Applied Intelligence (2022) 52:4961–4972
https://doi.org/10.1007/s10489-021-02635-5

# Confidence interval for micro-averaged $F_1$ and macro-averaged $F_1$ scores

Kanae Takahashi[1,2] · Kouji Yamamoto[3] · Aya Kuchiba[4,5] · Tatsuki Koyama[6]

Accepted: 19 June 2021 / Published online: 31 July 2021


**Abstract**
A binary classification problem is common in medical field, and we often use sensitivity, specificity, accuracy, negative and positive predictive values as measures of performance of a binary predictor. In computer science, a classifier is usually evaluated with precision (positive predictive value) and recall (sensitivity). As a single summary measure of a classifier's performance, $F_1$ score, defined as the harmonic mean of precision and recall, is widely used in the context of information retrieval and information extraction evaluation since it possesses favorable characteristics, especially when the prevalence is low. Some statistical methods for inference have been developed for the $F_1$ score in binary classification problems; however, they have not been extended to the problem of multi-class classification. There are three types of $F_1$ scores, and statistical properties of these $F_1$ scores have hardly ever been discussed. We propose methods based on the large sample multivariate central limit theorem for estimating $F_1$ scores with confidence intervals.

**Keywords** Precision · Recall · Machine learning · $F_1$ measures · Multi-class classification · Delta-method

## 1 Introduction

In medical field, a binary classification problem is common, and we often use sensitivity, specificity, accuracy, negative and positive predictive values as measures of performance of a binary predictor. In computer science, a classifier is usually evaluated with precision and recall, which are equal to positive predictive value and sensitivity, respectively. For measuring the performance of text classification in the field of information retrieval and of a classifier in machine learning, the $F$ score ($F$ measure) has been widely used. In particular, the $F_1$ score has been popular, which is defined as the harmonic mean of precision and recall [1, 2]. The $F_1$ score is rarely used in diagnostic studies in medicine despite its favorable characteristics. As a single performance measure, the $F_1$ score may be preferred to specificity and accuracy, which may be artificially high even for a poor classifier with a high false negative probability when disease prevalence is low. The $F_1$ score is especially useful when identification of true negatives is relatively unimportant because the true negative rate is not included in the computation of either precision or recall.

To evaluate a multi-class classification, a single summary measure is often sought. And as extensions of the $F_1$ score for the binary classification, there exist two types of such measures: a micro-averaged $F_1$ score and a macro-averaged $F_1$ score [2]. The micro-averaged $F_1$ score pools per-sample classifications across classes, and then calculates the overall $F_1$ score. Contrarily, the macro-averaged $F_1$ score computes a simple average of the $F_1$ scores over classes. Sokolova and Lapalme [3] gave an alternative definition of the macro-averaged $F_1$ score as the harmonic mean of the simple averages of the precision and recall over classes. Both micro-averaged and macro-averaged $F_1$ scores have a

✉ Tatsuki Koyama
tatsuki.koyama@vumc.org

1 Department of Medical Statistics, Osaka City University Graduate School of Medicine, Osaka, Japan

2 Department of Biostatistics, Hyogo College of Medicine, Hyogo, Japan

3 Department of Biostatistics, School of Medicine, Yokohama City University, Yokohama, Japan

4 Graduate School of Health Innovation, Kanagawa University of Human Services, Kanagawa, Japan

5 Division of Biostatistical Research, Center for Public Health Sciences / Biostatistics Division, Center for Research Administration and Support, National Cancer Center, Tokyo, Japan

6 Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, USA

simple interpretation as an average of precision and recall, with different ways of computing averages. Moreover, as will be shown in Section 2, the micro-averaged $F_1$ score has an additional interpretation as the total probability of true positive classifications.

For binary classification, some statistical methods for inference have been proposed for the $F_1$ scores (e.g., [4]); however, the methodology has not been extended to the multi-class $F_1$ scores. To our knowledge, methods for computing variance estimates of the micro-averaged $F_1$ score and macro-averaged $F_1$ score have not been reported. Thus, computing confidence intervals for the multi-class $F_1$ scores is not possible, and the inference about them is usually solely based on point estimates, and thus highly limited in practical utility. For example, consider the results of an analysis reported by Dong et al. [5]. In this analysis, the authors calculated the point estimates of macro-averaged $F_1$ scores for four classifiers, and they concluded a classifier outperformed the others by comparing the point estimates without taking into account their uncertainty. Others have also used multi-class $F_1$ scores but only reported point estimates without confidence intervals [6–16].

To address this knowledge gap, we provide herein the methods for computing variances of these multi-class $F_1$ scores so that estimating the micro-averaged $F_1$ score and macro-averaged $F_1$ score with confidence intervals becomes possible in multi-class classification.

The rest of the manuscript is organized as follows: The definitions of the micro-averaged $F_1$ score and macro-averaged $F_1$ score are reviewed in Section 2. In Section 3, variance estimates and confidence intervals for the multi-class $F_1$ scores are derived. A simulation study to investigate the coverage probabilities of the proposed confidence intervals is presented in Section 4. Then, our method is applied to a real study as an example in Section 5 followed by a brief discussion in Section 6.

## 2 Averaged $F_1$ scores

This section introduces notations and definitions of multi-class $F_1$ scores, namely, macro-averaged and micro-averaged $F_1$ scores. Consider an $r \times r$ contingency table for a nominal categorical variable with $r$ classes ($r \geq 2$). The columns indicate the true conditions, and rows indicate the predicted conditions. It is called the binary classification when $r = 2$, and the multi-class classification when $r > 2$. Such a table is also called a confusion matrix. We consider multi-class classification, i.e., $r > 2$, and denote cell probabilities and marginal probabilities by $p_{ij}$, $p_{i\cdot}$, and $p_{\cdot j}$, respectively ($i, j = 1, \cdots, r$). For each class $i$ ($i = 1, \cdots, r$), the true positive rate ($TP_i$), the false positive rate ($FP_i$), and the false negative rate ($FN_i$) are defined as follows:

$$
\begin{aligned}
TP_i &= p_{ii}, \\
FP_i &= \sum_{\substack{j=1 \\ j \neq i}}^{r} p_{ij}, \\
FN_i &= \sum_{\substack{j=1 \\ j \neq i}}^{r} p_{ji}.
\end{aligned}
$$

$TP_i$ is the $i$-th diagonal element, $FP_i$ is the sum of off-diagonal elements of the $i$-th row, and $FN_i$ is the sum of off-diagonal elements of the $i$-th column. Note that $TP_i + FP_i = p_{i\cdot}$, and $TP_i + FN_i = p_{\cdot i}$.

In the current and following sections, we will use the simple 3-by-3 confusion matrix in Table 1 as an example to demonstrate various computations. Columns represent the true state, and rows represent the predicted classification. The total sample size is 100.

The within-class probabilities are:

$$
\begin{aligned}
TP_1 &= 0.02, & TP_2 &= 0.70, & TP_3 &= 0.15. \\
FP_1 &= 0.04, & FP_2 &= 0.07, & FP_3 &= 0.02. \\
FN_1 &= 0.05, & FN_2 &= 0.04, & FN_3 &= 0.04.
\end{aligned}
$$

**Micro-averaged $F_1$ score** The micro-averaged precision ($miP$) and micro-averaged recall ($miR$) are defined as

$$
\begin{aligned}
miP &= \frac{\sum_{i=1}^{r} TP_i}{\sum_{i=1}^{r} (TP_i + FP_i)} = \frac{\sum p_{ii}}{\sum p_{i\cdot}} = \sum_{i=1}^{r} p_{ii}, \\
miR &= \frac{\sum_{i=1}^{r} TP_i}{\sum_{i=1}^{r} (TP_i + FN_i)} = \frac{\sum p_{ii}}{\sum p_{\cdot i}} = \sum_{i=1}^{r} p_{ii}.
\end{aligned}
$$

Note that for both $miP$ and $miR$, the denominator is the sum of all the elements (diagonal and off-diagonal) of the confusion matrix, and it is 1. Finally, the micro-averaged $F_1$

**Table 1** Numeric example

| | | True Classification | | | |
|---|---|---|---|---|---|
| | | Class 1 | Class 2 | Class 3 | |
| a: Frequencies | | | | | |
| | Class 1 | 2 | 2 | 2 | |
| Prediction | Class 2 | 5 | 70 | 2 | |
| | Class 3 | 0 | 2 | 15 | |
| b: Proportions | | | | | |
| | Class 1 | 0.02 | 0.02 | 0.02 | 0.06 |
| Prediction | Class 2 | 0.05 | 0.70 | 0.02 | 0.77 |
| | Class 3 | 0.00 | 0.02 | 0.15 | 0.17 |
| | | 0.07 | 0.74 | 0.19 | |

score is defined as the harmonic mean of these quantities:

$$miF_1 = 2\frac{miP \times miR}{miP + miR} = \sum_{i=1}^{r} p_{ii}. \quad (1)$$

This definition is commonly used (e.g., [6, 8–12, 14, 15]).

By definition, we have $miP$, $miR$, and $miF_1$ all equal to the sum of the diagonal elements, which, in our example, is 0.87.

**Macro-averaged $F_1$ score** To define the macro-averaged $F_1$ score ($maF_1$), first consider the following precision ($P_i$) and recall ($R_i$) within each class, $i = 1, \cdots, r$:

$$P_i = \frac{TP_i}{(TP_i + FP_i)} = p_{ii}/p_{i\cdot},$$
$$R_i = \frac{TP_i}{(TP_i + FN_i)} = p_{ii}/p_{\cdot i}.$$

For our example, simple calculation shows:

$$P_1 = 0.33, \quad P_2 = 0.91, \quad P_3 = 0.88,$$
$$R_1 = 0.29, \quad R_2 = 0.95, \quad R_3 = 0.79.$$

And $F_1$ score within each class ($F_{1i}$) is defined as the harmonic mean of $P_i$ and $R_i$, that is,

$$F_{1i} = 2\frac{P_i \times R_i}{P_i + R_i} = 2\frac{p_{ii}}{p_{i\cdot} + p_{\cdot i}}.$$

The macro-averaged $F_1$ score is defined as the simple arithmetic mean of $F_{1i}$:

$$maF_1 = \frac{1}{r}\sum_{i=1}^{r} F_{1i} = \frac{2}{r}\sum_{i=1}^{r}\frac{p_{ii}}{p_{i\cdot} + p_{\cdot i}}. \quad (2)$$

This score, like $miF_1$, is frequently reported (e.g., [5–10, 13]).

$F_{1i}$ and $maF_1$ in our example are:

$$F_{11} = 0.308, \quad F_{12} = 0.927, \quad F_{13} = 0.833.$$
$$maF_1 = (0.308 + 0.927 + 0.833)/3 = 0.689.$$

**Alternative definition of Macro-averaged $F_1$ score** Sokolova and Lapalme [3] gave an alternative definition of the macro-averaged $F_1$ score ($maF_1^*$). First, macro-averaged precision ($maP$) and macro-averaged recall ($maR$) are defined as simple arithmetic means of the within-class precision and within-class recall, respectively.

$$maP = \frac{1}{r}\sum_{i=1}^{r}\frac{TP_i}{TP_i + FP_i} = \frac{1}{r}\sum_{i=1}^{r}\frac{p_{ii}}{p_{i\cdot}},$$
$$maR = \frac{1}{r}\sum_{i=1}^{r}\frac{TP_i}{TP_i + FN_i} = \frac{1}{r}\sum_{i=1}^{r}\frac{p_{ii}}{p_{\cdot i}}.$$

And $maF_1^*$ is defined as the harmonic mean of these quantities.

$$maF_1^* = 2\frac{maP \times maR}{maP + maR}. \quad (3)$$

This version of macro-averaged $F_1$ score is less frequently used (e.g., [11, 12, 16]). For our example,

$$maP = (0.02/0.06 + 0.70/0.77 + 0.15/0.17)/3 = 0.708.$$
$$maR = (0.02/0.07 + 0.70/0.74 + 0.15/0.19)/3 = 0.674.$$
$$maF_1^* = 0.691.$$

In this example, the micro-averaged $F_1$ score is higher than the macro-averaged $F_1$ scores because both within-class precision and recall are much lower for the first class compared to the other two. Micro-averaging puts only a small weight on the first column because the sample size there is relatively small. This numeric example shows a shortcoming of summarizing a performance of a multi-class classification with a single number when within-class precision and recall vary substantially. However, aggregate measures such as the micro-averaged and macro-averaged $F_1$ scores are useful in quantifying the performance of a classifier as a whole.

## 3 Variance estimate and confidence interval

In this section, we derive the confidence interval for $miF_1$, $maF_1$, and $maF_1^*$. We assume that the observed frequencies, $n_{ij}$, for $1 \le i \le r$, $1 \le j \le r$, have a multinomial distribution with sample size $n$ and probabilities
$\boldsymbol{p} = (p_{11}, \cdots, p_{1r}, \ p_{21}, \cdots, p_{2r}, \ \cdots, \ p_{r1}, \cdots, p_{rr})^T$, where "$T$" represents the transpose, that is

$$(n_{11}, n_{12}, \cdots, n_{rr}) \sim Multinomial(n; \boldsymbol{p}).$$

The expectation, variance, and covariance for $i, j = 1, \cdots, r$, are:

$$E\left(n_{ij}\right) = np_{ij},$$
$$Var\left(n_{ij}\right) = np_{ij}(1 - p_{ij}),$$
$$Cov\left(n_{ij}, n_{kl}\right) = -np_{ij}p_{kl}, \text{ for } i \ne k \text{ or } j \ne l,$$

respectively, where $n = \sum_{i,j} n_{ij}$ is the overall sample size. The maximum likelihood estimate (MLE) of $p_{ij}$ is $\hat{p}_{ij} = n_{ij}/n$. Using the multivariate central limit theorem, we have

$$\sqrt{n}\left(\hat{\boldsymbol{p}} - \boldsymbol{p}\right) \dot{\sim} Normal\left(\mathbf{0}_{r^2}, diag(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right),$$

where $\mathbf{0}_{r^2}$ is $r^2 \times 1$ vector whose elements are all 0, $diag(\boldsymbol{p})$ is an $r^2 \times r^2$ diagonal matrix whose diagonal elements are $\boldsymbol{p}$, and "$\dot{\sim}$" represents "approximately distributed as."

By invariance property of MLE's, the maximum likelihood estimates of $miF_1$, $maF_1$, $maF_1^*$, and other quantities in the previous section can be obtained by substituting $p_{ij}$ by $\hat{p}_{ij}$. In the following subsections, we use the multivariate delta-method to derive large-sample distributions of $\widehat{miF_1}$, $\widehat{maF_1}$, and $\widehat{maF_1^*}$.

### 3.1 Confidence interval for $miF_1$

As shown in (1), $miF_1 = \sum p_{ii}$, and the maximum likelihood estimate (MLE) of $miF_1$ is

$$\widehat{miF_1} = \sum_{i=1}^{r} \hat{p}_{ii}.$$

Using the multivariate delta-method (Appendix A), we have

$$\widehat{miF_1} \dot{\sim} Normal\left(miF_1, Var(\widehat{miF_1})\right),$$

where variance of $\widehat{miF_1}$ is

$$Var\left(\widehat{miF_1}\right) = \left(\sum_{i=1}^{r} p_{ii}\right)\left(1 - \sum_{i=1}^{r} p_{ii}\right) \Big/ n. \tag{4}$$

And a $(1-\alpha)\times 100\%$ confidence interval of $miF_1$ is

$$\widehat{miF_1} \pm Z_{1-\alpha/2} \times \sqrt{\widehat{Var}\left(\widehat{miF_1}\right)},$$

where $\widehat{Var}\left(\widehat{miF_1}\right)$ is $Var\left(\widehat{miF_1}\right)$ with $\{p_{ii}\}$ replaced by $\{\hat{p}_{ii}\}$, and $Z_p$ denote the $100p$-th percentile of the standard normal distribution. Computation of $\widehat{Var}\left(\widehat{miF_1}\right)$ for our numeric example is straightforward using (4):

$$\begin{aligned}\widehat{Var}\left(\widehat{miF_1}\right) &= (0.02 + 0.70 + 0.15) \\ &\quad \times \{1 - (0.02 + 0.70 + 0.15)\}/100 \\ &= 0.0336^2.\end{aligned}$$

And a 95% confidence interval for $miF_1$ is

$$0.87 \pm 1.960 \times 0.0336 = (0.804,\ 0.936).$$

### 3.2 Confidence interval for $maF_1$

The MLE of $maF_1$ can be obtained by substituting $p_{ii}$, $p_{\cdot i}$ and $p_{\cdot i}$ by their MLE's in (2).

$$\widehat{maF_1} = \frac{2}{r}\sum_{i=1}^{r} \frac{\hat{p}_{ii}}{\hat{p}_{i\cdot} + \hat{p}_{\cdot i}}.$$

Again by the multivariate delta-method (Appendix B), we have the variance of $\widehat{maF_1}$ as

$$\begin{aligned}Var\left(\widehat{maF_1}\right) &= \frac{2}{r^2}\left\{\sum_{i=1}^{r} \frac{F_{1i}(p_{i\cdot} + p_{\cdot i} - 2p_{ii})}{(p_{i\cdot} + p_{\cdot i})^2}\left(\frac{p_{i\cdot} + p_{\cdot i} - 2p_{ii}}{p_{i\cdot} + p_{\cdot i}} + \frac{F_{1i}}{2}\right)\right. \\ &\quad \left. + \sum_{i=1}^{r}\sum_{j\neq i} \frac{p_{ij}F_{1i}F_{1j}}{(p_{i\cdot} + p_{\cdot i})(p_{j\cdot} + p_{\cdot j})}\right\} \Big/ n.\end{aligned}$$

A $(1-\alpha)\times 100\%$ confidence interval of $maF_1$ is

$$\widehat{maF_1} \pm Z_{1-\alpha/2} \times \sqrt{\widehat{Var}\left(\widehat{maF_1}\right)},$$

where $\widehat{Var}\left(\widehat{maF_1}\right)$ is $Var\left(\widehat{maF_1}\right)$ with $\{p_{ij}\}$ replaced by $\{\hat{p}_{ij}\}$. This computation is complex even for a small 3 by 3 table; an R code (Appendix D) was used to compute the variance estimate and a 95% confidence interval of $maF_1$.

$$\begin{aligned}\widehat{Var}\left(\widehat{maF_1}\right) &= 0.0650^2, \\ 0.69 \pm 1.960 \times 0.0650 &= (0.562,\ 0.817).\end{aligned}$$

### 3.3 Confidence interval for $maF_1^*$

To obtain the MLE's of $maF_1^*$, we first substitute $p_{ii}$, $p_{i\cdot}$ and $p_{\cdot i}$ by their MLE's to get MLE's of $maP$ and $maR$ and use these in (3):

$$\widehat{maF_1^*} = 2\frac{\widehat{maP} \times \widehat{maR}}{\widehat{maP} + \widehat{maR}}.$$

As shown in Appendix C,

$$Var\left(\widehat{maF_1^*}\right) = 4n\frac{maR^4 Var\left(\widehat{maP}\right) + 2maP^2 maR^2 Cov\left(\widehat{maP}, \widehat{maR}\right) + maP^4 Var\left(\widehat{maR}\right)}{(maP + maR)^4} \Big/ n,$$

where

$$\begin{aligned}Var\left(\widehat{maP}\right) &= \frac{1}{r^2}\left(\sum_{i=1}^{r} \frac{p_{ii}\left(\sum_{j\neq i} p_{ij}\right)}{p_{i\cdot}^3}\right) \Big/ n, \\ Var\left(\widehat{maR}\right) &= \frac{1}{r^2}\left(\sum_{i=1}^{r} \frac{p_{ii}\left(\sum_{j\neq i} p_{ji}\right)}{p_{\cdot i}^3}\right) \Big/ n, \\ Cov\left(\widehat{maP}, \widehat{maR}\right) &= \frac{1}{r^2}\left(\sum_{i=1}^{r} \frac{\left(\sum_{j\neq i} p_{ij}\right)p_{ii}\left(\sum_{j\neq i} p_{ji}\right)}{p_{i\cdot}^2 p_{\cdot i}^2}\right. \\ &\quad \left. + \sum_{i=1}^{r}\sum_{j\neq i} \frac{p_{ii}p_{ij}p_{jj}}{p_{i\cdot}^2 p_{j\cdot}^2}\right) \Big/ n.\end{aligned}$$

A $(1-\alpha)\times 100\%$ confidence interval of $maF_1^*$ is

$$\widehat{maF_1^*} \pm Z_{1-\alpha/2} \times \sqrt{\widehat{Var}\left(\widehat{maF_1^*}\right)}.$$

Again to get $\widehat{Var}\left(\widehat{maF_1^*}\right)$, all components of $Var\left(\widehat{maF_1^*}\right)$ are replaced by their respective MLE's. Using the accompanying R code (Appendix D), we computed the variance estimate and a 95% confidence interval

of $maF_1^*$:

$$\widehat{Var}\left(\widehat{maF_1^*}\right) = 0.0649^2$$
$$0.69 \pm 1.960 \times 0.0649 = (0.563,\ 0.818)\,.$$

## 4 Simulation

We performed a simulation study to assess the coverage probability of the confidence intervals proposed in Section 3. We set $r = 3$ (class 1, 2, 3), and generated data according to the multinomial distributions with $\boldsymbol{p}$ summarized in Table 2. The total sample size, $n$, was set to 25, 50, 100, 500, 1,000, and 5,000. For each combination of the true distribution and sample size, we generated 1,000,000 data, each time computing 95% confidence intervals for $miF_1$, $maF_1$, and $maF_1^*$.

In scenario 1, the true conditions of class 1, 2, and 3 have the same probability (1/3), and the recall and precision are equal (80%). Thus $miP = maP = 0.80$, $miR = maR = 0.80$, and $miF_1 = maF_1 = maF_1^* = 0.80$.

In scenario 2, the true condition of class 1 has higher probability than the others (80% vs 10%), and the recall and precision of class 1 are also higher than the others (80% vs 40%, and 91% vs 27%, respectively). $miF_1$ gives equal weight to each per-sample classification decision, whereas $maF_1$ gives equal weight to each class. Thus, large classes dominate small classes in computing $miF_1$ [2], and $miF_1$ is larger than $maF_1$ ($miF_1 = 0.72$, $maF_1 = 0.50$, $maF_1^* = 0.51$) in scenario 2 because class 1 has higher probability and has higher precision and recall.

**Table 2** Simulation study: True cell probabilities

| | | True condition | | |
|---|---|---|---|---|
| | | 1 | 2 | 3 |
| Scenario 1 | | | | |
| | 1 | 8/30 | 1/30 | 1/30 |
| Predicted condition | 2 | 1/30 | 8/30 | 1/30 |
| | 3 | 1/30 | 1/30 | 8/30 |
| Scenario 2 | | | | |
| | 1 | 64/100 | 3/100 | 3/100 |
| Predicted condition | 2 | 8/100 | 4/100 | 3/100 |
| | 3 | 8/100 | 3/100 | 4/100 |
| Scenario 3 | | | | |
| | 1 | 32/100 | 1/100 | 1/100 |
| Predicted condition | 2 | 24/100 | 8/100 | 1/100 |
| | 3 | 24/100 | 1/100 | 8/100 |

In scenario 3, the true condition of class 1 has higher probability than the others (80% vs 10%). The precision of class 1 is higher than the others (94% vs 24%), and the recall of class 1 is lower than the others, (40% vs 80%). Compared to the other two scenarios, the diagonal entries are relatively small, which makes $miF_1$ small ($miF_1 = 0.48$, $maF_1 = 0.44$, and $maF_1^* = 0.55$).

Table 3 shows the coverage probability of the proposed 95% confidence intervals for each scenario. The coverage probabilities for both $miF_1$ and $maF_1$ are close to the nominal 95% when the sample size is large. When $n$ is small (25, 50), the coverage probability tends to be smaller than 95%, especially for $maF_1$ and $maF_1^*$. Moreover, computing a confidence interval for $maF_1^*$ for small $n$ is often impossible because $\widehat{maF_1^*}$ is undefined when either $p_{i\cdot} = 0$ or $p_{\cdot j} = 0$ for any $i$ or $j$. In typical applications where these $F$ scores are computed, $n$ is large, and the small $n$ problem is unlikely to occur.

## 5 Example

As an example, we applied our method to the temporal sleep stage classification data provided by Dong et al. [5]. They proposed a new approach based on a Mixed Neural Network (MNN) to classify sleep into five stages with one awake stage (W), three sleep stages (N1, N2, N3), and one rapid eye movement stage (REM). In addition to the MNN, they evaluated the following three classifiers: Support Vector Machine (SVM), Random Forest (RF), and Multilayer Perceptron (MLP). The data came from 62 healthy subjects, and classification by a single sleep expert was used as the gold standard. The staging is based on a 30-second window of the physiological signals called an EEG (electroencephalography) epoch. Thus, each subject contributes a large number of data to be classified. The total number of epochs depends on the classifiers, and it is about 59,000. Performance of each classifier was evaluated using $maF_1$ along with precision, recall, and overall accuracy. They concluded that the MNN outperformed the competitors by comparing the point estimates of $maF_1$ and overall accuracy. We provide here 95% confidence intervals for $miF_1$, $maF_1$, and $maF_1^*$ for each of the four methods, as summarized in Table 4. The confidence intervals of $miF_1$, $maF_1$, and $maF_1^*$ for the MNN do not overlap with the point estimates of other methods, providing further evidence that MNN is superior to the other method. For completeness we present 95% confidence intervals for other methods in Table 4 as well. As $n$ is large for this example, the confidence intervals are narrow, and the ones for MNN do not overlap with confidence intervals for other three methods.

**Table 3** Simulation study: Coverage probability

| | Scenario 1 | | | Scenario 2 | | | Scenario 3 | | |
|---|---|---|---|---|---|---|---|---|---|
| $n$ | $miF_1$ | $maF_1$ | $maF_1^*$ | $miF_1$ | $maF_1$ | $maF_1^*$ | $miF_1$ | $maF_1$ | $maF_1^*$ |
| 25 | 0.885 | 0.901 | 0.890 | 0.921 | 0.790 | 0.774 | 0.930 | 0.870 | 0.821 |
| 50 | 0.937 | 0.935 | 0.923 | 0.941 | 0.864 | 0.853 | 0.935 | 0.918 | 0.905 |
| 100 | 0.933 | 0.938 | 0.936 | 0.937 | 0.914 | 0.914 | 0.943 | 0.936 | 0.933 |
| 500 | 0.949 | 0.949 | 0.948 | 0.947 | 0.944 | 0.945 | 0.946 | 0.947 | 0.947 |
| 1000 | 0.946 | 0.948 | 0.948 | 0.947 | 0.947 | 0.947 | 0.947 | 0.949 | 0.947 |
| 5000 | 0.950 | 0.950 | 0.950 | 0.951 | 0.949 | 0.949 | 0.951 | 0.950 | 0.950 |

## 6 Discussion

We derived large sample variance estimates of $miF_1$, $maF_1$, and $maF_1^*$ in terms of the observed cell probabilities and sample size. This enabled us to derive large sample confidence intervals.

Coverage probabilities of the proposed confidence intervals were assessed through the simulation study. According to the result of the simulation, when $n$ is larger than 100, the coverage probability was close to the nominal level; however, for $n < 100$, the coverage probabilities tended to be smaller than the target. Moreover, with an extremely small sample size, $maF_1^*$ could not be estimated as computation of $maF_1^*$ requires all margins to be non-zero. Zhang et al. [17] have considered interval estimation of $miF_1$ and $maF_1$ and proposed the highest density interval through Bayesian framework. On the other hand, we have proposed confidence interval for $miF_1$, $maF_1$, and $maF_{1*}$ through frequentist framework using a large-sample approximation.

There is an inherit drawback of multi-class $F_1$ scores that these scores do not summarize the data appropriately when a large variability exists between classes. This was demonstrated in the numeric example in Section 2 for which the within-class $F_1$ values are 0.308, 0.927, and 0.833, and $miF_1$, $maF_1$, and $maF_1^*$ are 0.870, 0.689, and 0.691, respectively. Reporting multiple within-class $F_1$ scores may be an option as done in [18] and [19]; however, an aggregate measure is useful in evaluating an overall performance of a classifier across classes. Another limitation with $F_1$ scores is that they do not take into consideration the true negative rate, and they may not be an appropriate measure when true negatives are important.

For future works, we are working on developing hypothesis testing procedure for $miF_1$, $maF_1$ and, $maF_1^*$ based on the variance estimates proposed in this article.

An R code for computing confidence intervals for $miF_1$, $maF_1$, and $maF_1^*$, is available and presented in Appendix D.

## Appendix A: Derivation of the distribution and variance of $\widehat{miF_1}$

Let $\boldsymbol{p}$ be the ordered elements of a confusion matrix. $\boldsymbol{p} = (p_{11}, \cdots, p_{1r},\ p_{21}, \cdots, p_{2r},\ \cdots,\ p_{r1}, \cdots, p_{rr})^T$. Using the multivariate delta-method for $\hat{p}$, we get

$$\sqrt{n}\left(\widehat{miF_1} - miF_1\right) \dot{\sim} Normal\left(0, \left[\frac{\partial(miF_1)}{\partial(\boldsymbol{p})}\right]^T \left(diag(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right)\left[\frac{\partial(miF_1)}{\partial(\boldsymbol{p})}\right]\right). \tag{5}$$

Because $miF_1 = \sum_{i=1}^{r} p_{ii}$ we have

$$\frac{\partial(miF_1)}{\partial(p_{ii})} = 1 \quad \forall\, i = 1, \cdots r \text{ and}$$

$$\frac{\partial(miF_1)}{\partial(p_{ij})} = 0 \quad \text{if } i \neq j.$$

And

$$\frac{\partial(miF_1)}{\partial(\boldsymbol{p})} = (1, 0, \cdots, 0,\ 0, 1, 0, \cdots, 0,\ \cdots,\ 0, \cdots, 0, 1)^T.$$

**Table 4** Point estimates and confidence intervals for $miF_1$, $maF_1$, and $maF_1^*$

| Method | $n$ | $\widehat{miF_1}$ | 95% CI | $\widehat{maF_1}$ | 95% CI | $\widehat{maF_1^*}$ | 95% CI |
|---|---|---|---|---|---|---|---|
| MNN | 59,066 | 0.859 | (0.856, 0.862) | 0.805 | (0.801, 0.809) | 0.807 | (0.803, 0.811) |
| SVM | 59,255 | 0.797 | (0.794, 0.800) | 0.750 | (0.746, 0.754) | 0.756 | (0.752, 0.760) |
| RF | 59,193 | 0.817 | (0.814, 0.820) | 0.724 | (0.720, 0.729) | 0.746 | (0.741, 0.750) |
| MLP | 59,130 | 0.814 | (0.811, 0.817) | 0.772 | (0.768, 0.776) | 0.778 | (0.774, 0.782) |

Note that all the elements corresponding to the diagonal entries ($p_{ii}$) of the confusion matrix is 1. To evaluate the variance in (5), further note that

$$diag(\boldsymbol{p}) = \begin{pmatrix} p_{11} & 0 & 0 & \cdots & 0 \\ 0 & p_{12} & 0 & \cdots & 0 \\ 0 & 0 & p_{13} & \cdots & 0 \\ \vdots & & & \ddots & \\ 0 & 0 & 0 & \cdots & p_{rr} \end{pmatrix},$$

$$\boldsymbol{p}\boldsymbol{p}^T = \begin{pmatrix} p_{11}^2 & p_{11}p_{12} & p_{11}p_{13} & \cdots & p_{11}p_{rr} \\ p_{12}p_{11} & p_{12}^2 & p_{12}p_{13} & \cdots & p_{12}p_{rr} \\ p_{13}p_{11} & p_{13}p_{12} & p_{13}^2 & \cdots & p_{13}p_{rr} \\ \vdots & & & \ddots & \\ p_{rr}p_{11} & p_{rr}p_{12} & p_{rr}p_{13} & \cdots & p_{rr}^2 \end{pmatrix}.$$

Then we have

$$\begin{aligned} &\left[\frac{\partial(miF_1)}{\partial(\boldsymbol{p})}\right]^T (diag(\boldsymbol{p}))\left[\frac{\partial(miF_1)}{\partial(\boldsymbol{p})}\right] \\ &= (p_{11}, 0, \cdots, p_{22}, 0, \cdots, p_{33}, \cdots, p_{rr})\left[\frac{\partial(miF_1)}{\partial(\boldsymbol{p})}\right] \\ &= \sum_{i=1}^{r} p_{ii}, \\ &\left[\frac{\partial(miF_1)}{\partial(\boldsymbol{p})}\right]^T \left(\boldsymbol{p}\boldsymbol{p}^T\right)\left[\frac{\partial(miF_1)}{\partial(\boldsymbol{p})}\right] \\ &= \left(\sum_{i=1}^{r} p_{ii}p_{11}, \sum_{i=1}^{r} p_{ii}p_{12}, \cdots, \sum_{i=1}^{r} p_{ii}p_{rr}\right)\left[\frac{\partial(miF_1)}{\partial(\boldsymbol{p})}\right] \\ &= \left(\sum_{i=1}^{r} p_{ii}p_{11} + \sum_{i=1}^{r} p_{ii}p_{22} + \cdots + \sum_{i=1}^{r} p_{ii}p_{rr}\right) \\ &= \left(\sum_{i=1}^{r} p_{ii}\right)^2. \end{aligned}$$

Thus,

$$\begin{aligned} &\left[\frac{\partial(miF_1)}{\partial(\boldsymbol{p})}\right]^T \left(diag(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right)\left[\frac{\partial(miF_1)}{\partial(\boldsymbol{p})}\right] \\ &= \left(\sum_{i=1}^{r} p_{ii}\right)\left(1 - \sum_{i=1}^{r} p_{ii}\right). \end{aligned}$$

Finally,

$$Var\left(\widehat{miF_1}\right) = \left(\sum_{i=1}^{r} p_{ii}\right)\left(1 - \sum_{i=1}^{r} p_{ii}\right)\Big/ n.$$

And

$$\widehat{miF_1} \dot{\sim} Normal\left(miF_1, \left(\sum_{i=1}^{r} p_{ii}\right)\left(1 - \sum_{i=1}^{r} p_{ii}\right)\Big/ n\right).$$

## Appendix B: Derivation of the distribution and variance of $\widehat{maF_1}$

In a similar manner to Appendix A, using the multivariate delta-method, we get

$$\begin{aligned} &\sqrt{n}\left(\widehat{maF_1} - maF_1\right) \dot{\sim} Normal\left(0, \left[\frac{\partial(maF_1)}{\partial(\boldsymbol{p})}\right]^T \right. \\ &\left. \times \left(diag(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right)\left[\frac{\partial(maF_1)}{\partial(\boldsymbol{p})}\right]\right). \end{aligned}$$

Now we take the partial derivatives of (2) to get

$$\begin{aligned} \frac{\partial(maF_1)}{\partial(p_{ii})} &= \frac{2}{r}\left(\frac{p_{i\cdot} + p_{\cdot i} - 2p_{ii}}{(p_{i\cdot} + p_{\cdot i})^2}\right), \quad \forall i = 1, \cdots r, \\ \frac{\partial(maF_1)}{\partial(p_{ij})} &= \frac{2}{r}\left(\frac{-p_{ii}}{(p_{i\cdot} + p_{\cdot i})^2} + \frac{-p_{jj}}{(p_{j\cdot} + p_{\cdot j})^2}\right), \\ & i, j = 1, \cdots, r; i \neq j. \end{aligned}$$

Arranging these terms according to the order of the elements in $\boldsymbol{p}$, we have

$$\begin{aligned} \frac{\partial(maF_1)}{\partial(\boldsymbol{p})} = \frac{2}{r}\left(\frac{p_{1\cdot} + p_{\cdot 1} - 2p_{11}}{(p_{1\cdot} + p_{\cdot 1})^2}, \frac{-p_{11}}{(p_{1\cdot} + p_{\cdot 1})^2}\right. \\ \left. + \frac{-p_{22}}{(p_{2\cdot} + p_{\cdot 2})^2}, \cdots, \frac{p_{r\cdot} + p_{\cdot r} - 2p_{rr}}{(p_{r\cdot} + p_{\cdot r})^2}\right)^T. \end{aligned}$$

Next, we note

$$\left[\frac{\partial(maF_1)}{\partial(\boldsymbol{p})}\right]^T \left(\boldsymbol{p}\boldsymbol{p}^T\right)\left[\frac{\partial(maF_1)}{\partial(\boldsymbol{p})}\right] = 0$$

because

$$\begin{aligned} \left[\frac{\partial(maF_1)}{\partial(\boldsymbol{p})}\right]^T \boldsymbol{p} &= \frac{2}{r}\left\{\sum_{i=1}^{r}\left(\frac{p_{i\cdot} + p_{\cdot i} - 2p_{ii}}{(p_{i\cdot} + p_{\cdot i})^2}\right)p_{ii} - \sum_{i=1}^{r}\left(\frac{\sum_{j\neq i} p_{ij}}{(p_{i\cdot} + p_{\cdot i})^2}\right)p_{ii} - \sum_{j=1}^{r}\left(\frac{\sum_{i\neq j} p_{ij}}{(p_{j\cdot} + p_{\cdot j})^2}\right)p_{jj}\right\} \\ &= \frac{2}{r}\left\{\sum_{i=1}^{r}\left(\frac{p_{i\cdot} + p_{\cdot i} - 2p_{ii}}{(p_{i\cdot} + p_{\cdot i})^2}\right)p_{ii} - \sum_{i=1}^{r}\left(\frac{\sum_{j\neq i}(p_{ij} + p_{ji})}{(p_{i\cdot} + p_{\cdot i})^2}\right)p_{ii}\right\} \\ &= 0. \end{aligned}$$

Therefore,

$$\left[\frac{\partial(maF_1)}{\partial(\boldsymbol{p})}\right]^T \left(diag(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right) \left[\frac{\partial(maF_1)}{\partial(\boldsymbol{p})}\right] = \left[\frac{\partial(maF_1)}{\partial(\boldsymbol{p})}\right]^T (diag(\boldsymbol{p})) \left[\frac{\partial(maF_1)}{\partial(\boldsymbol{p})}\right],$$

which can be shown to equal

$$= \frac{2}{r^2}\left\{\sum_{i=1}^{r} \frac{maF_{1i}(p_{i.}+p_{.i}-2p_{ii})}{(p_{i.}+p_{.i})^2}\left(\frac{p_{i.}+p_{.i}-2p_{ii}}{p_{i.}+p_{.i}} + \frac{maF_{1i}}{2}\right) + \sum_{i=1}^{r}\sum_{j\neq i} \frac{p_{ij}maF_{1i}maF_{1j}}{(p_{i.}+p_{.i})(p_{j.}+p_{.j})}\right\}.$$

Putting all together, we have

$$\widehat{maF_1} \dot{\sim} Normal\left(maF_1, Var\left(\widehat{maF_1}\right)\right),$$

where

$$Var\left(\widehat{maF_1}\right) = \frac{2}{r^2}\left\{\sum_{i=1}^{r} \frac{maF_{1i}(p_{i.}+p_{.i}-2p_{ii})}{(p_{i.}+p_{.i})^2}\left(\frac{p_{i.}+p_{.i}-2p_{ii}}{p_{i.}+p_{.i}} + \frac{maF_{1i}}{2}\right) + \sum_{i=1}^{r}\sum_{j\neq i} \frac{p_{ij}maF_{1i}maF_{1j}}{(p_{i.}+p_{.i})(p_{j.}+p_{.j})}\right\} \Big/ n.$$

## Appendix C: Derivation of the distribution and variance of $\widehat{maF_1^*}$

For marcro-averaged precision ($maP$) and macro-averaged recall ($maR$), let the vector $\boldsymbol{m}$ and its MLE $\hat{\boldsymbol{m}}$ be

$$\boldsymbol{m} = \begin{pmatrix} maP \\ maR \end{pmatrix}, \quad \hat{\boldsymbol{m}} = \begin{pmatrix} \widehat{maP} \\ \widehat{maR} \end{pmatrix},$$

respectively. Using the multivariate delta-method, we have

$$\sqrt{n}\left(\hat{\boldsymbol{m}} - \boldsymbol{m}\right) \dot{\sim} Normal\left(\boldsymbol{0}_2, \boldsymbol{\Sigma}\right),$$

where

$$\begin{aligned}\boldsymbol{\Sigma} &= \left[\frac{\partial(\boldsymbol{m})}{\partial(\boldsymbol{p}^T)}\right]\left(diag(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right)\left[\frac{\partial(\boldsymbol{m})}{\partial(\boldsymbol{p}^T)}\right]^T \\ &= \begin{pmatrix} \left[\frac{\partial(maP)}{\partial(\boldsymbol{p}^T)}\right]\left(diag(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right)\left[\frac{\partial(maP)}{\partial(\boldsymbol{p}^T)}\right]^T, & \left[\frac{\partial(maP)}{\partial(\boldsymbol{p}^T)}\right]\left(diag(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right)\left[\frac{\partial(maR)}{\partial(\boldsymbol{p}^T)}\right]^T \\ \left[\frac{\partial(maP)}{\partial(\boldsymbol{p}^T)}\right]\left(diag(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right)\left[\frac{\partial(maR)}{\partial(\boldsymbol{p}^T)}\right]^T, & \left[\frac{\partial(maR)}{\partial(\boldsymbol{p}^T)}\right]\left(diag(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right)\left[\frac{\partial(maR)}{\partial(\boldsymbol{p}^T)}\right]^T \end{pmatrix} \\ &= n\begin{pmatrix} Var(\widehat{maP}) & Cov(\widehat{maP}, \widehat{maR}) \\ Cov(\widehat{maP}, \widehat{maR}) & Var(\widehat{maR}) \end{pmatrix}.\end{aligned}$$

This is a $2 \times 2$ matrix with

$$Var\left(\widehat{maP}\right) = \frac{1}{r^2}\left(\sum_{i=1}^{r} \frac{p_{ii}\left(\sum_{j\neq i} p_{ij}\right)}{p_{i.}^3}\right) \Big/ n,$$

$$Var\left(\widehat{maR}\right) = \frac{1}{r^2}\left(\sum_{i=1}^{r} \frac{p_{ii}\left(\sum_{j\neq i} p_{ji}\right)}{p_{.i}^3}\right) \Big/ n,$$

$$Cov\left(\widehat{maP}, \widehat{maR}\right) = \frac{1}{r^2}\left\{\sum_{i=1}^{r} \frac{\left(\sum_{j\neq i} p_{ij}\right) p_{ii} \left(\sum_{j\neq i} p_{ji}\right)}{p_{i.}^2 p_{.i}^2} + \left(\sum_{i=1}^{r}\sum_{j\neq i} \frac{p_{ii}p_{ij}p_{jj}}{p_{i.}^2 p_{j.}^2}\right)\right\} \Big/ n.$$

Using the multivariate delta-method again, we get

$$\sqrt{n}\left(\widehat{maF_1^*} - maF_1^*\right) \dot{\sim} Normal\left(0, \left[\frac{\partial(maF_1^*)}{\partial(\boldsymbol{m})}\right]^T \boldsymbol{\Sigma} \left[\frac{\partial(maF_1^*)}{\partial(\boldsymbol{m})}\right]\right),$$

where

$$\frac{\partial(maF_1^*)}{\partial(\boldsymbol{m})} = \begin{pmatrix} \frac{2maR^2}{(maP+maR)^2} \\ \frac{2maP^2}{(maP+maR)^2} \end{pmatrix}.$$

Using this and $\boldsymbol{\Sigma}$ above, we obtain

$$\left[\frac{\partial(maF_1^*)}{\partial(\boldsymbol{m})}\right]^T \boldsymbol{\Sigma} \left[\frac{\partial(maF_1^*)}{\partial(\boldsymbol{m})}\right] = 4n\frac{maR^4Var\left(\widehat{maP}\right) + 2maP^2maR^2Cov\left(\widehat{maP},\widehat{maR}\right) + maP^4Var\left(\widehat{maR}\right)}{(maP+maR)^4}.$$

Finally, we have

$$\widehat{maF_1^*} \dot{\sim} Normal\left(maF_1^*, Var\left(\widehat{maF_1^*}\right)\right),$$

where

$$Var\left(\widehat{maF_1^*}\right) = 4n\frac{maR^4Var\left(\widehat{maP}\right) + 2maP^2maR^2Cov\left(\widehat{maP},\widehat{maR}\right) + maP^4Var\left(\widehat{maR}\right)}{(maP+maR)^4} \Big/ n\ .$$

## Appendix D: R code

The following R code computes point estimates and confidence intervals for $miF_1$, $maF_1$, and $maF_1^*$.

```
## Takahashi et al. ##
## Computation of F1 score and its confidence interval ##

f1scores <- function(mat, conf.level=0.95){
    ## This function computes point estimates and (conf.level*100%) confidence intervals
    ## for microF1, macroF1, and macroF1* scores.

    ## mat is an r by r matrix (confusion matrix).
    ## Rows indicate the predicted (fitted) conditions,
    ## and columns indicate the truth.
    ## miF1 is micro F1
    ## maF1 is macro F1
    ## maF2 is macro F1* (Sokolova and Lapalme)

    ## ###### ##
    ## Set up ##
    ## ###### ##
    r <- ncol(mat)
    n <- sum(mat) ## Total sample size
    p <- mat/n ## probabilities
        pii <- diag(p)
        pi. <- rowSums(p)
        p.i <- colSums(p)

    ## ################ ##
    ## Point estimates ##
    ## ################ ##
        miP <- miR <- sum(pii) ## MICRO precision, recall
    miF1 <- miP ## MICRO F1
```

```
    F1i <- 2*pii/(pi.+p.i)
maF1 <- sum(F1i)/r ## MACRO F1
    maP <- sum(pii/rowSums(p))/r ## MACRO precision
    maR <- sum(pii/colSums(p))/r ## MACRO recall
maF2 <- 2*(maP*maR)/(maP+maR) ## MACRO F1*

## ################### ##
## Variance estimates ##
## ################### ##

## ----------------- ##
## MICRO F1 Variance ##
## ----------------- ##
miF1.v <- sum(pii)*(1-sum(pii))/n
miF1.s <- sqrt(miF1.v)

## ----------------- ##
## MACRO F1 Variance ##
## ----------------- ##
```

```
for(i in 1:r){
        jj <- (1:r)[-i]
    for(j in jj){
        b <- b+ p[i,j]*F1i[i]*F1i[j]/((pi.[i]+p.i[i])*(pi.[j]+p.i[j]))
}}
maF1.v <- 2*(a+b)/(n*r^2)
maF1.s <- sqrt(maF1.v)

## ------------------ ##
## MACRO F1* Variance ##
## ------------------ ##
varmap <- sum(pii*(pi.-pii)/pi.^3) / r^2 / n
varmar <- sum(pii*(p.i-pii)/p.i^3) / r^2 / n
covmpr1 <- sum( ((pi.-pii) * pii * (p.i-pii)) / (pi.^2 * p.i^2) )
covmpr2 <- 0
    for(i in 1:r){
        covmpr2 <- covmpr2 + sum(pii[i] * p[i,-i] * pii[-i] / pi.[i]^2 / p.i[-i]^2)
    }
covmpr <- (covmpr1+covmpr2) / r^2 / n
maF2.v <- 4 * (maR^4*varmap + 2*maP^2*maR^2*covmpr + maP^4*varmar) / (maP+maR)^4
maF2.s <- sqrt(maF2.v)

## ##################### ##
## Confidence intervals ##
## ##################### ##
z <- qnorm(1-(1-conf.level)/2)
    miF1.ci <- miF1 + c(-1,1)*z*miF1.s
    maF1.ci <- maF1 + c(-1,1)*z*maF1.s
    maF2.ci <- maF2 + c(-1,1)*z*maF2.s
```

```
  ## ################## ##
  ##Formattnig output ##
  ## ################## ##
  pr <- data.frame(microPrecision=miP, microRecall=miR, macroPrecision=maP, macroRecall=maR)
  fss <- data.frame(
              rbind(miF1=c(miF1, miF1.s, miF1.ci),
                    maF1=c(maF1, maF1.s, maF1.ci),
                    maF1.star=c(maF2, maF2.s, maF2.ci)))
    names(fss) <- c('PointEst','Sd', 'Lower','Upper')
    out <- list(pr, fss)
    names(out) <- c('Precision.and.Recall', 'Confidence.Interval')
    out
}
```

```
## Example ##
## Table V from Dong et al. (2017) PMID: 28767373
mnn <- cbind(c(5022,577,188,19,395),
             c(407,2468,989,4,965),
             c(130,630,27254,1021,763),
             c(13,0,1236,6399,5),
             c(103,258,609,0,9611)
             )

f1scores(mnn)

## End ##
```

**Author Contributions** All authors contributed to the study conception. Mathematical derivation was primarily performed by Kanae Takahashi and Kouji Yamamoto, and it was confirmed by Tatsuki Koyama and Aya Kuchiba. The first draft of the manuscript was written by Kanae Takahashi with significant inputs from Kouji Yamamoto and Tatsuki Koyama. Aya Kuchiba made critical revisions to the first draft. The second draft was primarily written by Tatsuki Koyama, and all authors read and approved the final manuscript.

**Funding** This research was partially supported by Grant-in-Aid for Young Scientists No. 18K17325 (Takahashi), Grant-in-Aid for Scientific Research (C) No. 18K11195 (Yamamoto), and P30 CA068485 Cancer Center Support Grant (Koyama).

**Data Availability** Not applicable.

**Code Availability** The R code for computing point estimates and confidence intervals for $miF_1$, $maF_1$, and $maF_1^*$, is available in Appendix D.

## Declarations

**Conflict of Interests** None.

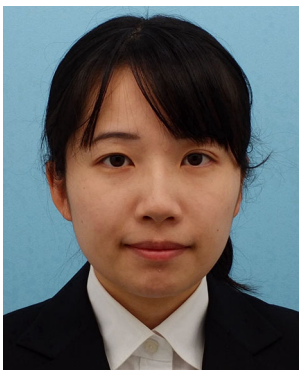


## References

1. van Rijsbergen CJ (1979) Information retrieval. Butterworths, Oxford
2. Manning CD, Raghavan P, Schütze H (2008) Introduction to information retrieval. Cambridge University Press, Cambridge
3. Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks. Inf Process Manage 45:427–437
4. Wang Y, Li J, Li Y, Wangi R, Yang X (2015) Confidence interval for $F_1$ measure of algorithm performance based on blocked $3 \times 2$ cross-validation. IEEE Trans Knowl Data Eng 27:651–659
5. Dong H, Supratak A, Pan W, Wu C, Matthews PM, Guo Y (2018) Mixed neural network approach for temporal sleep stage classification. IEEE Trans Neural Syst Rehabil Eng 26(2):324–333
6. Wang J, Zhang J, An Y, Lin H, Yang Z, Zhang Y, Sun Y (2016) Biomedical event trigger detection by dependency-based word embedding. BMC Med Genomics 2(9 Suppl):45

7. Socoró JC, Alías F, Alsina-Pagès RM (2017) An anomalous noise events detector for dynamic road traffic noise mapping in real-life urban and suburban environments. Sensors (Basel) 17(10)
8. Chowdhury S, Dong X, Qian L, Li X, Guan Y, Yang J, Yu Q (2018) A multitask bi-directional RNN model for named entity recognition on Chinese electronic medical records. BMC Bioinforma 19(Suppl 17):499
9. Troya-Galvis A, Gan˛carski P, Berti-Équille L (2018) Remote sensing image analysis by aggregation of segmentation-classification collaborative agents. Pattern Recognit 73:259–274
10. Hong N, Wen A, Stone DJ, Tsuji S, Kingsbury PR, Rasmussen LV, Pacheco JA, Adekkanattu P, Wang F, Luo Y, Pathak J, Liu H, Jiang G (2019) Developing a FHIRbased EHR phenotyping framework: A case study for identification of patients with obesity and multiple comorbidities from discharge summaries. J Biomed Inform 99:103310
11. Li L, Zhong B, Hutmacher C, Liang Y, Horrey WJ, Xu X (2020) Detection of driver manual distraction via image-based hand and ear recognition. Accid Anal Prev 137:105432
12. Zhou H, Ma Y, Li X (2020) Feature selection based on term frequency deviation rate for text classification. Appl Intell
13. Rashid MM, Kamruzzaman J, Hassan MM, Imam T, Gordon S (2020) Cyberattacks detection in IoT-based smart city applications using machine learning techniques. Int J Environ Res Public Health 17(24)
14. Wang SH, Nayak DR, Guttery DS, Zhang X, Zhang YD (2021) COVID-19 classification by CCSHNet with deep fusion using transfer learning and discriminant correlation analysis. Inf Fusion 68:131–148
15. Hao J, Yue K, Zhang B, Duan L, Fu X (2021) Transfer learning of bayesian network for measuring qos of virtual machines. Appl Intell
16. Li J, Lin M (2021) Ensemble learning with diversified base models for fault diagnosis in nuclear power plants. Ann Nucl Energy 158:108265
17. Zhang D, Wang J, Zhao X (2015) Estimating the uncertainty of average $F_1$ scores. In: Proceedings of the 2015 International conference on the theory of information retrieval
18. Zhu F, Li X, Mcgonigle D, Tang H, He Z, Zhang C, Hung GU, Chiu PY, Zhou W (2020) Analyze informant-based questionnaire for the early diagnosis of senile dementia using deep learning. IEEE J Transl Eng Health Med 8:2200106
19. Bhalla S, Kaur H, Kaur R, Sharma S, Raghava GPS (2020) Expression based biomarkers and models to classify early and late-stage samples of papillary thyroid carcinoma. PLoS One 15(4):e0231629

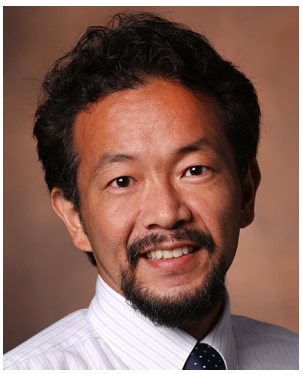


**Kanae Takahashi** is currently an Assistant Professor at Hyogo College of Medicine, Japan. She received the BS degree from Osaka University in 2008 and MPH degree from Kyoto University in 2010. Her research interests include design of clinical trials and diagnostic study.

**Kouji Yamamoto** is currently an Associate Professor at Yokohama City University, School of Medicine, Japan. He received the PhD in statistics from Tokyo University of Science in 2009. His research interests include design of clinical trials, diagnostic study, and categorical data analysis.

**Aya Kuchiba** is an Associate Professor at Graduate School of Health Innovation, Kanagawa University of Human Services, Japan. She received her PhD in Health Sciences (Biostatistics & Epidemiology) from University of Tokyo in 2008. Prior to joining Kanagawa University of Human Services, she was a Section Head of the Biostatistics Division at the National Cancer Center, Japan. Her research interest has focused on developing and applying statistical methods to cancer research in the areas of epidemiology with molecular and genetic data, diagnostic testing, and prevention, and in conducting clinical trials.



**Tatsuki Koyama** is an Associate Professor of Biostatistics at Vanderbilt University Medical Center. He received his PhD in statistics from University of Pittsburgh in 2003. His research interests are primarily centered on flexible experimental designs for clinical trials and inference from the data arising from such flexible and adaptive designs both in the Frequentist and Bayesian paradigms. His medical research interests include comparative effectiveness of treatments for localized prostate cancer, and association of ambient air pollution and acute lung injury.