient can wrongly reduce the likelihood of detecting the statistical significance of a difference between the performance levels of the two systems. It also demonstrates that the synchronized resampling algorithm of computing the correlation coefficient works well for the applications in which tests are designed in the way as described in Section 5-C.

8 Conclusions and discussion

For the datasets originated from ROC analysis with data dependency, the addition of two scores does not make any sense, and no statistic can be used to describe any characteristics related to sets of scores, which are created by different subjects. To analyze the related bootstrap approach, the traditional statistical methods of data analysis via computing variances cannot be applied. Hence, the hierarchical method as discussed in Ref. [16] is not relevant.

Indeed, to reduce the bootstrap variance and ensure the computation accuracy, our prior rigorous statistical research was carried out, such as the bootstrap variability studies that took months of CPU time to determine the appropriate number of bootstrap replications, the validation study by comparing the SEs of AUC estimated using the bootstrap algorithm on large i.i.d. datasets against those computed using the well-established analytical Mann-Whitney statistic method, using the multinomial probabilities to determine which bootstrap approach in the two-layer data structure should be used, and so on [4, 10-11, 17-22].

In this article, a new ROC analysis with data dependency for decision making of classifiers is dealt with, in which three types of nonparametric distributions of scores were created, two decision thresholds were involved, and a measure was defined to be an average of the two weighted sums of the probabilities of type I and type II errors at the two thresholds correspondingly.

To generate more samples because of limited resources, data dependency may be inevitable due to the need for multiple use of the same subjects. Thus, a new two-layer data structure was constructed after dataset optimization: for each type of scores, scores are grouped into sets based on the subjects, and the numbers of scores are kept the same for all sets. Such a data structure preserves the data dependency while the bootstrap resampling takes place. Moreover, scores of each type can be selected with equal probability, and the numbers of scores of each type resampled at each iteration while using the bootstrap algorithm can be the same.

The performance levels, i.e., CFs were computed before and after dataset optimization that cut the data sizes almost in half. When the data size reaches to over tens of thousands in ROC analysis, such dataset optimization has little impact on the performance levels of classifiers as shown in Section 6 [24]. Indeed, if data dependency is involved, datasets should be created based on the descriptions stated in Section 2 at the very beginning rather than using the dataset optimization later, which can save lots of resources.

The SEs of measures in the new ROC analysis with data dependency were computed using the nonparametric three-sample two-layer bootstrap algorithm applied on a two-layer data structure constructed after dataset optimization with about half of the raw data, based on the prior studies on bootstrap methods and bootstrap variability [3-5, 17, 20-22]. The bootstrap resampling takes place randomly WR on the three score distributions – not only on sets but subsequently on scores within the sets.

By contrast, the nonparametric three-sample bootstrap method conducted on all the raw data with the i.i.d. assumption underestimates the SEs of CFs, even after taking account of the factor of the data-size reduction. And the analytical approach cannot deal with the issues such as the data dependency and the covariance caused by the probabilities of type I error and type II error at a decision threshold, and it usually underestimates the SE of the measure.

Based on such SEs, to investigate the statistical significance of differences for evaluation and comparison of classifiers, the one-classifier and two-classifier significance testing in statistics can play an important role in providing quantitative information in terms of the significance level, i.e., *p*-value. To this end, when the 95% CI contains the performance criterion for evaluation and the two 95% CIs overlap for comparison, only using the 95% CI of the performance level of a classifier is far less than enough.

Further, because all scores of different systems are created on a common set of speakers and speech segments and all systems tend to assign higher or lower scores to the same trials, the resulting CFs of systems derived from these scores are positively correlated. Indeed, this is usually the way how the test is designed. As a result, the synchronized resampling algorithm of computing the correlation coefficient stated in this article is valid for this kind of test.

To conduct the two-classifier significance testing for performance comparison of two classifiers, the positive correlation coefficient must be taken into consideration as shown in Eq. (14). Otherwise, the likelihood of detecting the statistical significance of difference between the performance levels of two classifiers can be wrongly reduced. As demonstrated by all examples in this article, the synchronized resampling algorithm works well.

As far as the comparison is concerned, in this article, the pairwise comparison was conducted as shown in Section 7-B. If the multiple comparisons issue is of interest, then some procedures, such as Tukey's method, Scheffe's method, Bonferroni's method and so on, might need to be explored [28-29, and references therein].

Acknowledgment

The authors would like to thank Dr. Alvin F. Martin for providing the data and discussion regarding the test implementation issues, and William F. Guthrie, the Statistical Engineering Division Chief at the National Institute of Standards and Technology, for reviewing our manuscript and his valuable comments.

References

1. J.A. Hanley and B.J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.
2. T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no 8, pp. 861–74, 2006.
3. J.C. Wu, A.F. Martin, and R.N. Kacker, "Measures, uncertainties, and significance test in operational ROC analysis," *J. Res. Natl. Inst. Stand. Technol*., vol. 116, no. 1, pp. 517–537, 2011.
4. J.C. Wu, A.F. Martin, C.S. Greenberg, and R.N. Kacker, "The impact of data dependence on speaker recognition evaluation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 1, pp. 5-18, 2017.
5. J.C. Wu, M. Halter, R.N. Kacker, J.T. Elliott, and A.L. Plant, "A novel measure and significance testing in data analysis of cell image segmentation," *BMC Bioinformatics*, 18:168, pp. 1-13, 2017.
6. J.C. Wu and C.L. Wilson, "Nonparametric analysis of fingerprint data on large data sets," *Pattern Recognition*, vol. 40, no. 9, pp. 2574-2584, 2007.
7. "The NIST Speaker Recognition Evaluation," 2020. [Online]. Available: https://www.nist.gov/system/files/documents/itl/iad/mig/NIST_SRE12_evalplan-v17-r1.pdf
8. N. Brummer and J. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech and Language*, vol. 20, pp. 230-275, 2006.
9. G.R. Doddington, M.A. Przybocki, A.F. Martin, and D.A. Reynolds, "The NIST speaker recognition evaluation – Overview, methodology, systems, results, perspective," *Speech Communication*, vol. 31, pp. 225–254, 2000.
10. B. Efron, "Bootstrap methods: Another look at the Jackknife," *Ann. Statistics*, vol. 7, no. 1, pp. 1-26, 1979.

11. B. Efron and R.J. Tibshirani, *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.
12. J.A. Rice, *Mathematical Statistics and Data Analysis, third ed.* Thomson Brooks/Cole, Belmont, 2007.
13. E.L. Lehmann and J.P. Romano, *Testing Statistical Hypotheses, third ed.* Springer, New York, 2005.
14. J. A. Hanley and B. J. McNeil, “A method of comparing the areas under receiver operating characteristic curves derived from the same cases,” *Radiology*, vol. 148, pp. 839-843, 1983.
15. R.Y. Liu and K. Singh, “Moving blocks jackknife and bootstrap capture weak dependence,” in *Exploring the Limits of Bootstrap*, ed. by LePage and Billard, John Wiley, New York, 1992.
16. A.C. Davison and D.V. Hinkley, *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge, 1997.
17. J.C. Wu, A.F. Martin, and R.N. Kacker, “Validation of nonparametric two-sample bootstrap in ROC analysis on large datasets,” *Communications in Statistics – Simulation and Computation*, vol. 45, no. 5, pp. 1689-1703, 2016.
18. P. Hall, “On the number of bootstrap simulations required to construct a confidence interval,” *Ann. Statistics*, vol. 14, no. 4, pp. 1453-1462, 1986.
19. B. Efron, “Better bootstrap confidence intervals,” *J. Amer. Statist. Assoc.*, vol. 82, no. 397, pp. 171-185, 1987.
20. J.C. Wu, A.F. Martin, and R.N. Kacker, “Monte Carlo studies of bootstrap variability in ROC analysis with data dependency,” *Communications in Statistics – Simulation and Computation*, vol. 48, no. 2, pp. 317-333, 2019.
21. J.C. Wu, A.F. Martin, and R.N. Kacker, “Bootstrap variability studies in ROC analysis on large datasets,” *Communications in Statistics – Simulation and Computation*, vol. 43, no. 1, pp. 225–236, 2014.
22. J.C. Wu, “Studies of operational measurement of ROC curve on large fingerprint data sets using two-sample bootstrap,” National Institute of Standards and Technology, *NISTIR 7449*, Sep. 2007.
23. B. Ostle and L.C. Malone, *Statistics in Research: Basic Concepts and Techniques for Research Workers, fourth ed.* Iowa State University Press, Ames, 1988.
24. J.C. Wu and C.L. Wilson, “An empirical study of sample size in ROC-curve analysis of fingerprint data,” in *Biometric Technology for Human Identification III,* Proc. SPIE 6202, 620207, 2006.
25. R.J. Hyndman and Y. Fan, “Sample quantiles in statistical packages,” *The American Statistician*, vol. 50, no. 4, pp. 361-365, 1996.
26. G.E.P. Box, J.S. Hunter, and W.G. Hunter, *Statistics for experimenters: design, innovation, and discovery*, *second ed.* John Wiley & Sons, Inc., New York, 2005.
27. *R: A Language and Environment for Statistical Computing*, The R Development Core Team, Version 4.0.1, 2020. [Online]. Available: https://www.r-project.org/
28. R.G. Miller, *Simultaneous Statistical Inference, second ed.* Springer Verlag, New York, 1981.
29. *NIST/SEMATECH e-Handbook of Statistical Methods*, The National Institute of Standards and Technology, 2020. [Online]. Available: https://www.itl.nist.gov/div898/handbook/