HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2024 October 17.

Published in final edited form as:

Stat Med. 2023 October 15; 42(23): 4177–4192. doi:10.1002/sim.9853.

Hypothesis testing procedure for binary and multi-class F_1-scores in the paired design

Kanae Takahashi[1], **Kouji Yamamoto**[2], **Aya Kuchiba**[3], **Ayumi Shintani**[4], **Tatsuki Koyama**[5]

[1]Department of Biostatistics, Hyogo Medical University, Hyogo, Japan

[2]Department of Biostatistics, School of Medicine, Yokohama City University, Kanagawa, Japan

[3]Graduate School of Health Innovation, Kanagawa University of Human Services, Kanagawa, Japan

[4]Department of Medical Statistics, Osaka Metropolitan University Graduate School of Medicine, Osaka, Japan

[5]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, Tennessee, USA

Abstract

In modern medicine, medical tests are used for various purposes including diagnosis, disease screening, prognosis, and risk prediction. To quantify the performance of the binary medical test, we often use sensitivity, specificity, and negative and positive predictive values as measures. Additionally, the F_1-score, which is defined as the harmonic mean of precision (positive predictive value) and recall (sensitivity), has come to be used in the medical field due to its favorable characteristics. The F_1-score has been extended for multi-class classification, and two types of F_1-scores have been proposed for multi-class classification: a micro-averaged F_1-score and a macro-averaged F_1-score. The micro-averaged F_1-score pools per-sample classifications across classes and then calculates the overall F_1-score, whereas the macro-averaged F_1-score computes an arithmetic mean of the F_1-scores for each class. Additionally, Sokolova and Lapalme[1] gave an alternative definition of the macro-averaged F_1-score as the harmonic mean of the arithmetic means of the precision and recall over classes. Although some statistical methods of inference for binary and multi-class F_1-scores have been proposed, the methodology development of hypothesis testing procedure for them has not been fully progressing yet. Therefore, we aim to develop hypothesis testing procedure for comparing two F_1-scores in paired study design based on the large sample multivariate central limit theorem.

Keywords

delta-method; F_1 measures; multi-class classification; precision; recall

Correspondence: Tatsuki Koyama, Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, USA. tatsuki.koyama@vumc.org.

CONFLICT OF INTEREST STATEMENT

The authors have declared no conflict of interest.

1 | INTRODUCTION

Medical tests are important for the early detection and treatment of disease in modern medicine. Tests are used for various purposes including diagnosis, disease screening, prognosis, and risk prediction. Some measures exist to quantify the test performance; sensitivity, specificity, and positive and negative predictive values are commonly used for binary tests. Additionally, the F_1-score for binary data (binary F_1-score), which is defined as the harmonic mean of precision (positive predictive value) and recall (sensitivity), has been used in the medical field.[1,2]

The binary F_1-score is especially useful when evaluation of true negatives is relatively unimportant because true negatives are not included in computation of either precision or recall. In addition, the binary F_1-score performs well for a poor diagnostic test that identifies majority of the data as positive. In this situation, a simple arithmetic mean of precision and recall may be as high as 0.50 because recall will be 1.00 if all the data are diagnosed as positive. However, the binary F_1-score will be appropriately low in these instances: it will be 0.18 and 0.02 when the precision is 0.10 and 0.01, respectively, even if recall is 1.00. Therefore, F_1-score is a better statistic to report.[2]

Most of measures for performance of medical tests are only applicable to binary classification data, and multi-class classification data need to be dichotomized to compute these measures. In the motivating example,[3] for instance, skin cancer images were originally classified into six categories (malignant melanoma (MM), basal cell carcinoma (BCC), nevus, seborrheic keratosis (SK), senile lentigo (SL) and hematoma/hemangioma (H/H)), and the classification performances of board-certified dermatologists and dermatologic trainees were compared. The classification performance was assessed by accuracy, sensitivity, specificity, false negative rate, false positive rate, and positive predictive value after dichotomizing the six categories (MM and BCC vs. nevus, SK, SL, and H/H). However, evaluating the performance with the original six categories would have been preferable because dichotomization led to loss of information regarding the performance of the this classification.[4,5]

As measures of multi-class classification performance, a micro-averaged F_1-score and a macro-averaged F_1-score have been proposed.[2] The micro-averaged F_1-score calculates the overall F_1-score by pooling per-sample classifications across classes. Contrarily, the macro-averaged F_1-score computes an arithmetic mean of the F_1-scores for each class. In addition, Sokolova and Lapalme[6] proposed an alternative macro-averaged F_1-score as the harmonic mean of the arithmetic mean of the precisions and recalls for each class.

Although F_1-scores for binary and multi-class classifications have been originally used for measuring the performance of text classification in the field of information retrieval or of a classifier in machine learning, it has become frequently used in medicine.[7–14] Some statistical methods for inference have been proposed for the binary F_1-score,[15] and the methods for estimating confidence intervals of the micro-averaged F_1-scores and macro-averaged F_1-scores has been developed.[16,17] However, these previous methods are for

inference from one-sample. To our knowledge, no method is available for hypothesis testing of F_1-scores for paired samples as in our motivating example or two independent samples. Thus, we aim to provide the methods for comparing the binary F_1-scores, micro-averaged F_1-scores and macro-averaged F_1-scores in the paired-design setting. For two-independent-sample setting, the proposed method is readily applicable by setting the covariance part of the test statistics to 0.

The layout of this article is as follows: In Section 2, the definitions of the binary F_1-score, micro-averaged F_1-score and macro-averaged F_1-score are reviewed. Test statistics for comparing those scores are derived in Section 3. Then, the simulation results of the proposed statistics and the application to the motivating example are presented in Sections 4 and 5, respectively. Finally, our brief discussions are provided in Section 6.

2 | REVIEW OF F1-SCORES

This section introduces notations and definitions of binary F_1-score (biF), micro-averaged F_1-score (miF), and macro-averaged F_1-score (maF). Consider an $r \times r \times r$ table of data for a nominal categorical variable with r levels ($r \geq 2$). Each true class $1, \ \dots \ , \ r$ has an $r \times r$ table representing prediction frequencies of the two tests to be compared.

This arrangement of data represents the binary classification when $r = 2$, and the multi-class classification when $r > 2$. Table 1 shows general notations for each cell probability p_{ijk}, where i indicates the class of Test 1, j indicates the class of Test 2, and k indicates the true condition. Let Test 1 be a new medical test and Test 2 be an existing medical test. We consider a hypothesis testing to compare F_1-scores of Test 1 and Test 2. Using these notations, the true positive rate (TP_a), the false positive rate (FP_a), and the false negative rate (FN_a) for each class $a \ (a = 1, \ \dots \ , r)$ in Test 1 are defined as follows:

$$TP_{1a} = p_{a\,.\,a}, \quad FP_{1a} = \sum_{\substack{k=1 \\ k \neq a}}^{r} p_{a\,.\,k}, \quad FN_{1a} = \sum_{\substack{i=1 \\ i \neq a}}^{r} p_{i\,.\,a}.$$

Note that $TP_{1a} + FP_{1a} = p_{a\,.\,.}$, and $TP_{1a} + FN_{1a} = p_{.\,.\,a}$. Similarly, $TP_a, \ FP_a, \ FN_a$ for each class $a \ (a = 1, \ \dots \ , r)$ for Test 2 are defined as follows:

$$TP_{2a} = p_{.\,aa}, \quad FP_{2a} = \sum_{\substack{k=1 \\ k \neq a}}^{r} p_{.\,ak}, \quad FN_{2a} = \sum_{\substack{j=1 \\ j \neq a}}^{r} p_{.\,ja}.$$

Note that $TP_{2a} + FP_{2a} = p_{.\,a\,.}$, and $TP_{2a} + FN_{2a} = p_{.\,.\,a}$.

2.1 | Binary F_1-score

When $r = 2$, we consider the following precision (biP) and recall (biR) for Test 1 as:

$$biP_1 = \frac{TP_{11}}{TP_{11} + FP_{11}} = p_{1.1}/p_{1..},$$

$$biR_1 = \frac{TP_{11}}{TP_{11} + FN_{11}} = p_{1.1}/p_{..1}.$$

And binary F_1-score for Test 1 (biF_1) is defined as the harmonic mean of biP_1 and biR_1, that is,

$$biF_1 = 2\frac{biP_1 \times biR_1}{biP_1 + biR_1} = 2\frac{p_{1.1}}{p_{1..} + p_{..1}}. \quad (1)$$

Similarly, the binary F_1-score for Test 2 (biF_2) is as follows:

$$biP_2 = \frac{TP_{21}}{TP_{21} + FP_{21}} = p_{.11}/p_{.1.},$$

$$biR_2 = \frac{TP_{21}}{TP_{21} + FN_{21}} = p_{.11}/p_{..1},$$

$$biF_2 = 2\frac{biP_2 \times biR_2}{biP_2 + biR_2} = 2\frac{p_{.11}}{p_{.1.} + p_{..1}}. \quad (2)$$

2.2 | Micro-averaged F_1-score

When $r > 2$ the micro-averaged precision (miP) and micro-averaged recall (miR) are obtained from the sum of each class of TP_i, FP_i, FN_i. miP and miR for Test 1 can be written as

$$miP_1 = \frac{\sum_{a=1}^{r} TP_{1a}}{\sum_{a=1}^{r}(TP_{1a} + FP_{1a})} = \frac{\sum p_{a.a}}{\sum p_{a..}} = \sum_{a=1}^{r} p_{a.a},$$

$$miR_1 = \frac{\sum_{a=1}^{r} TP_{1a}}{\sum_{a=1}^{r}(TP_{1a} + FN_{1a})} = \frac{\sum p_{a.a}}{\sum p_{..a}} = \sum_{a=1}^{r} p_{a.a}.$$

Finally, as the harmonic mean of miP_1 and miR_1, we have the micro-averaged F_1-score for Test 1 (miF_1) as

$$miF_1 = 2\frac{miP_1 \times miR_1}{miP_1 + miR_1} = \sum_{a=1}^{r} p_{a\cdot a}. \tag{3}$$

Similarly, the micro-averaged F_1-score for Test 2 (miF_2) is

$$miP_2 = \frac{\sum_{a=1}^{r} TP_{2a}}{\sum_{a=1}^{r}(TP_{2a} + FP_{2a})} = \frac{\sum p_{\cdot aa}}{\sum p_{\cdot a\cdot}} = \sum_{a=1}^{r} p_{\cdot aa},$$

$$miR_2 = \frac{\sum_{a=1}^{r} TP_{2a}}{\sum_{a=1}^{r}(TP_{2a} + FN_{2a})} = \frac{\sum p_{\cdot aa}}{\sum p_{\cdot\cdot a}} = \sum_{a=1}^{r} p_{\cdot aa}.$$

$$miF_2 = 2\frac{miP_2 \times miR_2}{miP_2 + miR_2} = \sum_{a=1}^{r} p_{\cdot aa}. \tag{4}$$

2.3 | Macro-averaged F_1-score

When $r > 2$, to define the macro-averaged F_1-score for Test 1 (maF_1), first consider the following precision (P_{1a}) and recall (R_{1a}) within each class, $a = 1, \ldots, r$:

$$P_{1a} = \frac{TP_{1a}}{TP_{1a} + FP_{1a}} = p_{a\cdot a}/p_{a\cdot\cdot}, \tag{5}$$

$$R_{1a} = \frac{TP_{1a}}{TP_{1a} + FN_{1a}} = p_{a\cdot a}/p_{\cdot\cdot a}. \tag{6}$$

And F_1-score within each class for Test 1 (F_{1a}) is defined as the harmonic mean of P_{1a} and R_{1a}, that is,

$$F_{1a} = 2\frac{P_{1a} \times R_{1a}}{P_{1a} + R_{1a}} = 2\frac{p_{a\cdot a}}{p_{a\cdot\cdot} + p_{\cdot\cdot a}}.$$

The macro-averaged F_1-score for Test 1 (maF_1) is the simple arithmetic mean of F_{1a}:

$$maF_1 = \frac{1}{r}\sum_{a=1}^{r} F_{1a} = \frac{2}{r}\sum_{a=1}^{r} \frac{p_{a\cdot a}}{p_{a\cdot\cdot} + p_{\cdot\cdot a}}.$$

(7)

Similarly, the macro-averaged F_1-score for Test 2 (maF_2) is

$$P_{2a} = \frac{TP_{2a}}{(TP_{2a} + FP_{2a})} = p_{.aa}/p_{.a.},$$

$$R_{2a} = \frac{TP_{2a}}{(TP_{2a} + FN_{2a})} = p_{.aa}/p_{..a}.$$

$$F_{2a} = 2\frac{P_{2a} \times R_{2a}}{P_{2a} + R_{2a}} = 2\frac{p_{.aa}}{p_{.a.} + p_{..a}}.$$

$$maF_2 = \frac{1}{r}\sum_{a=1}^{r} F_{2a} = \frac{2}{r}\sum_{a=1}^{r}\frac{p_{.aa}}{p_{.a.} + p_{..a}}.$$

(8)

2.4 | Alternate definition of macro-averaged F_1-score

Sokolova and Lapalme[6] gave an alternative definition of the macro-averaged F_1. First, macro-averaged precision (maP) and macro-averaged recall (maR) for Test 1 are defined as simple arithmetic means of the within-class precision and within-class recall in (5) and (6), respectively.

$$maP_1 = \frac{1}{r}\sum_{a=1}^{r} P_{1a} = \frac{1}{r}\sum_{a=1}^{r}\frac{p_{a.a}}{p_{a..}},$$

$$maR_1 = \frac{1}{r}\sum_{a=1}^{r} R_{1a} = \frac{1}{r}\sum_{a=1}^{r}\frac{p_{a.a}}{p_{..a}}.$$

And the alternate definition of macro-averaged F_1-score for Test 1 (maF_1^*) is the harmonic mean of these quantities.

$$maF_1^* = 2\frac{maP_1 \times maR_1}{maP_1 + maR_1}.$$

(9)

Similarly, the alternate definition of macro-averaged F_1-score for Test 2 (maF_2^*) is

$$maP_2 = \frac{1}{r}\sum_{a=1}^{r} P_{2a} = \frac{1}{r}\sum_{a=1}^{r} \frac{p_{\cdot aa}}{p_{\cdot a \cdot}},$$

$$maR_2 = \frac{1}{r}\sum_{a=1}^{r} R_{2a} = \frac{1}{r}\sum_{a=1}^{r} \frac{p_{\cdot aa}}{p_{\cdot \cdot a}}.$$

$$maF_2^* = 2\frac{maP_2 \times maR_2}{maP_2 + maR_2}. \tag{10}$$

3 | PROPOSED HYPOTHESIS TESTING PROCEDURE

In this section, we derive the test statistics for comparing two F_1-scores (biF_1 and biF_2; miF_1 and miF_2; maF_1 and maF_2; and maF_1^* and maF_2^*). We assume that the observed frequencies, n_{ijk} for $1 \le i \le r$, $1 \le j \le r$, $1 \le k \le r$, have a multinomial distribution with overall sample size $N = \sum_{i,j,k} n_{ijk}$ and probabilities $\boldsymbol{p} = [p_{111}, \ldots, p_{1r1}, \cdots, p_{rr1}, \ldots, p_{rrr}]^T$, where i indicates the class of Test 1, j indicates the class of Test 2, k indicates the true condition, and "T" represents the transpose. The maximum likelihood estimate (MLE) of p_{ijk} is $\hat{p}_{ijk} = n_{ijk}/N$. That is

$$(n_{111}, n_{121}, \ldots, n_{rrr}) \sim Multinomial(N; \boldsymbol{p}).$$

By invariance property of MLE's, the maximum likelihood estimate of biF, miF, maF, maF^*, and other quantities in the previous section can be obtained by substituting p_{ijk} by $\hat{p}_{ijk}$.

3.1 | Test statistic for comparing two $biFs$

Let $\boldsymbol{biF} = (biF_1, biF_2)^T$ be a vector whose components are the $biFs$ of the two medical tests, and let $\widehat{\boldsymbol{biF}}$ be the MLE of $\boldsymbol{biF}$. $\widehat{\boldsymbol{biF}}$ can be obtained by substituting p_{ijk} by their MLE's in (1) and (2).

$$\widehat{biF_1} = 2\frac{\hat{p}_{1.1}}{\hat{p}_{1..} + \hat{p}_{..1}} = 2\frac{n_{1.1}}{n_{1..} + n_{..1}}, \qquad \widehat{biF_2} = 2\frac{\hat{p}_{.11}}{\hat{p}_{.1.} + \hat{p}_{..1}} = 2\frac{n_{.11}}{n_{.1.} + n_{..1}}.$$

Using the delta-method and the multivariate central limit theorem, we have

$$\sqrt{N}\left(\widehat{\boldsymbol{biF}} - \boldsymbol{biF}\right) \rightsquigarrow Normal\left(0, \left[\frac{\partial(\boldsymbol{biF})}{\partial(\boldsymbol{p})}\right]^T \left[diag(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right] \left[\frac{\partial(\boldsymbol{biF})}{\partial(\boldsymbol{p})}\right]\right),$$

where $diag(\boldsymbol{p})$ is an $r^2 \times r^2 \times r^2$ diagonal matrix whose elements are the diagonal elements of $\boldsymbol{p}$, and "$\sim$" represents "approximately distributed as". The Wald statistic for testing $H_0: biF_1 = biF_2$ vs. $H_1: biF_1 \neq biF_2$, therefore, is

$$T_{biF}^{W} = \frac{\left(\widehat{biF}_1 - \widehat{biF}_2\right)^2}{\widehat{Var}_{biFd}},$$

where $\widehat{Var}_{biFd}$ is the variance of $\left(\widehat{biF}_1 - \widehat{biF}_2\right)$ with $\{p_{ijk}\}$ replaced by $\{\hat{p}_{ijk}\}$. Derivation of the variance of $\left(\widehat{biF}_1 - \widehat{biF}_2\right)$ appear in Appendix A.1. The test statistic is distributed asymptotically as a χ^2 distribution with one degree of freedom under the null hypothesis.

As a side note, the confidence interval of biF for each test can be derived in the same way. A $(1 - \alpha) \times 100\%$ confidence interval of biF_1 and biF_2 is

$$\widehat{biF}_1 \pm Z_{1-\alpha/2} \times \sqrt{\widehat{Var}_{bF_1}},$$

$$\widehat{biF}_2 \pm Z_{1-\alpha/2} \times \sqrt{\widehat{Var}_{biF_2}},$$

where Z_p denote the $100p$-th percentile of the standard normal distribution, and $\widehat{Var}_{biF_1}$ and $\widehat{Var}_{biF_2}$ are the variance of $\widehat{biF}_1$ and the variance of $\widehat{biF}_2$ with $\{p_{ijk}\}$ replaced by $\{\hat{p}_{ijk}\}$. These simple formulas based on the multinomial distribution have not been proposed yet. Wang et al. proposed a confidence interval of biF based on the beta prime distribution and associated calculations using the bootstrap method.[15,18]

For the score statistic, we consider the MLE of $\{p_{ijk}\}$ under the null hypothesis that could be obtained, for example by applying the Newton-Raphson method to the log-likelihood equations. The score statistic for testing $H_0: biF_1 = biF_2$ vs. $H_1: biF_1 \neq biF_2$ is

$$T_{biF}^{S} = \frac{\left(\widehat{biF}_1 - \widehat{biF}_2\right)^2}{\widetilde{Var}_{biFd}},$$

where $\widetilde{Var}_{biFd}$ is the variance of $\left(\widehat{biF}_1 - \widehat{biF}_2\right)$ with $\{p_{ijk}\}$ replaced by $\{\tilde{p}_{ijk}\}$, that is calculated from the MLE of $\{p_{ijk}\}$ under the null hypothesis.

3.2 | Test statistic for comparing two $miFs$

As shown in (3) and (4), $miF_1 = \sum p_{a.a}$, $miF_2 = \sum p_{.aa}$, and the MLE of miF_1 and miF_2 are

$$\widehat{miF}_1 = \sum_{a=1}^{r} \hat{p}_{a.a} = \sum_{a=1}^{r} \frac{n_{a.a}}{N}, \widehat{miF}_2 = \sum_{a=1}^{r} \hat{p}_{.aa} = \sum_{a=1}^{r} \frac{n_{.aa}}{N}.$$

Again by the delta-method and multivariate central limit theorem (Appendix A.2), the Wald statistic for testing $H_0: miF_1 = miF_2$ versus $H_1: miF_1 \neq miF_2$ is

$$T^W_{miF} = \frac{\left(\widehat{miF}_1 - \widehat{miF}_2\right)^2}{\widehat{Var}_{miFd}},$$

where $\widehat{Var}_{miFd}$ is the variance of $\left(\widehat{miF}_1 - \widehat{miF}_2\right)$ with $\{p_{ijk}\}$ replaced by $\{\hat{p}_{ijk}\}$. The test statistic is distributed a symptotically as a χ^2 distribution with one degree of freedom under the null hypothesis.

Again to develop the score statistic, we consider the MLE of $\{p_{ijk}\}$ under the null hypothesis as in the case of biF. The score statistic for testing $H_0: miF_1 = miF_2$ versus $H_1: miF_1 \neq miF_2$ is

$$T^S_{miF} = \frac{\left(\widehat{miF}_1 - \widehat{miF}_2\right)^2}{\widetilde{Var}_{miFd}},$$

where $\widetilde{Var}_{miFd}$ is the variance of $\left(\widehat{miF}_1 - \widehat{miF}_2\right)$ with $\{p_{ijk}\}$ replaced by $\{\tilde{p}_{ijk}\}$, that is calculated from the MLE of $\{p_{ijk}\}$ under the null hypothesis.

3.3 | Test statistic for comparing two $maFs$

The MLE of maF_1 and maF_2 can be obtained by substituting $p_{a..}$, $p_{.aa}$, $p_{a..}$, $p_{.a.}$ and $p_{..a}$ by their MLE's in (7) and (8).

$$\widehat{maF}_1 = \frac{2}{r}\sum_{a=1}^{r}\frac{\hat{p}_{a..}}{\hat{p}_{a..} + \hat{p}_{..a}} = \frac{2}{r}\sum_{a=1}^{r}\frac{n_{a..}}{n_{a..} + n_{..a}}, \quad \widehat{maF}_2 = \frac{2}{r}\sum_{a=1}^{r}\frac{\hat{p}_{.aa}}{\hat{p}_{.a.} + \hat{p}_{..a}} = \frac{2}{r}\sum_{a=1}^{r}\frac{n_{.aa}}{n_{.a.} + n_{..a}}.$$

Again by the delta-method and multivariate central limit theorem (Appendix A.3), we have the Wald statistic for testing $H_0: maF_1 = maF_2$ versus $H_1: maF_1 \neq maF_2$ as

$$T^W_{maF} = \frac{\left(\widehat{maF}_1 - \widehat{maF}_2\right)^2}{\widehat{Var}_{maFd}},$$

where $\widehat{Var}_{maFd}$ is the variance of $\left(\widehat{maF}_1 - \widehat{maF}_2\right)$ with $\{p_{ijk}\}$ replaced by $\{\hat{p}_{ijk}\}$. The test statistic is distributed asymptotically as a χ^2 distribution with one degree of freedom under the null hypothesis.

For the score statistic, we consider the MLE of $\{p_{ijk}\}$ under the null hypothesis as in the case of biF and miF. The score statistic for testing $H_0: maF_1 = maF_2$ versus $H_1: maF_1 \neq maF_2$ is

$$T^S_{maF} = \frac{\left(\widehat{maF}_1 - \widehat{maF}_2\right)^2}{\widetilde{Var}_{maFd}},$$

where $\widetilde{Var}_{maFd}$ is the variance of $\left(\widehat{maF}_1 - \widehat{maF}_2\right)$ replaced by $\{\tilde{p}_{ijk}\}$, that is calculated from the MLE of $\{p_{ijk}\}$ under the null hypothesis.

3.4 | Test statistic for comparing two maF^*s

To obtain the MLEs of maF_1^* and maF_2^*, we first substitute $p_{a..a}$, $p_{.aa}$, $p_{a..}$, $p_{.a.}$ and $p_{..a}$ by their MLE's to get MLE's of maP and maR and use these in (9) and (10):

$$\widehat{maF}_1^* = 2\frac{\widehat{maP}_1 \times \widehat{maR}_1}{\widehat{maP}_1 + \widehat{maR}_1}, \quad \widehat{maF}_2^* = 2\frac{\widehat{maP}_2 \times \widehat{maR}_2}{\widehat{maP}_2 + \widehat{maR}_2}.$$

Using the delta-method and multivariate central limit theorem (Appendix A.4), we have the Wald statistic for testing $H_0: maF_1^* = maF_2^*$ versus $H_1: maF_1^* \neq maF_2^*$ as

$$T_{maF}^W = \frac{\left(\widehat{maF}_1^* - \widehat{maF}_2^*\right)^2}{\widehat{Var}_{maFd^*}},$$

Again to get $\widehat{Var}_{maFd^*}$, all components of the variance of $\left(\widehat{maF}_1^* - \widehat{maF}_2^*\right)$ are replaced by their respective MLE's. The test statistic is distributed asymptotically as a χ^2 distribution with one degree of freedom under the null hypothesis.

On the other hand, for the score statistic, we consider the MLE of $\{p_{ijk}\}$ under the null hypothesis as in the case of biF, miF, and maF. The score statistic for testing $H_0: maF_1^* = maF_2^*$ versus $H_1: maF_1^* \neq maF_2^*$ is

$$T_{maF}^S = \frac{\left(\widehat{maF}_1^* - \widehat{maF}_2^*\right)^2}{\widetilde{Var}_{maFd^*}},$$

where $\widetilde{Var}_{maFd^*}$ is the variance of $\left(\widehat{maF}_1^* - \widehat{maF}_2^*\right)$ replaced by $\{\tilde{p}_{ijk}\}$, that is calculated from the MLE of $\{p_{ijk}\}$ under the null hypothesis.

4 | SIMULATION

4.1 | Simulation setup

A simulation study was conducted to evaluate the performance of the test statistics proposed in Section 3. We set $r = 3$ (class 1, 2, 3), and generated data according to the multinomial distributions with $\boldsymbol{p}$ shown in Table 2. Classes 2 and 3 were combined when calculating biF. The total sample size, N, was set to 100, 300, 500, and 1,000. The nominal type I error rate was set to 0.05 (two-sided test). We used the empirical type I error rate and empirical power as performance measures. For each combination of the scenario and sample size, we performed 100,000 repeated simulations.

Scenarios 1 and 2 are set up to evaluate the empirical type I error rate of the proposed test statistics, while scenario 3 and 4 are designed to assess their empirical power. In

scenario 1, the true conditions of classes 1, 2, and 3 have the same probability (1⁄3), and the recalls and precisions within each class are equal in the two tests (60%). Thus, $maR_1 = maR_2 = maP_1 = maP_2 = 0.60$, and $F_{1a} = F_{2a} = 0.60$ for each class, $a = 1, 2, 3$. Then, $maF_1 = maF_2 = maF_1^* = maF_2^* = 0.60$. Because classes 2 and 3 are combined to calculate $biF_1 = F_{11} = 0.60$ and $biF_2 = F_{21} = 0.60$. Also, $p_{a\cdot a} = p_{\cdot aa} = 0.20$ for each class $a = 1, 2, 3$, and $miF_1 = miF_2 = 0.60$.

In scenario 2, the true condition of class 1 has higher probability than the others (60% vs. 20%), and performances of two tests are equal: $biF_1 = biF_2 = 0.69$, $miF_1 = miF_2 = 0.60$, $maF_1 = maF_2 = 0.56$, and $maF_1^* = maF_2^* = 0.58$. Although the distributions in scenario 2 are the same as those in scenario 1 for each class, the value of biF, maF and maF^* are different between scenarios because $TP_a/(TP_a + FP_a)$ is large in the true class = 1 and, conversely, relatively small in the true classes 2 and 3. In contrast, miF in scenario 2 is the same as that in scenario 1 because $p_{a\cdot a}$ and $p_{\cdot aa}$ for each class $a = 1, 2, 3$ in scenarios 1 and 2 are equal.

The true conditions of classes 1, 2, and 3 have the same probability (1⁄3) in scenario 3. However, maR and maP of Test 2 are lower than Test 1 (60% vs. 50%), F_{2a} are lower than F_{1a} (60% vs. 50%), and $p_{\cdot aa}$ is lower than $p_{a\cdot a}$ for each class $a = 1, 2, 3$ (20% vs. 17%). Therefore, $biF_1 = miF_1 = maF_1 = maF_1^* = 0.60$, whereas $biF_2 = miF_2 = maF_2 = maF_2^* = 0.50$.

In scenario 4, the true condition of class 1 has higher probability than the others (60% vs. 20%) as in scenario 2. However, the performance of two tests are different: $biF_1 = 0.69$ versus $biF_2 = 0.60$, $miF_1 = 0.60$ versus $miF_2 = 0.50$, $maF_1 = 0.56$ versus $maF_2 = 0.47$, and $maF_1^* = 0.58$ versus $maF_2^* = 0.49$.

4.2 | Simulation result

Table 3 shows the empirical type I error rates of the proposed tests for scenarios 1 and 2. The empirical type I error rates for both test statistics were close to nominal type I error rate of 0.05 when the sample size is large (300, 500, 1000). When N is relatively small (100), the empirical type I error rates tended to be slightly larger than 0.05, especially for Wald statistics. Contrarily, the empirical type I error rates with score statistics are close to the nominal type I error rate of 0.05 for all sample sizes. Table 4 shows the empirical power of the proposed tests for scenarios 3 and 4. As shown in Table 4, the empirical powers increase with the sample size. The empirical powers of Wald statistics and score statistics are similar, especially when the sample size is large.

5 | EXAMPLE

We describe an application of the proposed hypothesis testing procedure to the motivating example.[3] In this study, a skin cancer classification system with faster, region-based convolutional neural network algorithm (FRCNN) for brown to black pigmented skin lesions was developed using a deep learning method. The target diseases were malignant tumors (malignant melanoma (MM) and basal cell carcinoma (BCC)) and benign tumors (nevus, seborrheic keratosis (SK), senile lentigo (SL) and hematoma/hemangioma (H/H)), and 2000

images were evaluated. The 2000 images were obtained by randomly sampling 200 images from the 666 images 10 times. For illustration, all images were treated as independent in this study. The data are shown in Tables B1–B3, Appendix B. Although images were classified into six categories (MM, BCC, nevus, SK, SL, H/H), accuracy was the only performance measure computed for six-class classification data in the motivating example. Other performance measures, sensitivity, specificity, false negative, false positive, and positive predictive value, were calculated for two-class classification data after combining malignant tumors (MM and BCC) and benign tumors (nevus, SK, SL, and H/H). The accuracy of six-class classification by the FRCNN (86.2% ± 2.95%) was statistically higher than that of board-certified dermatologists (BCD) (79.5% ± 5.27%, $p = 0.0081$) and that of dermatologic trainees (75.1% ± 2.18%, $p < 0.0001$).

We compared the performance of skin cancer classification between the FRCNN and BCD using biFs, miFs, maFs, and maF*s with the proposed hypothesis testing procedures. miFs, maFs, and maF*s were calculated from six-class classification data, while biFs were calculated from two-class classification data (malignant tumors vs. benign tumors). The results are shown in Table 5. All biFs, miFs, maFs, and maF*s of six-class classification by FRCNN were significantly higher than those by BCD.

6 | DISCUSSION

We developed hypothesis testing procedures for comparing two F_1-scores (biF_1 and biF_2, miF_1 and miF_2, maF_1 and maF_2, and maF_1^* and maF_2^*) in paired study design. Through the simulation study and motivating example, we assessed the performance and feasibility of those testing procedures. We conclude that the method based on the score statistics (T^S_{biF}, T^S_{miF}, T^S_{maF}, and T^S_{maF*}) is slightly better compared to the method based on the Wald statistics (T^W_{biF}, T^W_{miF}, T^W_{maF}, and T^W_{maF*}) because the empirical type I error rate is closer to the nominal level even when the sample size is small. However, when multi-class classification is considered, typical sample size is much larger than 100, and both approaches perform equally well in such scenarios.

We did not observe a substantial disparity in the empirical powers of the two approaches.

At present, others have not studied hypothesis testing procedure of biFs, miFs, maFs, and maF*s, and only the point estimates of these scores were reported in most studies. Han et al[19] applied one sample t-test for comparison of biFs; however, this approach may not be appropriate because biF is the harmonic mean of precision and recall, and the distribution of the difference between two biFs is unlikely to follow a Student's t-distribution.

A limitation of this work is that the proposed procedures are based on the large sample theory, and thus require a large sample size to provide strict control of the type I error rate. For future works, we are working on the exact test for comparing two biFs, miFs, maFs, and maF*s based on the methods presented in this article.

An R code for computing point estimates, Wald statistics, score statistics, and p-values for biFs, miFs, maFs, and maFs of each statistic in the paired design, is available on the lead

author's GitHub page: https://github.com/kanaet52/f1score/blob/main/R/F1score_test.R. For two-sample designs, the F_1-scores can be compared by setting the covariance part of the test statistic (C^W_{biF}, C^W_{miF}, C^W_{maF}, C^W_{maF*}, see Appendix A) to 0. Note that for the score statistics, the MLE of $\{p_{ijk}\}$ under each null hypothesis is obtained by applying the Newton-Raphson method to the log-likelihood equations in the code.

ACKNOWLEDGEMENTS

The authors would like to thank Dr Shunichi Jinnai for providing the motivating example data. This research was partially supported by grant-in-aid for Scientific Research (C) No. 18K11195 and 21K11790 (Yamamoto), grant-in-aid for Research Activity start-up no. 21K21170 (Takahashi), and P30CA068485 Cancer Center Support grant (Koyama).

Funding information

Cancer Center Support, Grant/Award Number: P30CA068485; Japan Society for the Promotion of Science, Grant/Award Numbers: 18K11195, 21K11790, 21K21170

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

APPENDIX A.: DERIVATION OF VARIANCES

A.1 Variance of biF

The derivation of the variance of $\left(\widehat{biF}_1 - \widehat{biF}_2\right)$ is as follows:

$$\boldsymbol{biF} = \begin{bmatrix} biF_1 \\ biF_2 \end{bmatrix},$$

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

$$\left[\frac{\partial(\boldsymbol{biF})}{\partial(\boldsymbol{p})}\right]^T \left(diag\left(\boldsymbol{p}\right) - \boldsymbol{p}\boldsymbol{p}^T\right) \left[\frac{\partial(\boldsymbol{biF})}{\partial(\boldsymbol{p})}\right] = \left[\frac{\partial(\boldsymbol{biF})}{\partial(\boldsymbol{p})}\right]^T \left(diag(\boldsymbol{p})\right) \left[\frac{\partial(\boldsymbol{biF})}{\partial(\boldsymbol{p})}\right] = \begin{bmatrix} A^W_{biF} & C^W_{biF} \\ C^W_{biF} & B^W_{biF} \end{bmatrix},$$

with

$$A^W_{biF} = \frac{1}{(p_{1..} + p_{..1})^2}\left[p_{1.1}(2(1 - biF_1))^2 + (p_{1.2} + p_{2.1})biF_1^2\right]$$

$$B^W_{biF} = \frac{1}{(p_{.1.} + p_{..1})^2}\left[p_{.11}(2(1 - biF_2))^2 + (p_{.12} + p_{.21})biF_2^2\right]$$

$$C^W_{biF} = \frac{1}{(p_{1..} + p_{..1})(p_{.1.} + p_{..1})}\left[2^2 p_{111}(1 - biF_1)(1 - biF_2) - 2p_{121}(1 - biF_1)biF_2 \right.$$
$$\left. -2p_{211}biF_1(1 - biF2) + (p_{221} + p_{112})biF_1 biF_2\right].$$

Therefore, the variance of $\widehat{biF_1}$ is

$$\frac{1}{(p_{1..} + p_{.\cdot 1})^2}\left[p_{1.1}(2(1-biF_1))^2 + (p_{1.2}+p_{2.1})biF_1^2\right]/n,$$

the variance of $\widehat{biF_2}$ is

$$\frac{1}{(p_{.1.} + p_{.\cdot 1})^2}\left[p_{.11}(2(1-biF_2))^2 + (p_{.12}+p_{.21})biF_2^2\right]/n,$$

the variance of $\left(\widehat{biF_1} - \widehat{biF_2}\right)$ is

$$\left(A_{biF}^W + B_{biF}^W - 2C_{biF}^W\right)/n\,.$$

A.2 Variance of miF

The derivation of the variance of $\left(\widehat{miF_1} - \widehat{miF_2}\right)$ is as follows:

$$\boldsymbol{miF} = \begin{bmatrix} miF_1 \\ miF_2 \end{bmatrix},$$

$$\left[\frac{\partial(\boldsymbol{miF})}{\partial(\boldsymbol{p})}\right]^T \left(\boldsymbol{p}\boldsymbol{p}^T\right)\left[\frac{\partial(\boldsymbol{miF})}{\partial(\boldsymbol{p})}\right] = \begin{bmatrix} \left(\sum_{a=1}^r p_{a.a}\right)^2 & \left(\sum_{a=1}^r p_{a.a}\right)\left(\sum_{a=1}^r p_{.aa}\right) \\ \left(\sum_{a=1}^r p_{a.a}\right)\left(\sum_{a=1}^r p_{.aa}\right) & \left(\sum_{a=1}^r p_{.aa}\right)^2 \end{bmatrix} = \begin{bmatrix} miF_1^2 & miF_1 miF_2 \\ miF_1 miF_2 & miF_2^2 \end{bmatrix},$$

$$\left[\frac{\partial(\boldsymbol{miF})}{\partial(\boldsymbol{p})}\right]^T (diag(\boldsymbol{p}))\left[\frac{\partial(\boldsymbol{miF})}{\partial(\boldsymbol{p})}\right] = \begin{bmatrix} \sum_{a=1}^r p_{a.a} & \sum_{a=1}^r p_{aaa} \\ \sum_{a=1}^r p_{aaa} & \sum_{a=1}^r p_{.aa} \end{bmatrix} = \begin{bmatrix} miF_1 & \sum_{a=1}^r p_{aaa} \\ \sum_{a=1}^r p_{aaa} & miF_2 \end{bmatrix},$$

$$\left[\frac{\partial(\boldsymbol{miF})}{\partial(\boldsymbol{p})}\right]^T \left(diag(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right)\left[\frac{\partial(\boldsymbol{miF})}{\partial(\boldsymbol{p})}\right] = \begin{bmatrix} A_{miF}^W & C_{miF}^W \\ C_{miF}^W & B_{miF}^W \end{bmatrix},$$

with

$$A_{miF}^W = miF_1(1 - miF_1),$$

$$B_{miF}^W = miF_2(1 - miF_2),$$

$$C_{miF}^W = \sum_{a=1}^r p_{aaa} - miF_1 miF_2\,.$$

Therefore, the variance of $\left(\widehat{miF}_1 - \widehat{miF}_2\right)$ is

$$\left(A^W_{miF} + B^W_{miF} - 2C^W_{miF}\right)/n\,.$$

A.3 Variance of maF

The derivation of the variance of $\left(\widehat{maF}_1 - \widehat{maF}_2\right)$ is as follows.

$$\begin{aligned}
&\boldsymbol{maF} = \begin{bmatrix} maF_1 \\ maF_2 \end{bmatrix}, \\
&\left[\frac{\partial(\boldsymbol{maF})}{\partial(\boldsymbol{p})}\right]^T \left(\boldsymbol{p}\boldsymbol{p}^T\right)\left[\frac{\partial(\boldsymbol{maF})}{\partial(\boldsymbol{p})}\right] = \boldsymbol{0}, \\
&\left[\frac{\partial(\boldsymbol{maF})}{\partial(\boldsymbol{p})}\right]^T \left(diag\left(\boldsymbol{p}\right) - \boldsymbol{p}\boldsymbol{p}^T\right)\left[\frac{\partial(\boldsymbol{maF})}{\partial(\boldsymbol{p})}\right] = \left[\frac{\partial(\boldsymbol{maF})}{\partial(\boldsymbol{p})}\right]^T (diag(\boldsymbol{p}))\left[\frac{\partial(\boldsymbol{maF})}{\partial(\boldsymbol{p})}\right] \\
&\quad = \frac{1}{r^2}\begin{bmatrix} A^W_{maF} & C^W_{maF} \\ C^W_{maF} & B^W_{maF} \end{bmatrix},
\end{aligned}$$

with

$$\begin{aligned}
A^W_{maF} &= \sum_{a=1}^{r} p_{a..a}\left(\frac{2(1 - F_{1a})}{p_{a..} + p_{..a}}\right)^2 + \sum_{a=1}^{r}\sum_{b\neq a} p_{a..b}\left(\frac{F_{1a}}{p_{a..} + p_{..a}} + \frac{F_{1b}}{p_{b..} + p_{..b}}\right)^2, \\
B^W_{maF} &= \sum_{a=1}^{r} p_{.aa}\left(\frac{2(1 - F_{2a})}{p_{.a.} + p_{..a}}\right)^2 + \sum_{a=1}^{r}\sum_{b\neq a} p_{.ab}\left(\frac{F_{2a}}{p_{.a.} + p_{..a}} + \frac{F_{2b}}{p_{.b.} + p_{..b}}\right)^2, \\
C^W_{maF} &= \sum_{a=1}^{r} p_{aaa}\frac{2^2(1 - F_{1a})(1 - F_{2a})}{(p_{a..} + p_{..a})(p_{.a.} + p_{..a})} \\
&\quad - \sum_{a=1}^{r}\sum_{b\neq a}\left\{p_{aba}\frac{2(1 - F_{1a})}{(p_{a..} + p_{..a})}\left(\frac{F_{2a}}{(p_{.a.} + p_{..a})} + \frac{F_{2b}}{(p_{.b.} + p_{..b})}\right) + p_{baa}\frac{2(1 - F_{2a})}{(p_{.a.} + p_{..a})}\left(\frac{F_{1a}}{(p_{a..} + p_{..a})} + \frac{F_{1b}}{(p_{b..} + p_{..b})}\right)\right\} \\
&\quad + \sum_{a=1}^{r}\sum_{b\neq a}\sum_{c\neq a} p_{bca}\left(\frac{F_{1a}}{(p_{a..} + p_{..a})} + \frac{F_{1b}}{(p_{b..} + p_{..b})}\right)\left(\frac{F_{2a}}{(p_{.a.} + p_{..a})} + \frac{F_{2c}}{(p_{.c.} + p_{..c})}\right).
\end{aligned}$$

Therefore, the variance of $\left(\widehat{maF}_1 - \widehat{maF}_2\right)$ is

$$\frac{1}{r^2}\left(A^W_{\mathrm{maF}} + B^W_{\mathrm{maF}} - 2C^W_{\mathrm{maF}}\right)/n\,.$$

A.4 Variance of maF^*

The derivation of the variance of $\left(\widehat{maF^*}_1 - \widehat{maF^*}_2\right)$ is as follows.

$$\boldsymbol{maF}^* = \begin{pmatrix} maF_1^* \\ maF_2^* \end{pmatrix},$$

$$\left[\frac{\partial(\boldsymbol{maF}^*)}{\partial(\boldsymbol{p})}\right]^T \left(\boldsymbol{pp}^T\right)\left[\frac{\partial(\boldsymbol{maF}^*)}{\partial(\boldsymbol{p})}\right] = \mathbf{0},$$

$$\left[\frac{\partial(\boldsymbol{maF}^*)}{\partial(\boldsymbol{p})}\right]^T \left(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{pp}^T\right)\left[\frac{\partial(\boldsymbol{maF}^*)}{\partial(\boldsymbol{p})}\right] = \left[\frac{\partial(\boldsymbol{maF}^*)}{\partial(\boldsymbol{p})}\right]^T (\mathrm{diag}(\boldsymbol{p}))\left[\frac{\partial(\boldsymbol{maF}^*)}{\partial(\boldsymbol{p})}\right]$$
$$= \begin{pmatrix} A_{maF^*}^W & C_{maF^*}^W \\ C_{maF^*}^W & B_{maF^*}^W \end{pmatrix},$$

with

$$A_{maF^*}^W = \frac{2^2}{r^2(maP_1 + maR_1)^4}\left[\sum_{a=1}^{r} p_{a.a}\left(\frac{(p_{a..} - p_{a.a})maR_1^2}{p_{a..}^2} + \frac{(p_{\cdot.a} - p_{a.a})maP_1^2}{p_{\cdot.a}^2}\right)^2 + \sum_{a=1}^{r}\sum_{b\neq a} p_{a.b}\left(\frac{p_{a.a}maR_1^2}{p_{a..}^2} + \frac{p_{b.b}maP_1^2}{p_{\cdot.b}^2}\right)^2\right],$$

$$B_{maF^*}^W = \frac{2^2}{r^2(maP_2 + maR_2)^4}\left[\sum_{a=1}^{r} p_{\cdot aa}\left(\frac{(p_{\cdot a.} - p_{\cdot aa})maR_2^2}{p_{\cdot a.}^2} + \frac{(p_{\cdot.a} - p_{\cdot aa})maP_2^2}{p_{\cdot.a}^2}\right)^2 + \sum_{a=1}^{r}\sum_{b\neq a} p_{\cdot ab}\left(\frac{p_{\cdot aa}maR_2^2}{p_{\cdot a.}^2} + \frac{p_{\cdot bb}maP_2^2}{p_{\cdot.b}^2}\right)^2\right],$$

$$C_{maF^*}^W = \frac{2^2}{r^2(maP_1 + maR_1)^2(maP_2 + maR_2)^2}$$
$$\times\left[\sum_{a=1}^{r} p_{aaa}\left(\frac{p_{a..} - p_{a.a}}{p_{a..}^2}maR_1^2 + \frac{p_{\cdot.a} - p_{a.a}}{p_{\cdot.a}^2}maP_1^2\right)\left(\frac{p_{\cdot a.} - p_{\cdot aa}}{p_{\cdot a.}^2}maR_2^2 + \frac{p_{\cdot.a} - p_{\cdot aa}}{p_{\cdot.a}^2}maP_2^2\right)\right.$$
$$-\sum_{a=1}^{r}\sum_{b\neq a}\left\{p_{aba}\left(\frac{p_{a..} - p_{a.a}}{p_{a..}^2}maR_1^2 + \frac{p_{\cdot.a} - p_{a.a}}{p_{\cdot.a}^2}maP_1^2\right)\left(\frac{p_{\cdot.bb}}{p_{\cdot.b}^2}maR_2^2 + \frac{p_{\cdot.bb}}{p_{\cdot.b}^2}maP_2^2\right)\right.$$
$$\left.+\ p_{baa}\left(\frac{p_{\cdot a.} - p_{\cdot aa}}{p_{\cdot a.}^2}maR_2^2 + \frac{p_{\cdot.a} - p_{\cdot aa}}{p_{\cdot.a}^2}maP_2^2\right)\left(\frac{p_{b.b}}{p_{b..}^2}maR_1^2 + \frac{p_{b.b}}{p_{\cdot.b}^2}maP_1^2\right)\right\}$$
$$\left.+\sum_{a=1}^{r}\sum_{b\neq a}\sum_{c\neq a} p_{bca}\left(\frac{p_{b.b}}{p_{b..}^2}maR_1^2 + \frac{p_{a.a}}{p_{\cdot.a}^2}maP_1^2\right)\left(\frac{p_{\cdot.cc}}{p_{\cdot.c}^2}maR_2^2 + \frac{p_{\cdot.aa}}{p_{\cdot.a}^2}maP_2^2\right)\right].$$

Therefore, the variance of $\left(\widehat{maF^*}_1 - \widehat{maF^*}_2\right)$ is

$$\left(A_{maF^*}^W + B_{maF^*}^W - 2C_{maF^*}^W\right)/n\,.$$

APPENDIX B.: EXAMPLE DATA

Tables B1, B2, and B3 here.

TABLE B1

Example data.

		True condition					
FRCNN	**BCD**	**MM**	**BCC**	**Nevus**	**SK**	**H/H**	**SL**
MM	MM	289	2	20	6	2	0
	BCC	9	2	2	5	0	0
	Nevus	10	0	14	0	0	0
	SK	14	2	4	10	0	0
	H/H	3	0	2	0	1	0
	SL	2	0	0	0	0	0
BCC	MM	6	6	0	1	0	0
	BCC	2	95	0	6	0	0
	Nevus	0	2	6	0	0	0
	SK	1	5	0	2	0	0
	H/H	0	0	0	0	0	0
	SL	0	0	0	0	0	0
Nevus	MM	32	1	108	0	1	0
	BCC	1	6	8	1	0	0
	Nevus	11	1	789	8	1	0
	SK	3	3	50	27	0	0
	H/H	0	1	9	0	16	0
	SL	1	0	3	0	0	0
SK	MM	13	1	3	11	0	0
	BCC	0	1	1	12	0	0
	Nevus	1	0	11	9	0	0
	SK	7	4	14	186	0	1
	H/H	0	0	0	0	0	0
	SL	0	0	1	5	0	2
H/H	MM	0	0	0	0	6	0
	BCC	0	0	0	0	1	0
	Nevus	0	0	3	0	5	0
	SK	0	0	0	0	1	0
	H/H	0	0	0	0	44	0
	SL	0	0	0	0	0	0
SL	MM	0	0	0	0	0	0
	BCC	0	0	0	0	0	1
	Nevus	0	0	0	0	0	0
	SK	1	0	0	0	0	6
	H/H	0	0	0	0	0	0
	SL	2	0	0	0	0	35

TABLE B2

Example data (FRCNN only).

	True condition					
FRCNN	**MM**	**BCC**	**Nevus**	**SK**	**H/H**	**SL**
MM	327	6	42	21	3	0
BCC	9	108	6	9	0	0
Nevus	48	12	967	36	18	0
SK	21	6	30	223	0	3
H/H	0	0	3	0	57	0
SL	3	0	0	0	0	42

TABLE B3

Example data (BCD only).

	True condition					
BCD	**MM**	**BCC**	**Nevus**	**SK**	**H/H**	**SL**
MM	340	10	131	18	9	0
BCC	12	104	11	24	1	1
Nevus	22	3	823	17	6	0
SK	26	14	68	225	1	7
H/H	3	1	11	0	61	0
SL	5	0	4	5	0	37

REFERENCES

1. van Rijsbergen CJ. Information Retrieval. London: Butterworths; 1979.
2. Manning CD, Raghavan P, Schütze H. Introduction to Information Retrieval. Cambridge: Cambridge University Press; 2008.
3. Jinnai S, Yamazaki N, Hirano Y, Sugawara Y, Ohe Y, Hamamoto R. The development of a skin cancer classification system for pigmented skin lesions using deep learning. Biomolecules. 2020;10(8):1123. doi:10.3390/biom10081123 [PubMed: 32751349]
4. Altman DG, Royston P. The cost of dichotomising continuous variables. Bmj. 2006;332(7549):1080. [PubMed: 16675816]
5. Fedorov V, Mannino F, Zhang R. Consequences of dichotomization. Pharm Stat. 2009;8(1):50–61. [PubMed: 18389492]
6. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag. 2009;45:427–437.
7. Bhalla S, Kaur H, Kaur R, Sharma S, Raghava GPS. Expression based biomarkers and models to classify early and late-stage samples of papillary thyroid carcinoma. PloS One. 2020;15(4):e0231629. [PubMed: 32324757]
8. Chowdhury S, Dong X, Qian L, et al. A multitask bi-directional RNN model for named entity recognition on Chinese electronic medical records. BMC Bioinform. 2018;19(17):499.
9. Döring K, Qaseem A, Becer M, et al. Automated recognition of functional compound-protein relationships in literature. PloS One. 2020;15(3):e0220925. [PubMed: 32126064]

10. Hong N, Wen A, Stone DJ, et al. Developing a FHIR-based EHR phenotyping framework: a case study for identification of patients with obesity and multiple comorbidities from discharge summaries. J Biomed Inform. 2019;99:103310. [PubMed: 31622801]
11. Lee GH, Shin SY. Federated learning on clinical benchmark data: performance assessment. J Med Internet Res. 2020;22(10):e20891. [PubMed: 33104011]
12. Routray R, Tetarenko N, Abu-Assal C, et al. Application of augmented intelligence for pharmacovigilance case seriousness determination. Drug Saf. 2020;43(1):57–66. [PubMed: 31605285]
13. Wang J, Zhang J, An Y, et al. Biomedical event trigger detection by dependency-based word embedding. BMC Med Genom. 2016; 9(2):45.
14. Zhu F, Li X, Mcgonigle D, et al. Analyze informant-based questionnaire for the early diagnosis of senile dementia using deep learning. IEEE J Transl Eng Health Med. 2020;8:2200106. [PubMed: 31966933]
15. Wang Y, Li J, Li Y, Wangi R, Yang X. Confidence interval for F_1 measure of algorithm performance based on blocked 3 × 2 cross-validation. IEEE Trans Knowl Data Eng. 2015;27:651–659.
16. Zhang D, Wang J, Zhao X. Estimating the uncertainty of average F_1 scores. Proceedings of the 2015 International Conference on the Theory of Information Retrieval. 2015.
17. Takahashi K, Yamamoto K, Kuchiba A, Koyama T. Confidence interval for micro-averaged F_1 and macro-averaged F_1 scores. Appl Intell (Dordr). 2022;52(5):4961–4972. [PubMed: 35317080]
18. Attia ZI, Noseworthy PA, Lopez-Jimenez F, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. Lancet. 2019;394(10201):861–867. [PubMed: 31378392]
19. Han SS, Moon IJ, Lim W, et al. Keratinocytic skin cancer detection on the face using region-based convolutional neural network. JAMA Dermatol. 2020;156(1):29–37. [PubMed: 31799995]

TABLE 1

General notations.

		True condition = 1					⋯	**True condition = *r***					
		Test 2						**Test 2**					
		1	**2**	⋯	**r**			**1**	**2**	⋯	**r**		
Test 1	1	p_{111}	p_{121}	⋯	p_{1r1}	$p_{1.1}$		p_{11r}	p_{12r}	⋯	p_{1rr}	$p_{1.r}$	$p_{1..}$
	2	p_{211}	p_{221}	⋯	p_{2r1}	$p_{2.1}$		p_{21r}	p_{22r}	⋯	p_{2rr}	$p_{2.r}$	$p_{2..}$
	⋮	⋮	⋮	⋱	⋮	⋮		⋮	⋮	⋱	⋮	⋮	⋮
	r	p_{r11}	p_{r21}	⋯	p_{rr1}	$p_{r.1}$		p_{r1r}	p_{r2r}	⋯	p_{rrr}	$p_{r.r}$	$p_{r..}$
		$p_{.11}$	$p_{.21}$	⋯	$p_{.r1}$	$p_{..1}$		$p_{.1r}$	$p_{.2r}$	⋯	$p_{.rr}$	$p_{..r}$	1

TABLE 2

Simulation study: True cell probabilities.

		True class = 1			True class = 2			True class = 3		
		Test 2			Test 2			Test 2		
Scenario 1		**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
Test 1	1	40/300	10/300	10/300	5/300	10/300	5/300	5/300	5/300	10/300
	2	10/300	5/300	5/300	10/300	40/300	10/300	5/300	5/300	10/300
	3	10/300	5/300	5/300	5/300	10/300	5/300	10/300	10/300	40/300

$biF_1 = miF_1 = maF_1 = maF_1^* = 0.60$

$biF_2 = miF_2 = maF_2 = maF_2^* = 0.60$

		True class = 1			True class = 2			True class = 3		
		Test 2			Test 2			Test 2		
Scenario 2		**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
Test 1	1	120/500	30/500	30/500	5/500	10/500	5/500	5/500	5/500	10/500
	2	30/500	15/500	15/300	10/500	40/500	10/500	5/500	5/500	10/500
	3	30/500	15/500	15/500	5/500	10/500	5/500	10/500	10/500	40/500

$biF_1 = 0.69, miF_1 = 0.60, maF_1 = 0.56, maF_1^* = 0.58$

$biF_2 = 0.69, miF_2 = 0.60, maF_2 = 0.56, maF_2^* = 0.58$

		True class = 1			True class = 2			True class = 3		
		Test 2			Test 2			Test 2		
Scenario 3		**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
Test 1	1	30/300	15/300	15/300	5/300	10/300	5/300	5/300	5/300	10/300
	2	10/300	5/300	5/300	15/300	30/300	15/300	5/300	5/300	10/300
	3	10/300	5/300	5/300	5/300	10/300	5/300	15/300	15/300	30/300

$biF_1 = miF_1 = maF_1 = maF_1^* = 0.60$

$biF_2 = miF_2 = maF_2 = maF_2^* = 0.50$

		True class = 1			True class = 2			True class = 3		
		Test 2			Test 2			Test 2		
Scenario 4		**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
Test 1	1	90/500	45/500	45/500	5/500	10/500	5/500	5/500	5/500	10/500
	2	30/500	15/500	15/300	15/500	30/500	15/500	5/500	5/500	10/500
	3	30/500	15/500	15/500	5/500	10/500	5/500	15/500	15/500	30/500

$biF_1 = 0.69, miF_1 = 0.60, maF_1 = 0.56, maF_1^* = 0.58$

$biF_2 = 0.60, miF_2 = 0.50, maF_2 = 0.47, maF_2^* = 0.49$

TABLE 3

Simulation study: Empirical type I error rates.

Scenario	N	T^{W}_{biF}	T^{S}_{biF}	T^{W}_{miF}	T^{S}_{miF}	T^{W}_{maF}	T^{S}_{maF}	T^{W}_{maF*}	T^{S}_{maF*}
1	100	0.057	0.050	0.053	0.049	0.055	0.051	0.057	0.053
	300	0.052	0.050	0.051	0.050	0.052	0.051	0.052	0.051
	500	0.051	0.050	0.050	0.050	0.051	0.050	0.051	0.050
	1000	0.050	0.049	0.051	0.050	0.051	0.050	0.051	0.051
2	100	0.052	0.049	0.054	0.049	0.058	0.053	0.061	0.055
	300	0.052	0.050	0.051	0.050	0.054	0.052	0.054	0.052
	500	0.051	0.050	0.051	0.050	0.051	0.050	0.052	0.051
	1000	0.050	0.050	0.051	0.051	0.052	0.051	0.051	0.051

TABLE 4

Simulation study: Empirical power.

Scenario	N	T^{W}_{biF}	T^{S}_{biF}	T^{W}_{miF}	T^{S}_{miF}	T^{W}_{maF}	T^{S}_{maF}	T^{W}_{maF*}	T^{S}_{maF*}
3	100	0.192	0.174	0.304	0.289	0.309	0.297	0.310	0.300
	300	0.438	0.429	0.694	0.689	0.696	0.692	0.696	0.692
	500	0.641	0.635	0.890	0.888	0.889	0.888	0.889	0.888
	1000	0.905	0.904	0.995	0.995	0.995	0.995	0.995	0.995
4	100	0.235	0.226	0.305	0.291	0.291	0.278	0.271	0.256
	300	0.560	0.556	0.695	0.690	0.662	0.657	0.615	0.609
	500	0.773	0.771	0.889	0.887	0.865	0.863	0.826	0.824
	1000	0.969	0.969	0.995	0.995	0.992	0.992	0.984	0.984

TABLE 5

Example.

	FRCNN	BCD	Difference	Test statistics	p-value
biF (Wald)	0.840	0.776	0.064	19.4	< 0.001
biF (score)	0.840	0.776	0.064	18.9	< 0.001
miF (Wald)	0.862	0.795	0.067	41.9	< 0.001
miF (score)	0.862	0.795	0.067	41.0	< 0.001
maF (Wald)	0.846	0.768	0.078	26.2	< 0.001
maF (score)	0.846	0.768	0.078	24.5	< 0.001
maF* (Wald)	0.848	0.772	0.076	26.4	< 0.001
maF* (score)	0.848	0.772	0.076	23.0	< 0.001